Homework #3

Analysis of Arsenic in Rice Products

Antoine Baldassari

November 17, 2015

All .Rcode and relevant files can be accessed at https://github.com/kaskarn/Homework_3_779

1. Standard conditionally-conjugate specification of the hierarchical model

1.1 Model specification

In this model specification, the i^{th} arsenic reading of group j, y_{ij} is normally-distributed, so that

$$y_{ij} \sim \mathcal{N}\left(\theta_j, \sigma^2\right)$$

Where θ_j is the mean arsenic reading for the rice products indexed by j. θ_j is normally distributed, centered at the population mean μ, with between-group variance τ^2:

$$\theta_j \sim \mathcal{N}\left(\mu, \tau^2\right)$$

We use conditionally-conjugate Normal and Inverse-Gamma priors on the hyperparameters:

$$\begin{aligned} 1/\sigma^2 &\sim \text{gamma}\ \left(\nu_0/2, \nu_0\sigma_0^2/2\right) \\ 1/\tau^2 &\sim \text{gamma}\ \left(\eta_0/2, \eta_0\tau_0^2/2\right) \\ \mu &\sim \text{normal}\ \left(\mu_0, \gamma_0^2\right) \end{aligned}$$

The full conditional distribution of the parameters can be found to be (from the book):

$$\begin{aligned} \left\{\theta_j \,|\, \sigma^2, y_{j,1}, \ldots, y_{j,n}\right\} &\sim \mathcal{N}\left(\frac{n_j\bar{y}_j/\sigma^2 + \mu/\tau^2}{n_j/\sigma^2 + 1/\tau^2}, \left[n_j/\sigma^2 + 1/\tau^2\right]^{-1}\right) \\ \left\{\mu \,|\, \theta_1, \ldots, \theta_m, \tau\right\} &\sim \mathcal{N}\left(\frac{m\bar{\theta}/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \left[m/\tau^2 + 1/\gamma_0^2\right]^{-1}\right) \\ \left\{1/\tau^2 \,|\, \theta_1, \ldots, \theta_m, \mu\right\} &\sim \text{gamma}\left(\frac{\eta_0 + m}{2}, \frac{\eta_o\tau_0^2 + \sum\left(\theta_j - \mu\right)^2}{2}\right) \\ \left\{1/\sigma^2 \,|\, \boldsymbol{\theta}, y_1, \ldots, y_n\right\} &\sim \text{gamma}\left(\frac{1}{2}\left[\nu_0 + \sum_{j=1}^{m} n_j\right], \frac{1}{2}\left(\nu_0\sigma_0^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(y_{i,j} - \theta_j\right)^2\right)\right) \end{aligned}$$

1.2 Main analyses

We pick relatively uninformative priors, centering μ around 1 with somewhat large within and between sample variances: $\sigma_0^2 = 10, \nu_0 = 1, \tau_0^2 = 10, \eta_0 = 1, \gamma_0^2 = 10$. The marginal distributions of $\theta_1, \ldots, \theta_m, \mu, \sigma^2$ and τ^2 can be obtained from the full condition distributions using a Monte-Carlo Markov-Chain algorithm, Gibbs sampling, which we implement in R as follows::

First, we input the dataset downloaded from Sakai, modified in Stata to have numeric codes for rice products categories.

```
library(foreign)
Y <- read.dta(file="arsenicrice2.dta")
```

We set the weakly informative prior values

```
n <-  nrow(Y)
nu0 <- 1; eta0 <- 1;
t20 <- 10;
mu0 <- 1;
g20 <- s20 <- var(Y$arsenic)
```

We set initial values for algorithm

```
m <- length(unique(Y$food_num)) #number of groups
n <- sv <- ybar <- rep(NA,m)
for (i in 1:m)
{
  n[i] <- sum(Y$food_num==i)
  sv[i] <- var(Y$arsenic[which(Y$food_num==i)])
  ybar[i] <- mean(Y$arsenic[which(Y$food_num==i)])
}
theta <- ybar; s2 <- mean(sv)
mu <- mean(theta); tau2 <- var(theta)
```

We create a Markov chain for each parameter by sequentially sampling from their posterior over 10,000 iterations. Elements are stored in the chain at the end of each iteration.

```
#Setup MCMC
set.seed(0808)
S <- 10000
THETA <- matrix(nrow=S, ncol=m)
OTH <- matrix(nrow=S, ncol=3)
ALL <- matrix(nrow=S, ncol=3+m)

#Run algorithm
for(i in 1:S)
{
  #Get new values for parameters
  for(j in 1:m) theta[j] <- newTheta(n[j], ybar[j], s2, tau2, mu)
  s2 <- newSigma2(m, n, nu0, s20, theta, Y)
  mu <- newMu(m, theta, tau2, g20)
  tau2 <- newTau2(m, eta0, t20, theta, mu)
```

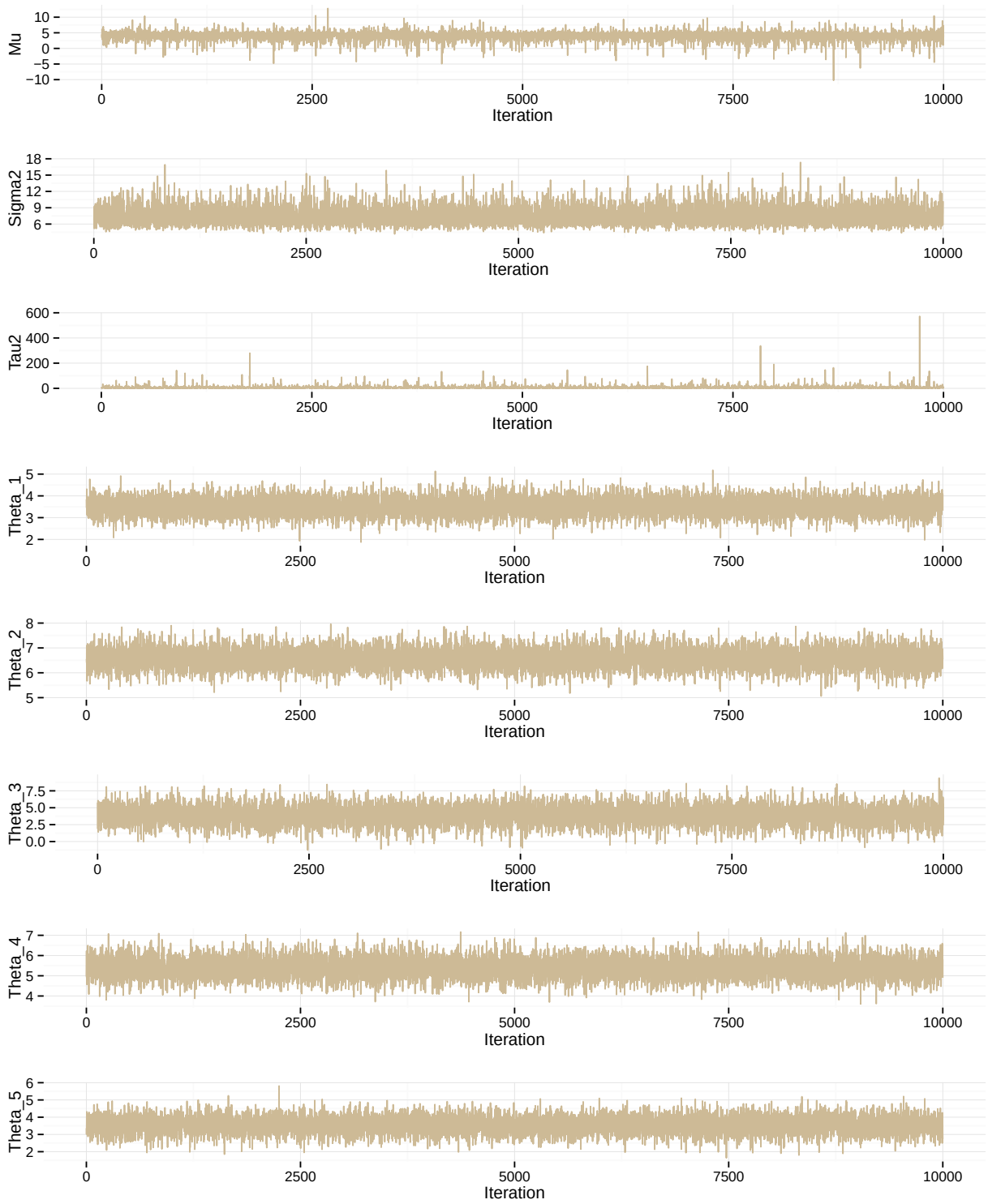
```
  #Store in chain
  THETA[i,] <- theta
  OTH[i,] <- c(mu,s2,tau2)
  ALL[i,] <- c(theta,mu,s2,tau2)
}
```

Where the functions updating the parameters follow the equations listed above:

```
newTheta <- function(n, ybar, s2, tau2, mu)
{
  v = 1/(n/s2 +1/tau2)
  e = v * (ybar*n/s2 +mu/tau2)
  new <- rnorm(1, e, sqrt(v))
  return(new)
}
newSigma2 <- function(m, n, nu0, s20, theta, Y)
{
  nun = nu0 + sum(n)
  ss <- nu0 * s20
  for(i in 1:m) ss = ss+sum((Y$arsenic[which(Y$food_num==i)] - theta[j])^2)
  sigma2 <- 1/rgamma(1, nun/2, ss/2)
  return(sigma2)
}
newMu <- function(m, theta, tau2, g20)
{
  v = 1/(m/tau2 + 1/g20)
  e = v *(m*mean(theta)/tau2 + mu0/g20)
  mu <- rnorm(1, e, v)
  return(mu)
}
newTau2 <- function(m, eta0, t20, theta, mu)
{
  etam = eta0 + m
  ss <- eta0*t20 + sum( (theta-mu) ^2 )
  tau2 <- 1/rgamma(1, etam/2, ss/2)
  return(tau2)
}
```

Before we go any further, we check that the MCMC model converged for all four statistics using ggplot2 (code used for μ repeated for other parameters):

```
library(ggplot2)
graphdata <- data.frame(
              "Iteration"=c(1:S), "Mu"=OTH[,1], "Sigma2"=OTH[,2], "Tau2"=OTH[,3],
              "Theta_1" = THETA[,1], "Theta_2" = THETA[,2], "Theta_3" = THETA[,3],
              "Theta_4" = THETA[,4], "Theta_5" = THETA[,5])

ggplot(graphdata,aes(x=Iteration,y=Mu)) +
theme_minimal(base_family = "") + geom_line(colour="wheat3")
```



We conclude from the graphs that convergence was achieved for all parameters.

1.3 Algorithm output

The estimated median values and 95% credible intervals for the parameters are as follow:

```
for(i in 1:length(ALL[1,])) print(round(unname(
  quantile(ALL[,i], probs=c(0.025, 0.5, 0.975))
  ),3))
```

Parameter	Credible Lower 95%	Median	Credible Upper 95%
θ_1 (Basmati)	2.810	3.569	4.308
θ_2 (Non-Basmati)	5.790	6.533	7.305
θ_3 (Beverage)	1.869	4.231	6.448
θ_4 (Cakes)	4.486	5.370	6.301
θ_5 (Cereal)	2.705	3.610	4.507
μ	3.193	4.697	6.182
σ^2	5.109	7.027	10.573
τ^2	0.693	2.251	12.920

1.4 Sensitivity analyses

Evaluation of sensitivity to priors: we try three separate scenarios each tuning prior distribution of parameters:

1. Large expected μ (Prior expectation of mad levels of arsenic)
2. Large σ^2 and ν_0 (High variability within products)
3. Large τ^2 and η_0 (High variability between products)

Scenario 1:

Parameter	Credible Lower 95%	Median	Credible Upper 95%
θ_1 (Basmati)	2.706	3.488	4.265
θ_2^2 (Non-Basmati)	5.902	6.671	7.460
$\theta_3(Beverage)$	0.643	3.774	6.932
$\theta_4(Cakes)$	4.511	5.452	6.429
θ_5 (Cereal)	2.548	3.496	4.456
μ	2.548	3.496	4.456
σ^2	88.793	100	111
τ^2	3021.927	8427	37962

Scenario 2:

Parameter	Credible Lower 95%	Median	Credible Upper 95%
θ_1 (Basmati)	1.237	4.396	7.069
θ_2^2 (Non-Basmati)	2.811	5.401	8.559
$\theta_3(Beverage)$	0.278	4.790	9.023
$\theta_4(Cakes)$	1.931	4.960	8.221
θ_5 (Cereal)	1.137	4.531	7.508
μ	2.42	4.84	7.19
σ^2	182	215	256
τ^2	0.399	1.876	22.3

Scenario 3:

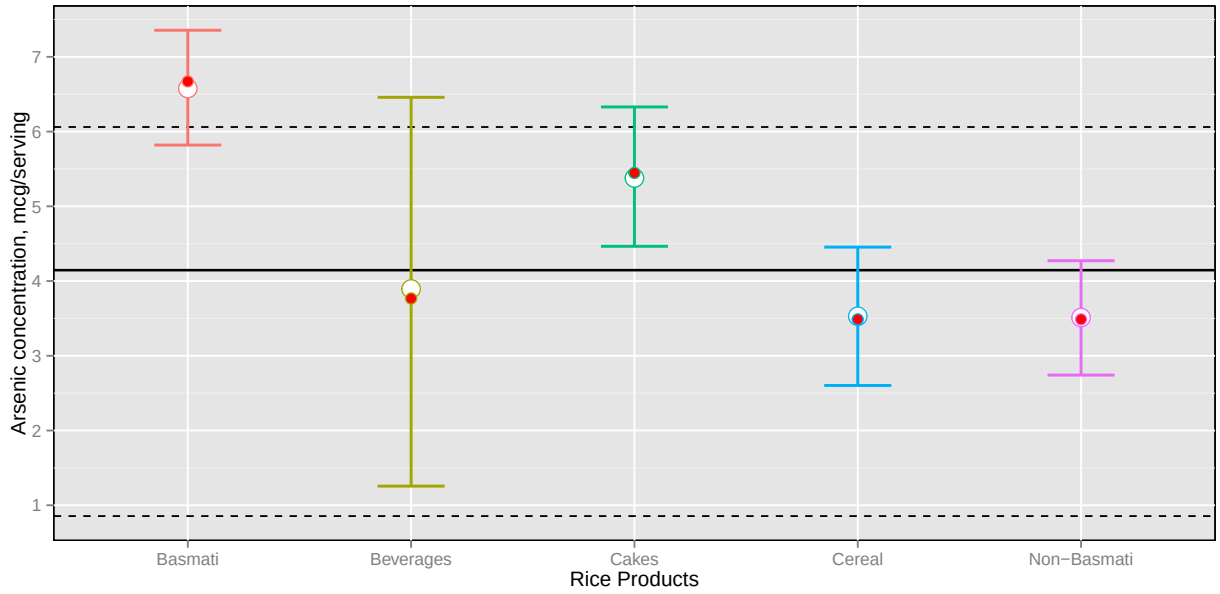
Parameter	Credible Lower 95%	Median	Credible Upper 95%
θ_1 (Basmati)	2.707	3.478	4.268
θ_2^2 (Non-Basmati)	5.907	6.674	7.451
$\theta_3(Beverage)$	0.676	3.739	6.9023
$\theta_4(Cakes)$	4.504	5.447	6.419
θ_5 (Cereal)	2.547	3.494	4.452
μ	-5.302	4.845	5.00
σ^2	5.197	7.327	11.37
τ^2	223	288	384

We observe that excessively large prior expectations of μ will drive up estimates of the within- and between-group variances but will have little effect on the magnitude of the estimates of within-groupmean estimates (although the precision may be negatively affected for groups with realtively few observations). A large prior within-sample variance will bring posterior within-group means closer to μ, as could be expected since the posterior estimates need to become more conservative. Increasing prior between-sample variance appears to drive up uncertainty on μ and bring it closer to 0, without however having a notable impact on the rest of the model.

1.5 Results presentation

Non-Basmati rice had the highest arsenic concentration, at an estimated 6.7 mcg/serving. Rice cakes came second, at 5.4 mcg/serving, and non-Basmati and rice cereal had comparatively low amounts, slightly below 3.5 mcg/serving. There lacked data to reliably evaluate arsenic concentration in rice beverages, whose 3.8 mcg/serving estimate was particulary imprecise (95% CI=0.64, 6.93). Posterior median estimates and observed mean concentrations of arsenic are presented by product type in the following graph. Markers are θ estimates with 95% credible interval lines; horizontal lines are the median estimate of μ (solid) and corresponding 95% credibal interval (dashed).

```
qmat=apply(THETA[,1:5],2,quantile,probs=c(0.025,.5,0.975))
mu_ci = quantile(OTH[,1], probs=(c(0.025, 0.5, 0.975)))
res <- data.frame("Rice"=c("Non-Basmati", "Basmati", "Beverages", "Cakes", "Cereal"),
                  "l95"=qmat[1,], "median"=qmat[2,], "u95"=qmat[3,], "mean"=ybar)
g <- ggplot(res, aes(x = Rice, group=Rice, colour=Rice)) +
  labs(x="Rice Products", y="Arsenic concentration, mcg/serving") +
  theme(legend.position="none", panel.background =element_rect(colour = "black")) +
  scale_y_continuous(breaks=seq(0, 7.5, 1)) +
  geom_hline(aes(yintercept=c(mu_ci[2])), size=0.7) +
  geom_hline(aes(yintercept=c(mu_ci[1])), linetype="dashed") +
  geom_hline(aes(yintercept=c(mu_ci[3])), linetype="dashed") +
  geom_errorbar(aes(ymin=l95, ymax=u95), width=.3, size=0.8) +
  geom_point(aes(y=median), fill="white", shape=21, size=5)  +
  geom_point(aes(y=mean), fill="red", shape=21, size=3)
g
```


Arsenic concentration, mcg/serving
7
6
5
4
3
2
1
Basmati
Beverages
Cakes
Cereal
Non−Basmati
Rice Products

2. Parameter-expanded specification of the hierarchical model

2.1 Model specification

Under this model specification, instead of group means we are interested in differences between groups and the population average μ, which is given by η_j for the group j, so that (under the prior belief that all groups will have equal mean):

$$\begin{aligned} y_{ij} &\sim \mathcal{N}\left(\mu + \xi\eta_j, \sigma_y^2\right) \\ \eta_i &\sim \mathcal{N}\left(0, \sigma_\eta^2\right) \end{aligned}$$

Word's on the street that well-behaved conditionally-conjugate specifications for the distributions of ξ and σ_η^2 are:

$$\begin{aligned} \xi &\sim \mathcal{N}(0,1) \\ 1/\sigma_\eta^2 &\sim \text{gamma}\left(\frac{\omega_0}{2}, \frac{\omega_0\sigma_{\eta 0}^2}{2}\right) \\ 1/\sigma_y^2 &\sim \text{gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_{y0}^2}{2}\right) \end{aligned}$$

The prior distribution of the population mean is still $\mu \sim \mathcal{N}\left(\mu_0, \gamma_0^2\right)$. We set out to find full conditionals: Reparametrizing the full conditionals in exercise 1 easily yields what we need:

Note that each of the N observation in the data supports, $\mu = y_{ij} - xi\eta_j$ given known ξ and η_j's, so that, summing over this expression and weighting it against the prior yields:

$$p\left(\mu, \sigma_y^2, \sigma_\eta^2, \xi, \eta_1, \ldots, \eta_m \,|\, \mathbf{y}\right) \alpha\, \mathcal{N}\left(\frac{\frac{\sum_j \sum_i y_{ij} - \xi\eta_j}{\sigma_y^2} + \frac{\mu_0}{\gamma_0^2}}{\frac{N}{\sigma_y^2} + \frac{1}{\gamma_0^2}}, \frac{1}{\frac{N}{\sigma_y^2} + \frac{1}{\gamma_0^2}}\right)$$

Posterior on the variances can be similarly rewritten. The sum of squares for $\mathbf{y}$ given μ, η_j and ξ is, of course, $\sum_j \sum_i y_{ij} - \mu - \xi\eta_j$, and the sum of squares of $\boldsymbol{\eta}$ given its $\mathbb{E} = 0$ is simply $\sum_j \eta_j^2$

$$\begin{aligned} p\left(1/\sigma_y^2 \,|\, \mathbf{y}, \mu, \sigma_\eta^2, \xi, \eta_1, \ldots, \eta_m\right) \alpha\ &\text{gamma}\ \left(\frac{\nu_0 + N}{2}, \frac{1}{2}\left(\nu_0\sigma_{y0}^2 + \sum_j \sum_i \left(y_{ij} - \mu - \xi\eta_j\right)^2\right)\right) \\ p\left(1/\sigma_\eta^2 \,|\, \mathbf{y}, \mu, \sigma_y^2, \xi, \eta_1, \ldots, \eta_m\right) \alpha\ &\text{gamma}\ \left(\frac{\omega_0 + m}{2}, \frac{1}{2}\left(\omega_0\sigma_{\eta_0}^2 + \sum_j \eta_j^2\right)\right) \end{aligned}$$

Getting full conditionals on $\boldsymbol{\eta}$ and ξ can be likewise achieved by manipulating $y_{ij} = \mu + \eta_j\xi$.

$$p\left(\eta_j \,|\, \mathbf{y}, \mu, \sigma_y^2, \sigma_\eta^2, \xi\right) \alpha\, \mathcal{N}\left(\frac{\frac{\xi \sum\limits_i y_{ij} - \mu}{\sigma_y^2}}{\frac{n_j \xi^2}{\sigma_y^2} + \frac{1}{\sigma_\eta^2}}, \frac{1}{\frac{n_j \xi^2}{\sigma_y^2} + \frac{1}{\sigma_\eta^2}}\right)$$

and

$$p\left(\xi \,|\, \mathbf{y}, \mu, \sigma_y^2, \sigma_\eta^2, \eta_1, \ldots, \eta_m\right) \alpha\, \mathcal{N}\left(\frac{\frac{\sum\limits_j \eta_j \sum\limits_i (y_{ij} - \mu)}{\sigma_y^2}}{\frac{\sum\limits_j n_j \eta_j^2}{\sigma_y^2} + 1}, \frac{1}{\frac{\sum\limits_j n_j \eta_j^2}{\sigma_y^2} + 1}\right)$$

2.2 Analyses

Similarly to exercise 1, we pick the priors $\sigma_{y0}^2 = 10, \nu_0 = 1, \omega_0 = 1, \sigma_{0\eta}^2 = 10$ and $\gamma_0^2 = 10$, and proceed with Gibbs sampling:

We read the data

```
library(foreign)
library(hdrcde)
```

```
## Loading required package: mvtnorm
## hdrcde 3.1 loaded
```

```
data <- read.dta(file="arsenicrice2.dta")
Y <- read.dta(file="arsenicrice2.dta")
```

We set prior values

```
n <-  nrow(Y)
nu0 <- 1; omega0 <- 1;
s2_eta0 <- 10; s2_y0 <- 10
mu0 <- mean(Y$arsenic);
g20 <- var(Y$arsenic)
```

We setup the MCMC

```
#Setup starting values
m <- length(unique(Y$food_num)) #number of groups
n <- sv_y <- ybar <- rep(NA,m) #create empty vectors for group descriptions
for (i in 1:m)
{
  n[i] <- sum(Y$food_num==i)
  sv_y[i] <- var(Y$arsenic[which(Y$food_num==i)])
  ybar[i] <-  mean(Y$arsenic[which(Y$food_num==i)])
}
eta <- ybar - mean(Y$arsenic); s2_y <- mean(sv_y)
```

```
mu <- mean(Y$arsenic); s2_eta <- var(eta)
xi <- 0

#Setup MCMC
set.seed(0808)
S <- 10000
THETA <- matrix(nrow=S, ncol=m)
RES <- matrix(nrow=S, ncol=4)
ALL <- matrix(nrow=S, ncol=4+m)

#Setup MCMC
set.seed(0808)
S <- 10000
ETA <- matrix(nrow=S, ncol=m)
RES <- matrix(nrow=S, ncol=4)
ALL <- matrix(nrow=S, ncol=4+m)
```

Our updating functions correspond to the full conditionals derived above:

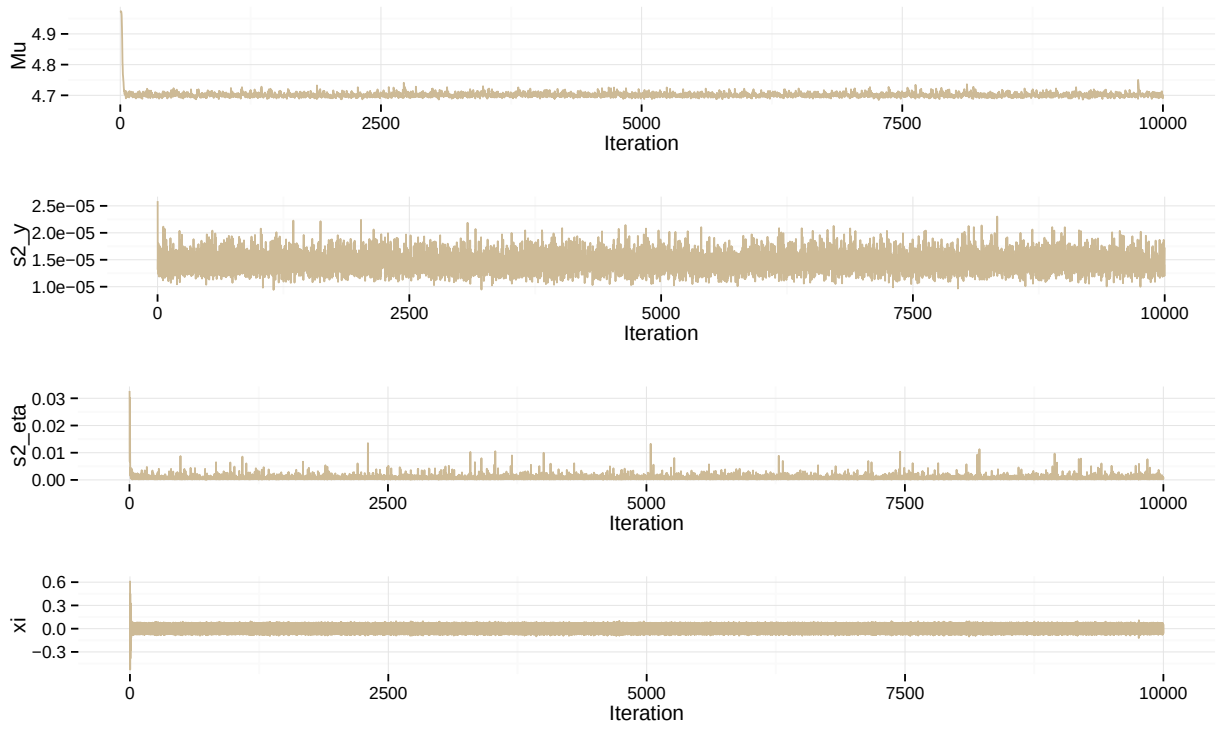
```
#FUNCTIONS
newS2eta <- function(m, s2_eta0, eta, omega0)
{
  ss <- s2_eta0*omega0 + sum( eta^2 )
  s2_eta <- 1/rgamma(1, shape=((omega0 + m)/2), scale=(ss/2))
  return(s2_eta)
}
newS2y <- function(nu0, n, s2_y0, y, m, eta, xi, mu)
{
  nun = nu0 + sum(n)
  ss = sum((y - mu - xi*rep(eta, times=n))^2) + nu0*s2_y0
  s2_y <- 1/rgamma(1, shape=(nun/2), scale=(ss/2))
  return(s2_y)
}
newMu <- function(y, xi, eta, s2_y, g20, mu0, n)
{
  v = ( sum(n)/s2_y + 1/g20 )
  ss = sum(y-rep(eta,times=n)*xi)/s2_y + mu0/g20
  e = (ss/s2_y + mu0/g20)
  mu = rnorm(1, ss/v, sqrt(1/v))
  return(mu)
}
newEta <- function(xi, mu, ybar, n, s2_y, s2_eta)
{
  v = 1/(n*xi^2/s2_y + 1/s2_eta)
  e = v*xi*(n*ybar - mu*n)/s2_y
  eta = rnorm(m, e, sqrt(v))
  return(eta)
}
newXi <- function(eta, m, y, n, mu, s2_y)
{
  v = (sum(n*eta^2)/s2_y + 1)
  e = sum((y-mu)*rep(eta,times=n))/s2_y
  xi = rnorm(1, e/v, sqrt(1/v))
```

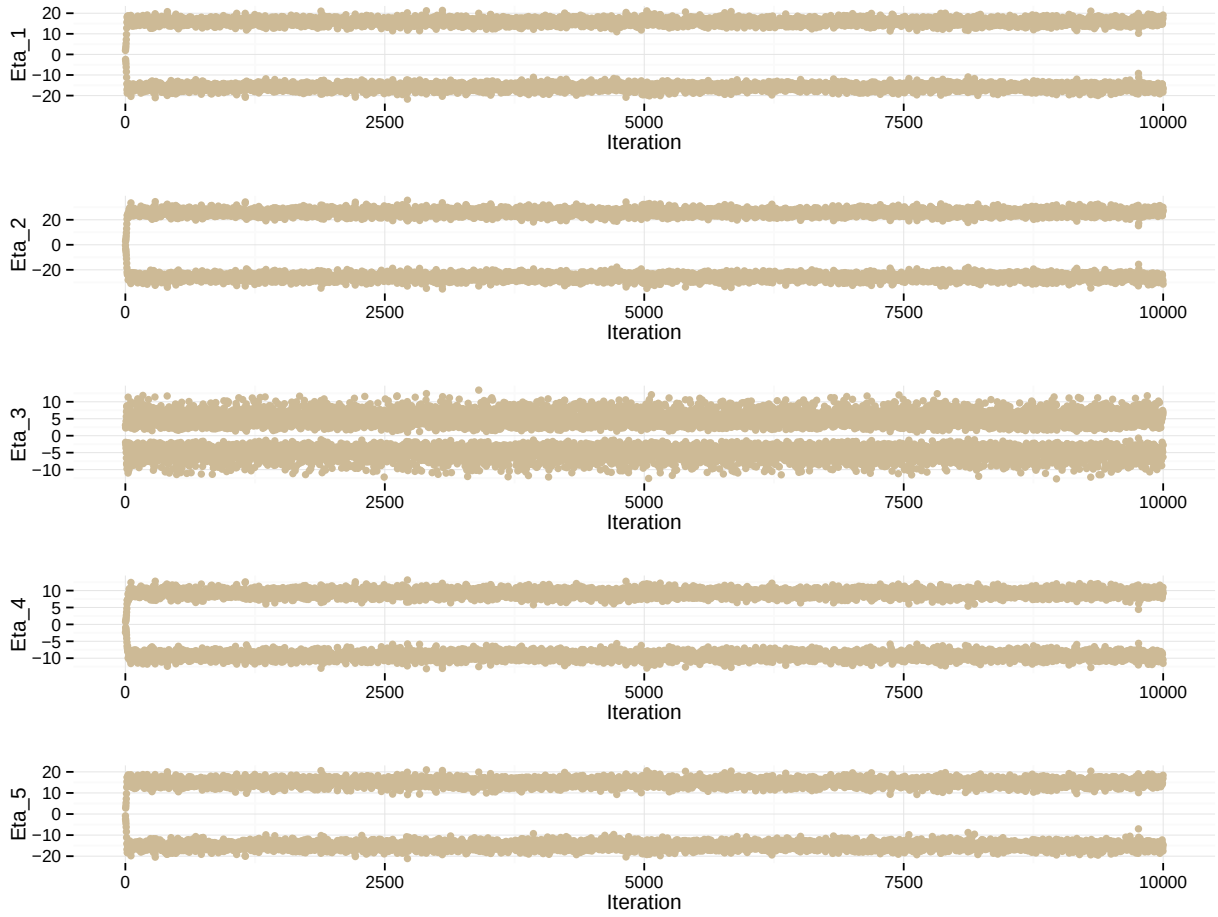
```
  return(xi)
}
```

We run the MCMC algorithm:

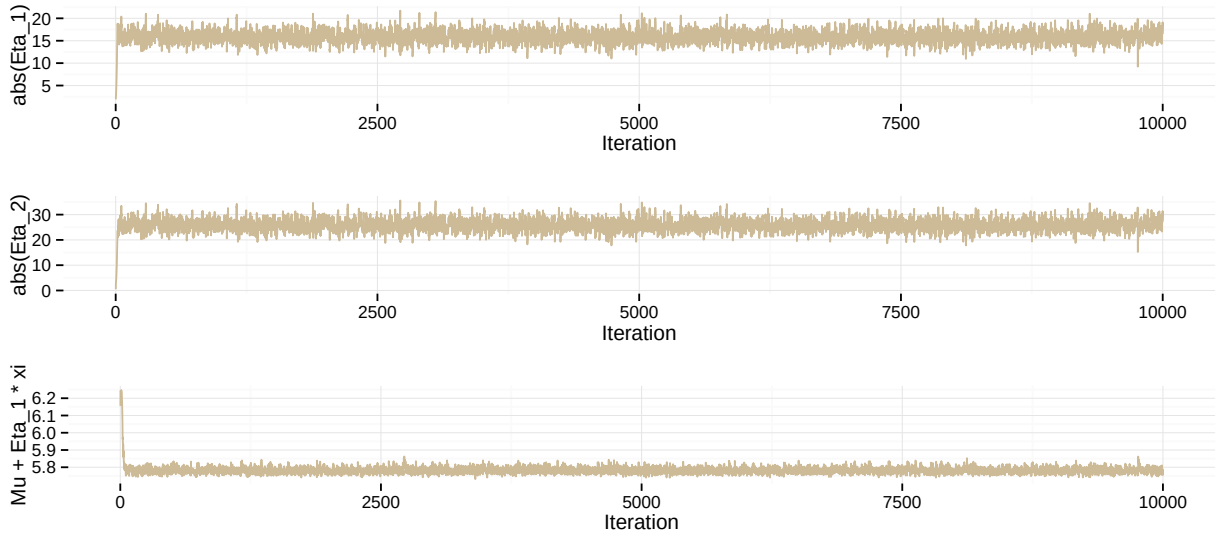
```
#RUN MCMC
for(i in 1:S)
{
  #Get new values for parameters
  eta <- newEta(xi, mu, ybar, n, s2_y, s2_eta)
  mu <- newMu(Y$arsenic, xi, eta, s2_y, g20, mu0, n)

  s2_eta <- newS2eta(m, s2_eta0, eta, omega0)
  s2_y <- newS2y(nu0, n, s2_y0, Y$arsenic, m, eta, xi, mu)

  xi <- newXi(eta, m, Y$arsenic, n, mu, s2_y)
  #Store in chain
  ETA[i,] <- eta
  RES[i,] <- c(mu,s2_y, s2_eta, xi)
  ALL[i,] <- c(eta, mu, xi, s2_eta, s2_y)
}
```

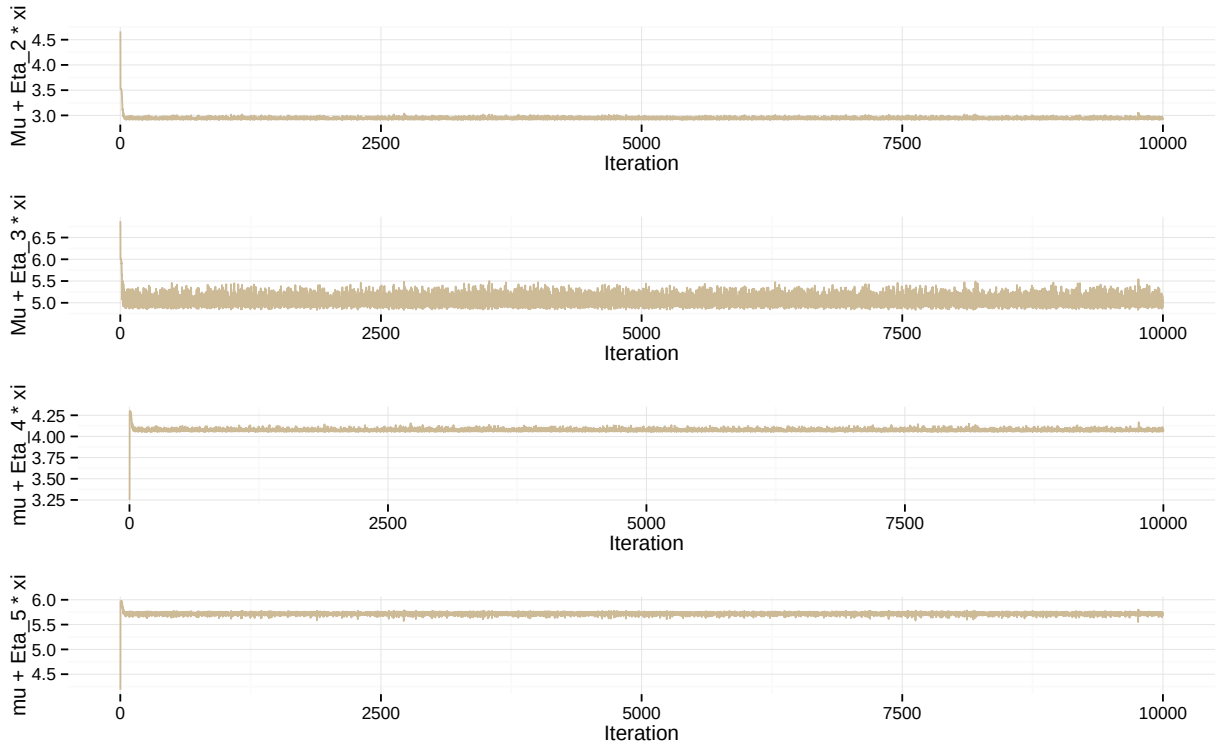
Again, before we get too psyched about results, we check MCMC convergence criteria (code not shown, see above):





Obviously, something is wrong with the η parameters, which could make sense since when the estimate of ξ crosses zero, the η parameters get updated on the other side of 0 as well. Taking the absolute value of ξ reassures us that the parameter space that is actually searched isn't terribad. We also check the convergence of $\mu + \xi\eta$, which is the mean concentration of arsenic in each group, and ultimately interests us. Reproduced below are the graphs for $|\eta_1|, |\eta_2|$, and $\boldsymbol{\theta}$





We are fairly satisfied with the outlook on the convergence of our θ's, although obviously the wacky errors we'll get mean the credible intervals we get for θ's will be meaningless.

Figure 1: I know, Hades isn't pleased either

2.3 Algorithm output

We provide the median estimate and 95% credible interval for the θ parameters in this expanded hierarchical specification model:

Parameter	Credible Lower 95%	Median	Credible Upper 95%
θ_1 (Basmati)	3.610	3.619	3.639
θ_2 (Non-Basmati)	6.420	6.454	6.503
θ_3 (Beverage)	4.072	4.361	4.544
θ_4 (Cakes)	5.308	5.331	5.345
θ_5 (Cereal)	3.649	3.670	3.747

2.4 Sensitivity analyses

We repeat the sensitivity analyses of section 1.4:

1. Large expected μ (Prior expectation of mad levels of arsenic)
2. Large σ^2 and ν_0 (High variability within products)
3. Large τ^2 and η_0 (High variability between products)

Scenario 1:

Parameter	Credible Lower 95%	Median	Credible Upper 95%
θ_1 (Basmati)	3.610	3.619	3.639
θ_2 (Non-Basmati)	6.421	6.454	6.503
θ_3 (Beverage)	4.072	4.361	4.544
θ_4 (Cakes)	5.308	5.332	5.345
θ_5 (Cereal)	3.649	3.670	3.747

Scenario 2:

Parameter	Credible Lower 95%	Median	Credible Upper 95%
θ_1 (Basmati)	3.610	3.619	3.639
θ_2 (Non-Basmati)	6.420	6.454	6.503
θ_3 (Beverage)	4.072	4.361	4.544
θ_4 (Cakes)	5.308	5.331	5.345
θ_5 (Cereal)	3.649	3.670	3.747

Scenario 3:

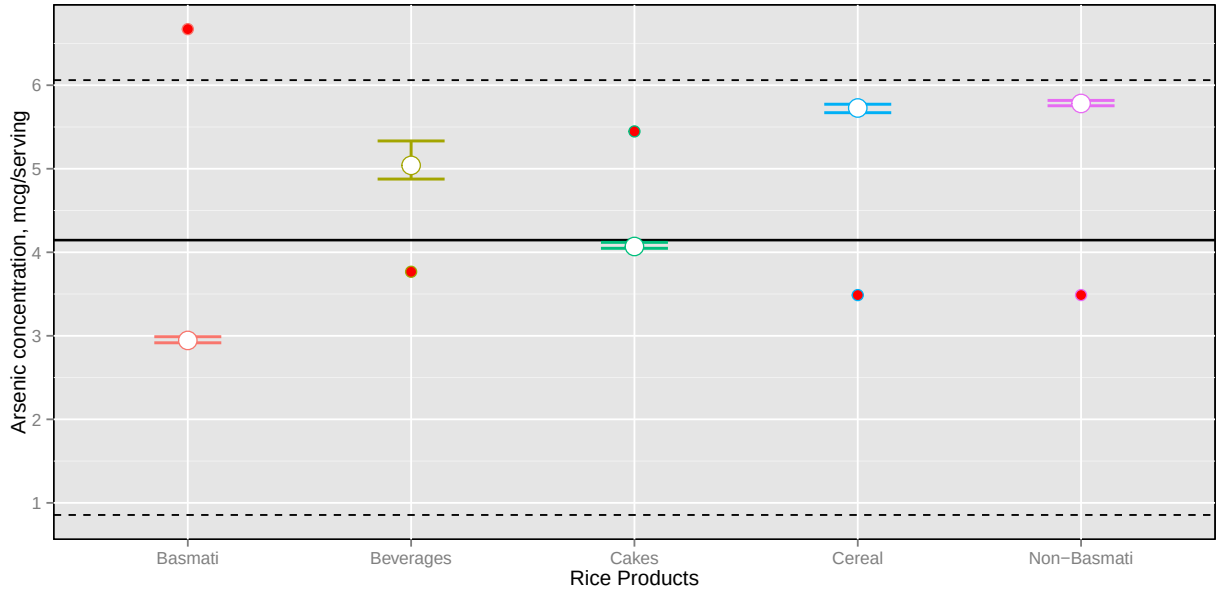
Parameter	Credible Lower 95%	Median	Credible Upper 95%
θ_1 (Basmati)	3.610	3.619	3.639
θ_2 (Non-Basmati)	6.420	6.454	6.503
θ_3 (Beverage)	4.072	4.361	4.544
θ_4 (Cakes)	5.308	5.331	5.345
θ_5 (Cereal)	3.649	3.670	3.747

We observe that outrageous standard errors will prevent any meaningful sensitivity analyses.

2.5 Results presentation

Based on our analyses, and under the assumption we did nothing incorrect, we are nearly positive that the type of rice with the greatest arsenic concentration is non-basmati rice, at 6.420 mcg/serving, followed by cake rice products, at 5.308 mcg/serving. Basmati and cereal products both neared 6.6mcg of arsening per serving, beverage products, which should have constituted an imprecise estimate (but did not), had an estimated arsenic concentration at 6.454 mcg/serving (95% CI=4.361,4.544).

Posterior median estimates and observed mean concentrations of arsenic are presented by product type in the following graph. Markers are θ estimates with 95% credible interval lines; horizontal lines are the median estimate of μ (solid) and corresponding 95% credibal interval (dashed, like my hopes and dreams).



3. Conditionally-conjugate specification of the hierarchical model with group-specific variances

3.1 Model specification

This model is similar to that laid out in section 1.1, with the exceptions that variance is allowed to vary between groups, with group-specific variances following a conjugate inverse-gamma distribution so that (from Hoff):

$$\{\theta_j \mid \sigma^2, y_{j,1}, \ldots, y_{j,n}\} \sim \mathcal{N}\left(\frac{n_j \bar{y}_j/\sigma^2 + \mu/\tau^2}{n_j/\sigma^2 + 1/\tau^2}, \left[n_j/\sigma^2 + 1/\tau^2\right]^{-1}\right)$$

$$\{1/\sigma_j^2 \mid \boldsymbol{\theta}, y_1, \ldots, y_n\} \sim \text{gamma}\left(\frac{1}{2}\left[\nu_0 + n_j\right], \frac{1}{2}\left(\nu_0\sigma_0^2 + \sum_{i=1}^{n_j}(y_{i,j} - \theta_j)^2\right)\right)$$

$$\{\sigma_0^2 \mid \boldsymbol{\sigma}, \nu_0\} \sim \text{gamma}\left(a + \frac{1}{2}m\nu_0, b + \frac{1}{2}\sum_j (1/\sigma_j^2)\right)$$

and

$$\{\nu_0 \,|\boldsymbol{\sigma}\,\} \sim \left(\frac{\left(\nu 0\sigma_0^2/2\right)^{\nu_0/2}}{\Gamma\left(\nu_0/2\right)}\right)^m \left(\prod_j \frac{1}{\sigma_j^2}\right)^{\nu_0/2-1} \times \exp\left\{-\nu_0\left(\alpha + \frac{1}{2}\sigma_0^2 \sum_j \left(\frac{1}{\sigma_j^2}\right)\right)\right\}$$

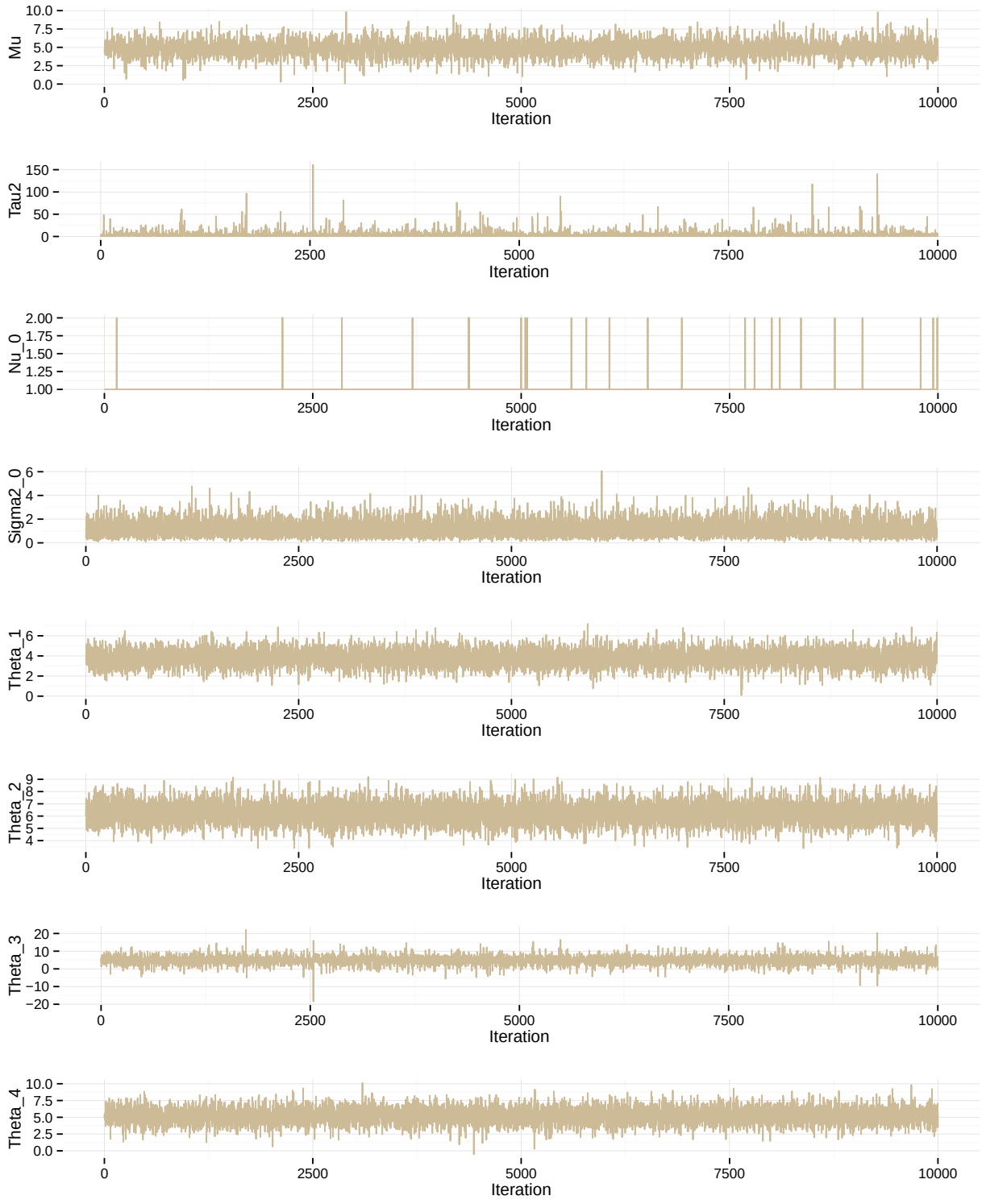
3.2 Analyses

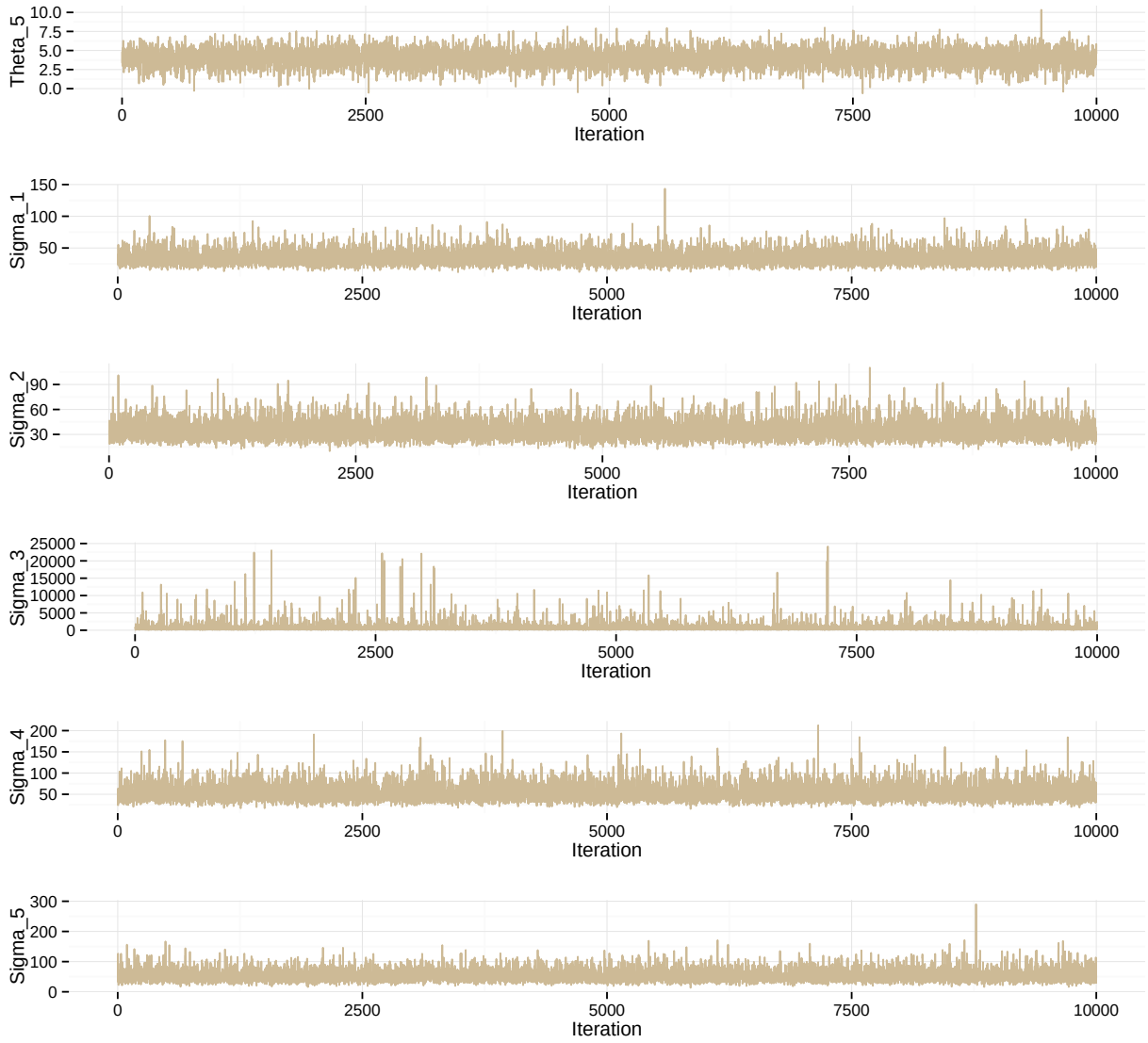
We keep the same priors for hyperparameters pertaining to μ and θ, and additionally specify $\alpha = 1, b = 3, a = 1$ for the prior distributions on ν_0 and σ_0^2. Code to run analyses is as follows, please refer to the earlier sections for added description.

```
data <- read.dta(file="arsenicrice2.dta")
Y <- read.dta(file="arsenicrice2.dta")

#We set weakly informative prior values
n <-  nrow(Y)
eta0 <- 1; t20 <- 3
mu0 <- mean(Y$arsenic)
g20  <- var(Y$arsenic)
a = 1; b = 3; alpha = 1

# mu0 <- 100 #scenario 1
# s20 <- 100*s20; nu0 <- 100 #scenario 2
# t20 <- 100*t20; eta0 <- 100 #scenario 3
#Setup starting values

m <- length(unique(Y$food_num)) #number of groups
n <- s2 <- ybar <- rep(NA,m)
for (i in 1:m)
{
  n[i] <- sum(Y$food_num==i)
  s2[i] <- var(Y$arsenic[which(Y$food_num==i)])
  ybar[i] <- mean(Y$arsenic[which(Y$food_num==i)])
}
theta <- ybar; s2_0 <- mean(sv)
mu <- mean(theta); tau2 <- var(theta)
nu0 <- 1

#Setup MCMC
set.seed(0808)
S <- 10000
THETA2 <- S2 <- matrix(nrow=S, ncol=m)
RES <- matrix(nrow=S, ncol=4)
ALL <- matrix(nrow=S, ncol=4+2*m)
NUMAX <- 10000

### Functions sampling from posteriors
newTheta <- function(n, ybar, s2, tau2, mu)
{
  v = 1/(n/s2 +1/tau2)
  e = v * (ybar*n/s2 + mu/tau2)
  new <- rnorm(length(n), e, sqrt(v))
```

```
  return(new)
}
newSigma2 <- function(m, n, nu0, s20, theta, y)
{
  nun = n + nu0
  ss <- nu0 * s20 + sum((y - rep(theta, times=n))^2)
  sigma2 <- 1/rgamma(m, nun/2, ss/2)
  return(sigma2)
}
newMu <- function(m, theta, tau2, g20)
{
  v = 1/(m/tau2 + 1/g20)
  e = v *(m*mean(theta)/tau2 + mu0/g20)
  mu <- rnorm(1, e, sqrt(v))
  return(mu)
}
newTau2 <- function(m, eta0, t20, theta, mu)
{
  etam = eta0 + m
  ss = eta0*t20 + sum( (theta-mu) ^2 )
  tau2 <- 1/rgamma(1, etam/2, ss/2)
  return(tau2)
}
newS20 <- function(m, nu0, s2, a, b)
{
  L = a + 0.5*m*nu0
  R = b + 0.5*sum(1/s2^2)
  s20 <- rgamma(1, L, R)
}

### Run MCMC
for(i in 1:S)
{
  #Get new values for parameters
  theta <- newTheta(n, ybar, s2, tau2, mu)
  s2 <- newSigma2(m, n, nu0, s20, theta, Y$arsenic)
  mu <- newMu(m, theta, tau2, g20)
  tau2 <- newTau2(m, eta0, t20, theta, mu)
  s20 <- newS20(m, nu0, s2, a, b)

  x <- 1:NUMAX
    lpnu0<-m*(.5*x*log(s20*x/2)-lgamma(x/2))+
      (x/2-1)*sum(log(1/s2))+
      -x*(alpha+.5*s20*sum(1/s2))
  nu0<-sample(x,1,prob=exp(lpnu0-max(lpnu0)))


  #Store in chain
  THETA2[i,] <- theta
  S2[i,] <- s2
  RES[i,] <- c(mu,tau2, nu0, s20)
  ALL[i,] <- c(theta,s2,mu,tau2, nu0, s20)
}
```

Trace plots for the parameters follow:





Other than σ_3^2 and ν_0, everything looks happy and converged. σ_3 probably has a large range due to the small number of non-missing data on rice beverages.

3.3 Algorithm output

Estimates for parameter medians and their 95% credible intervals are shown in the following table:

Figure 2: Other than those two, we're good

Parameter	Credible Lower 95%	Median	Credible Upper 95%
θ_1 (Basmati)	2.237	3.826	5.361
θ_2 (Non-Basmati)	4.713	6.240	7.868
θ_3 (Beverage)	0.394	4.841	9.097
θ_4 (Cakes)	3.108	5.200	7.355
θ_5 (Cereal)	1.827	4.042	6.122
σ_1 (Basmati)	18.545	33.132	61.644
σ_2 (Non-Basmati)	18.650	33.084	60.759
σ_3 (Beverage)	131.361	489.618	3600.339
σ_4 (Cakes)	27.028	50.380	100.924
σ_5 (Cereal)	26.807	50.369	101.083
μ (Basmati)	2.926	4.841	6.846
τ_2 (Non-Basmati)	0.573	2.448	16.913
ν_0 (Beverage)	1	1	1
$\sigma 2_0$ (Beverage)	0.277	1.060	2.712

3.4 Sensitivity analyses

We repeat the sensitivity analyses of section 1.4:

1. Large expected μ (Prior expectation of mad levels of arsenic)
2. Large b and α (High variability within products)
3. Large τ^2 and η_0 (High variability between products)

Scenario 1:

Parameter	Credible Lower 95%	Median	Credible Upper 95%	Big change?
θ_1 (Basmati)	0.913	3.505	6.207	Yes
θ_2 (Non-Basmati)	4.062	6.686	9.437	Yes
θ_3 (Beverage)	-32.754	8.074	71.878	Yes
θ_4 (Cakes)	1.528	5.485	9.821	Yes
θ_5 (Cereal)	-0.448	3.504	7.611	Yes
σ_1 (Basmati)	24.947	58.269	404.571	Yes
σ_2 (Non-Basmati)	25.104	57.722	404.574	Yes
σ_3 (Beverage)	194.759	931.623	12209.972	Yes
σ_4 (Cakes)	36.606	89.011	626.494	Yes
σ_5 (Cereal)	35.694	89.044	628.468	Yes
μ (Basmati)	95.120	99.723	104.288	Yes
τ_2 (Non-Basmati)	3033.227	8299.341	38038.052	Yes!!
ν_0 (Beverage)	1	1	1	No
$\sigma 2_0$ (Beverage)	0.277	1.060	2.712	Yes

Scenario 2:

Parameter	Credible Lower 95%	Median	Credible Upper 95%	Big change?
θ_1 (Basmati)	1.408	3.479	5.583	Meh
θ_2 (Non-Basmati)	4.618	6.665	8.690	Meh
θ_3 (Beverage)	-19.383	4.385	27.663	Meh
θ_4 (Cakes)	2.331	5.453	8.664	Meh
θ_5 (Cereal)	0.273	3.465	6.660	Yes
σ_1 (Basmati)	23.249	47.513	107.471	Meh
σ_2 (Non-Basmati)	23.565	47.531	106.423	Meh
σ_3 (Beverage)	173.775	706.454	5455.201	Meh
σ_4 (Cakes)	34.267	72.325	173.977	Meh
σ_5 (Cereal)	33.941	72.486	174.818	Meh
μ (Basmati)	0.456	4.872	9.262	Yes
τ_2 (Non-Basmati)	222.934	288.999	384.448	Yes!!
ν_0 (Beverage)	1	1	1	No
$\sigma 2_0$ (Beverage)	0.279	1.058	2.712	Yes

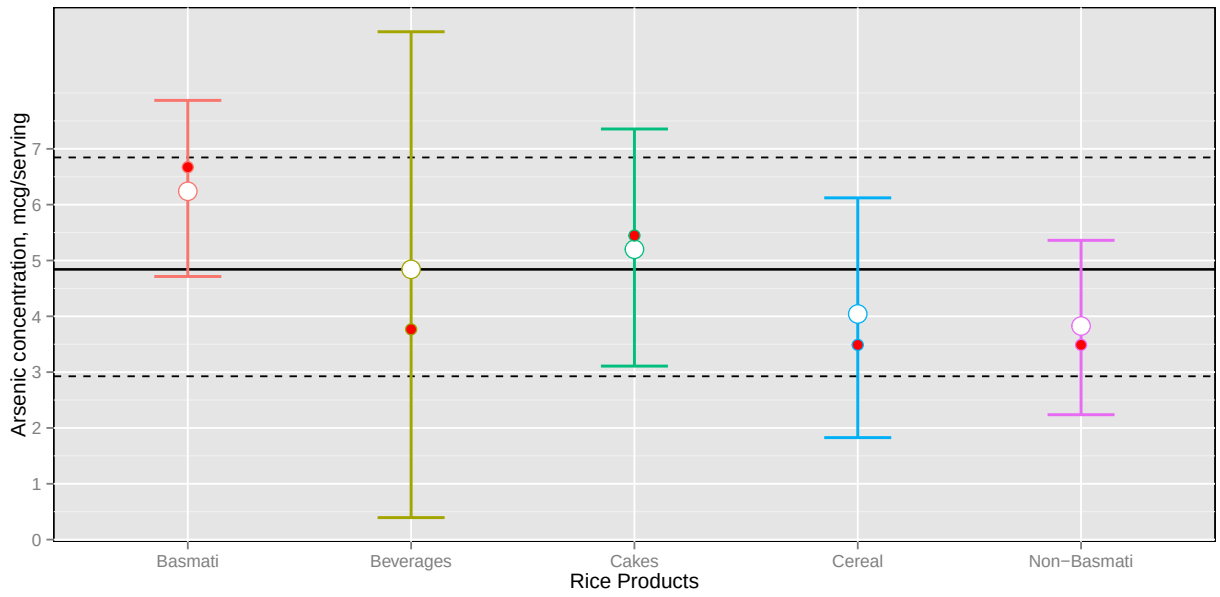
Scenario 3:

Parameter	Credible Lower 95%	Median	Credible Upper 95%	Big change?
θ_1 (Basmati)	1.408	3.479	5.583	No
θ_2 (Non-Basmati)	4.674	6.234	7.817	No
θ_3 (Beverage)	0.490	4.814	9.144	No
θ_4 (Cakes)	3.126	5.194	7.311	No
θ_5 (Cereal)	1.864	4.039	6.094	No
σ_1 (Basmati)	18.609	33.008	60.612	No
σ_2 (Non-Basmati)	18.807	33.083	61.010	No
σ_3 (Beverage)	131.637	487.250	3367.378	No
σ_4 (Cakes)	26.960	50.231	98.334	No
σ_5 (Cereal)	26.904	50.614	100.647	No
μ (Basmati)	2.926	4.828	6.741	No
τ_2 (Non-Basmati)	0.584	2.448	16.443	No
ν_0 (Beverage)	1	1	1	No
$\sigma 2_0$ (Beverage)	0.121	0.163	0.214	Yes

Results from these sensitivity analyses are analogous to those from section 1.4; we observe that increasing μ_0 increases shrinking towards μ and the error on θ's; increasing a and b increases the error on θ's and the

estimates of σ's; increasing τ_0^2 and η_0 enlarges the credible interval around estimates of σ's, along with, obviously, τ_0^2 itself.

3.5 Presentation of results



4. Discussion of shrinkage

Ignoring our failed attempt at the parameter-expanded model, we observed greater shrinkage when modelling group-specific variances than when setting them all equal, as the greater uncertainty on the models pulled them closer to the prior population mean.

5. Discussion of imputation

Replacing data with mean by-group observations would simply drive down the estimates on the errors around θ's, while reducing shrinkage, driving θ's towards $\bar{y}$. This follows from $\bar{y}$ remaining unchanged, but n_j increasing proportional to the amount of imputed data.

6. Model comparisons

Again, we must ignore the parameter-expanded model from these comparisons. Model 1 and 3 lead to similar inferences, although Model 3 has much greater uncertainty around the median estimates of θ's. Both models clearly suggest that Basmati rice products ought to receive special scrutiny when evaluating their arsenic contents, whereas urgent attention seems less warranted with non-basmati and ceraeal products. There lacks data to convincingly determine whether rice beverage products stand exactly with relation to others.