Homework #3

Analysis of Arsenic in Rice Products

Antoine Baldassari

November 17, 2015

1. Shared variance across groups

1.1 Definitions and derivations

Let the within- and between- group sampling models be normally-distributed with:

$$\phi_j = \{y|\phi_j\}\,,\ p(y|\phi_j) = \text{normal}(\theta_j, \sigma^2) \ \ \text{(within group)}$$
$$\psi = \left\{\mu, \tau^2\right\}\,,\ p(\theta_j|\psi) = \text{normal}(\mu, \tau^2) \ \ \text{(between-group)}$$

In this conditionally-conjugate specification:

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

The full conditional distribution of the parameters can be found to be:

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

1.2 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.

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

We set the weakly informative prior values

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

We set initial values for algorithm

```r
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.

```r
#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)
}
```

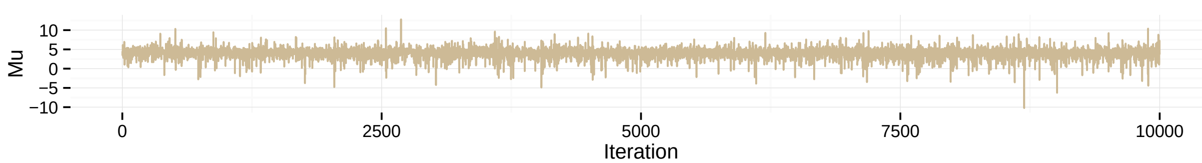
Where the functions updating the parameters follow the equations listed above:

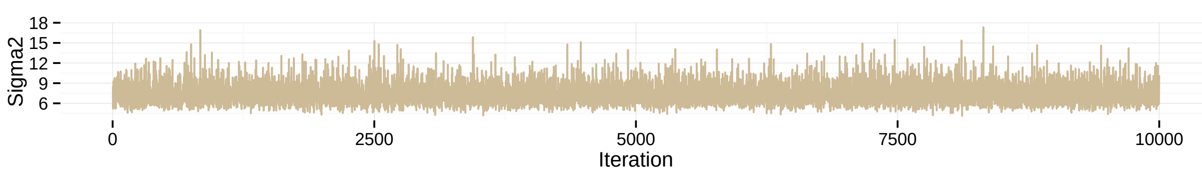
```
newTheta <- function(n, ybar, s2, tau2, mu)
{
  v = 1/(n/s2 +1/tau2)
  e = v * (ybar*n/s2 +mu/tau2)
  new <- rnorm(1, e, sqrt(v))
  return(new)
}
newSigma2 <- function(m, n, nu0, s20, theta, Y)
{
  nun = nu0 + sum(n)
  ss <- nu0 * s20
  for(i in 1:m) ss = ss+sum((Y$arsenic[which(Y$food_num==i)] - 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)
}
```

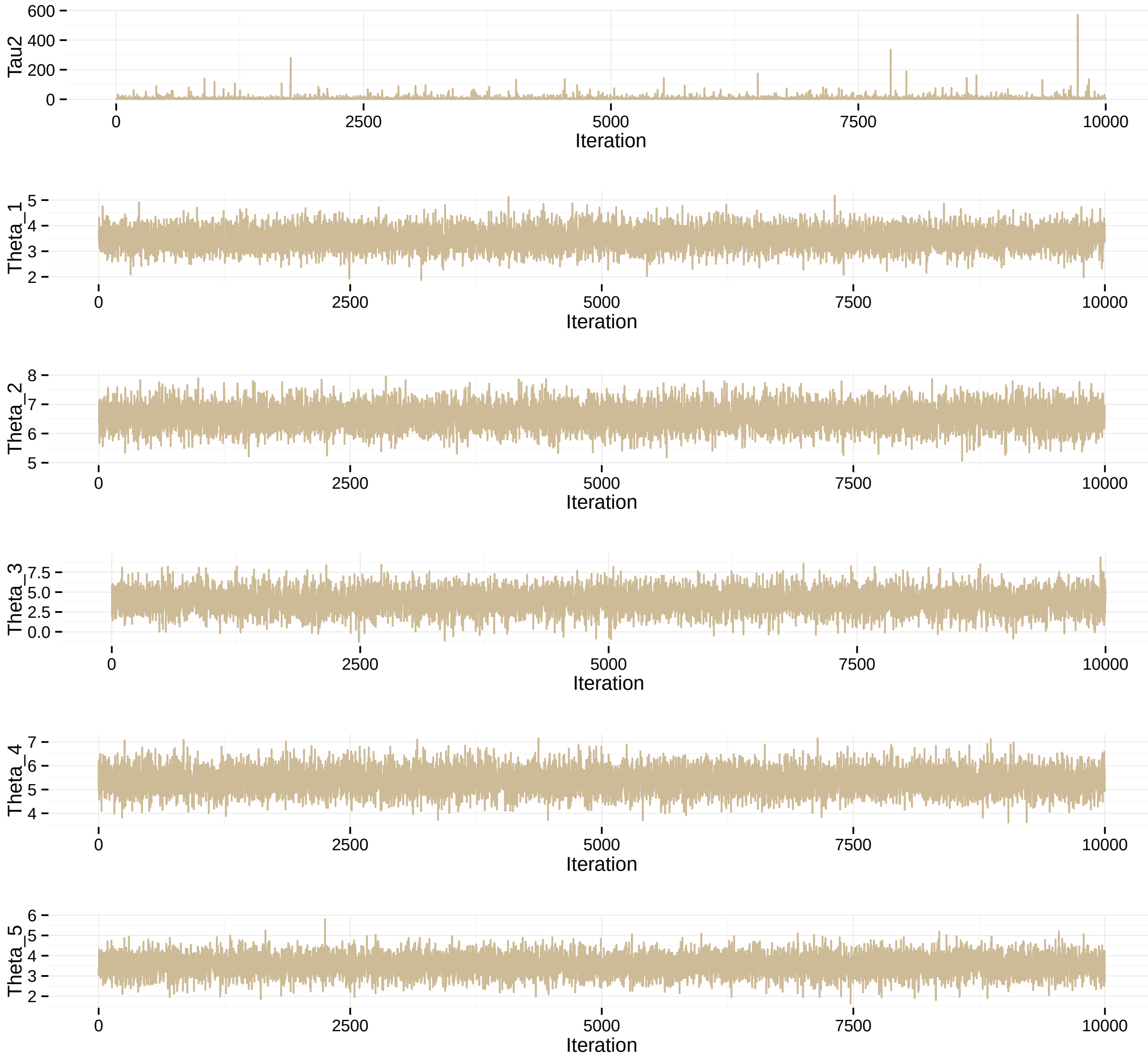
Before we go any further, we check that the MCMC model converged for all four statistics using ggplot2 (code used for μ repeated for other parameters):

```
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 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).

```
library(plotrix)
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,])
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)
g
```