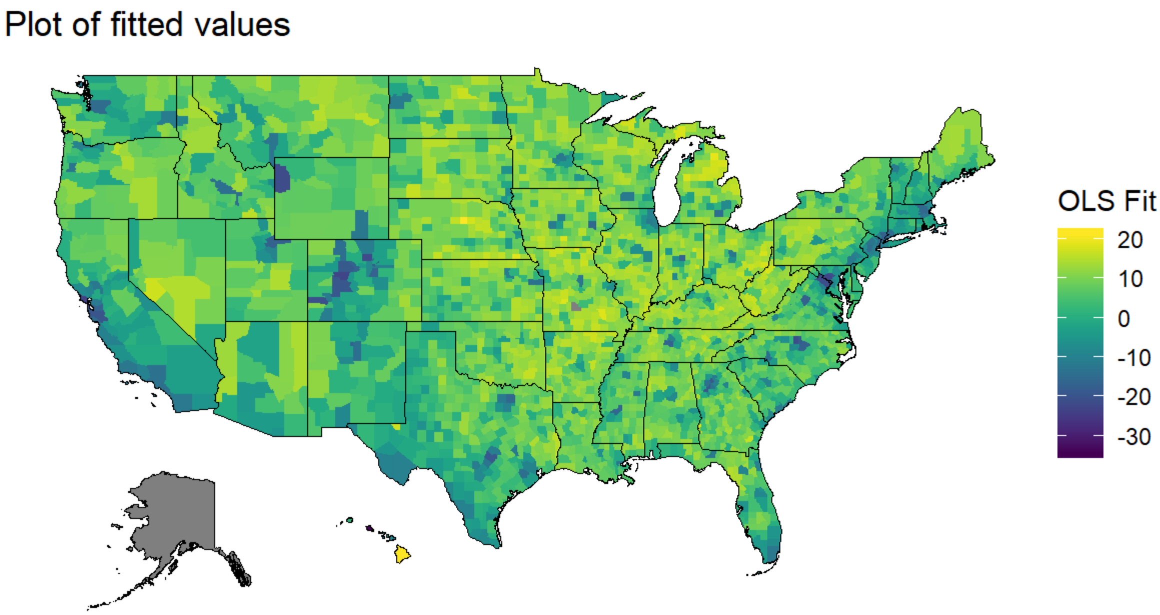
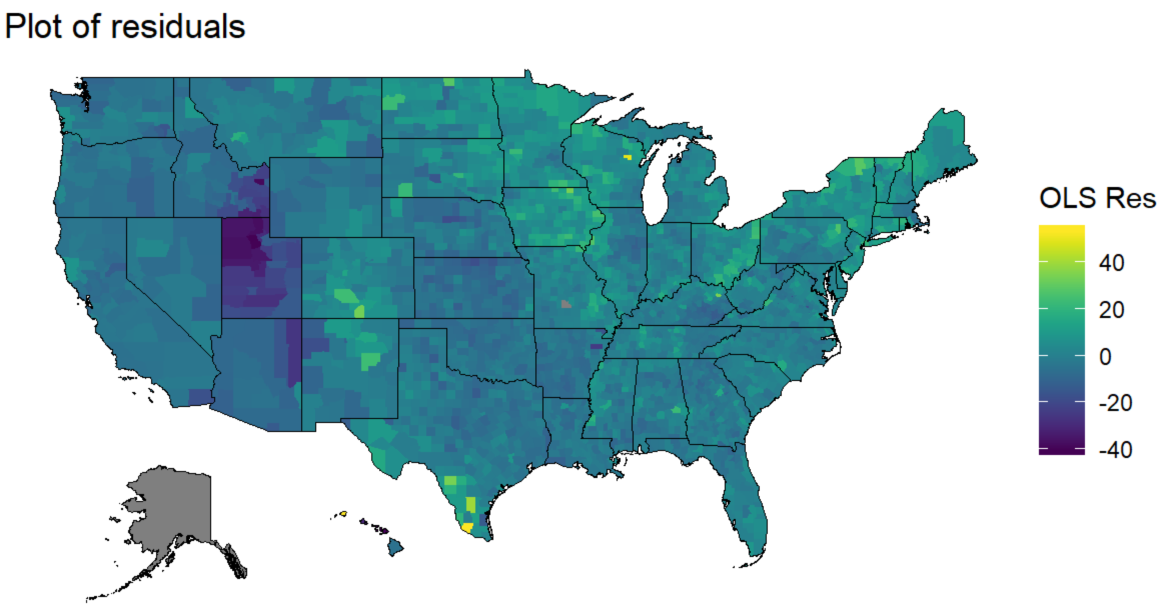
ST 433/533: Homework 5 Solutions

Non-spatial linear regression





CAR model

Loading [MathJax]/jax/output/HTML-CSS/jax.js

Model description

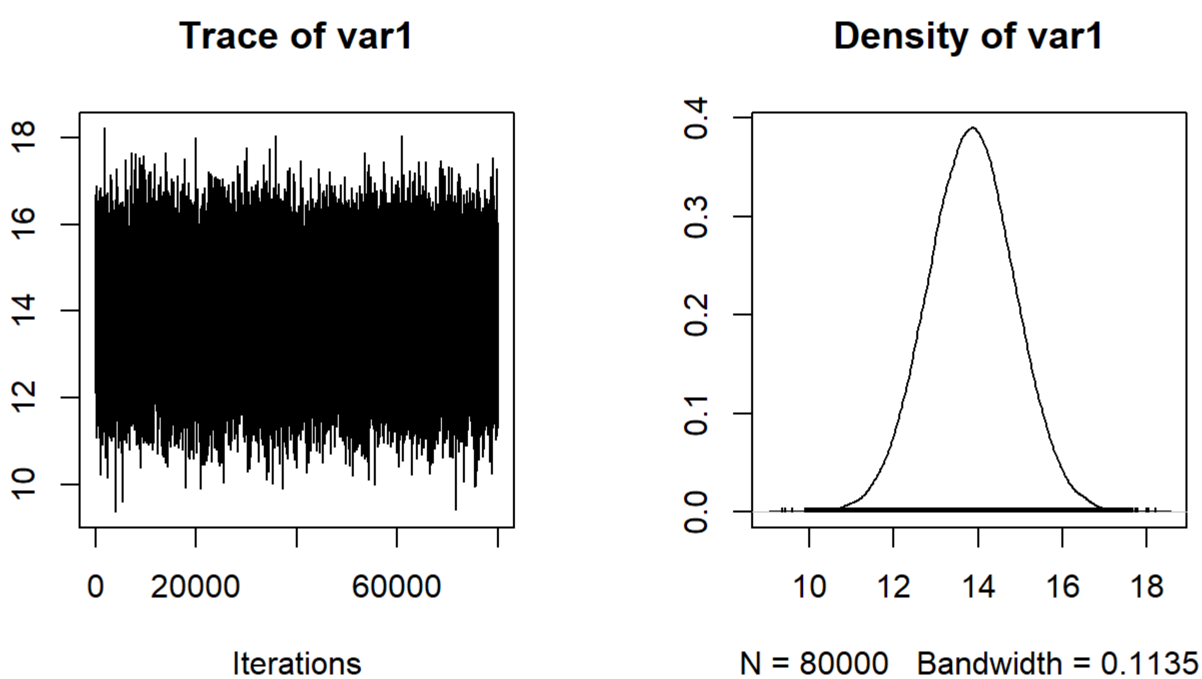
We are fitting the model $Y_i = X_i\beta + Z_i + \epsilon_i\ ,\ i \quad \{1, 2, ..n\}$. The uncorrelated nugget error is $\epsilon_i \quad \text{Normal}(0, \tau^2)$. The spatial term is $Z = (Z_1, Z_2, ... Z_n)^T \quad \text{Normal}(0, \Sigma)$.

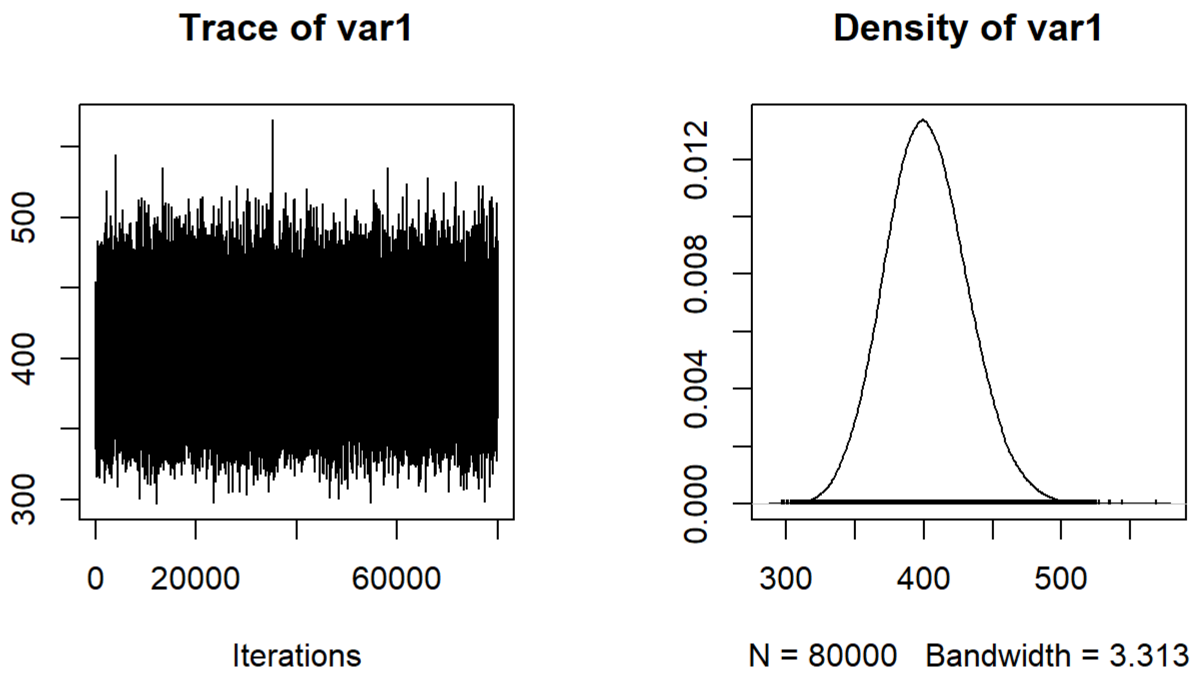
$Z_i | Z_{-i} \quad \text{Normal}(\rho \bar{Z}_i, \sigma^2/m_i)$, where Z_{-i} is the collection of the $n-1$ other spatial terms and $\bar{Z}_i$ is the mean of Z_j over the m_i regions that neighbor region i. $\rho \quad (0, 1)$ determines the strength of spatial correlation. σ^2 is the variance parameter. Using the Leroux parameterization, we have $\Sigma = \sigma^2[(1-\rho)I_n + \rho(M-W)]^{-1}$, where M is a diagonal matrix with $m_1, m_2, ..m_n$ in the diagonal and W is the adjacency matrix.

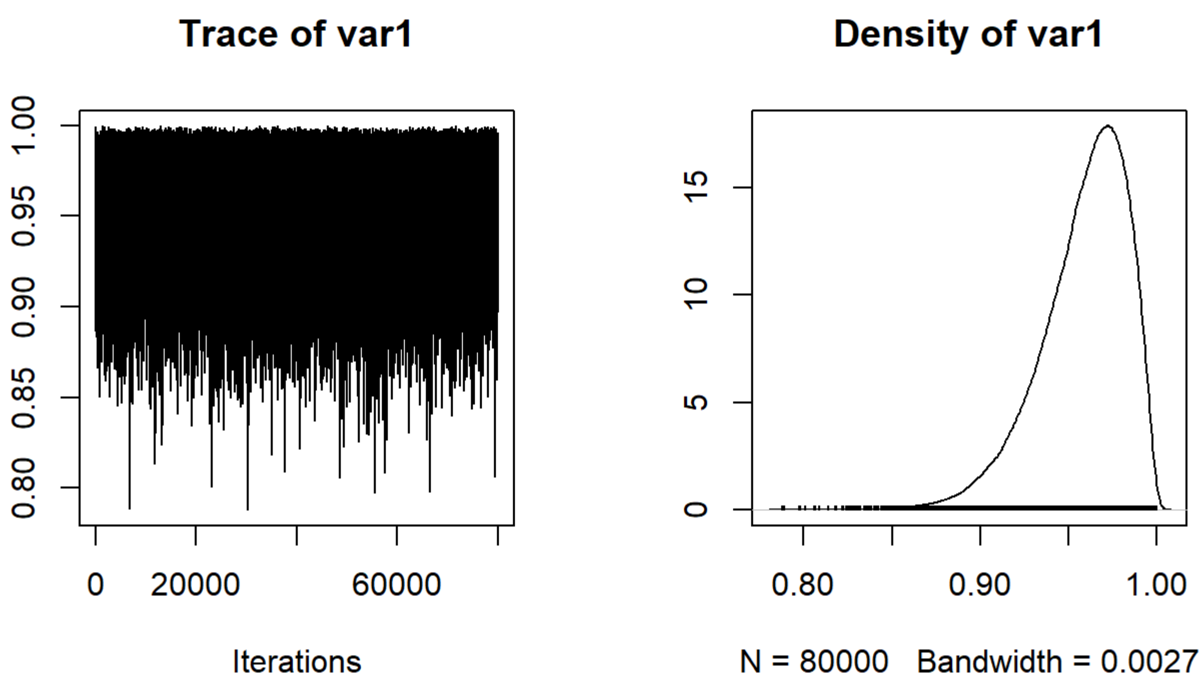
Y_i is the response and X_i is the set of covariates for county i. The estimates of β's $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{10}$ are interpreted as the change in response Y_i with a unit change in the corresponding covariate, keeping the effect of other covariates fixed ($\hat{\beta}_0$ gives the estimate of Y_i when other covariates are unaccounted for).

MCMC convergence

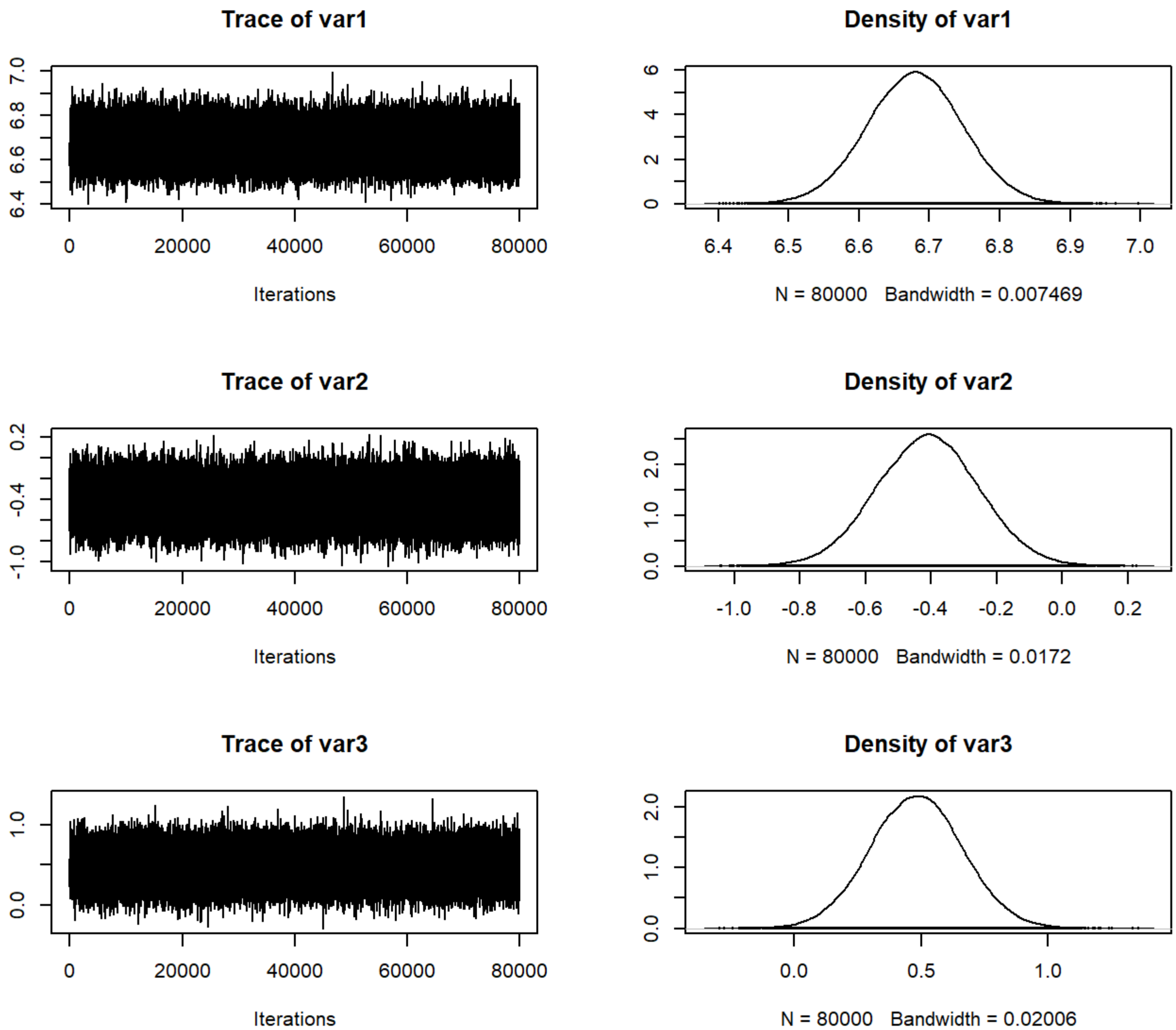
For σ^2, τ^2 and ρ:

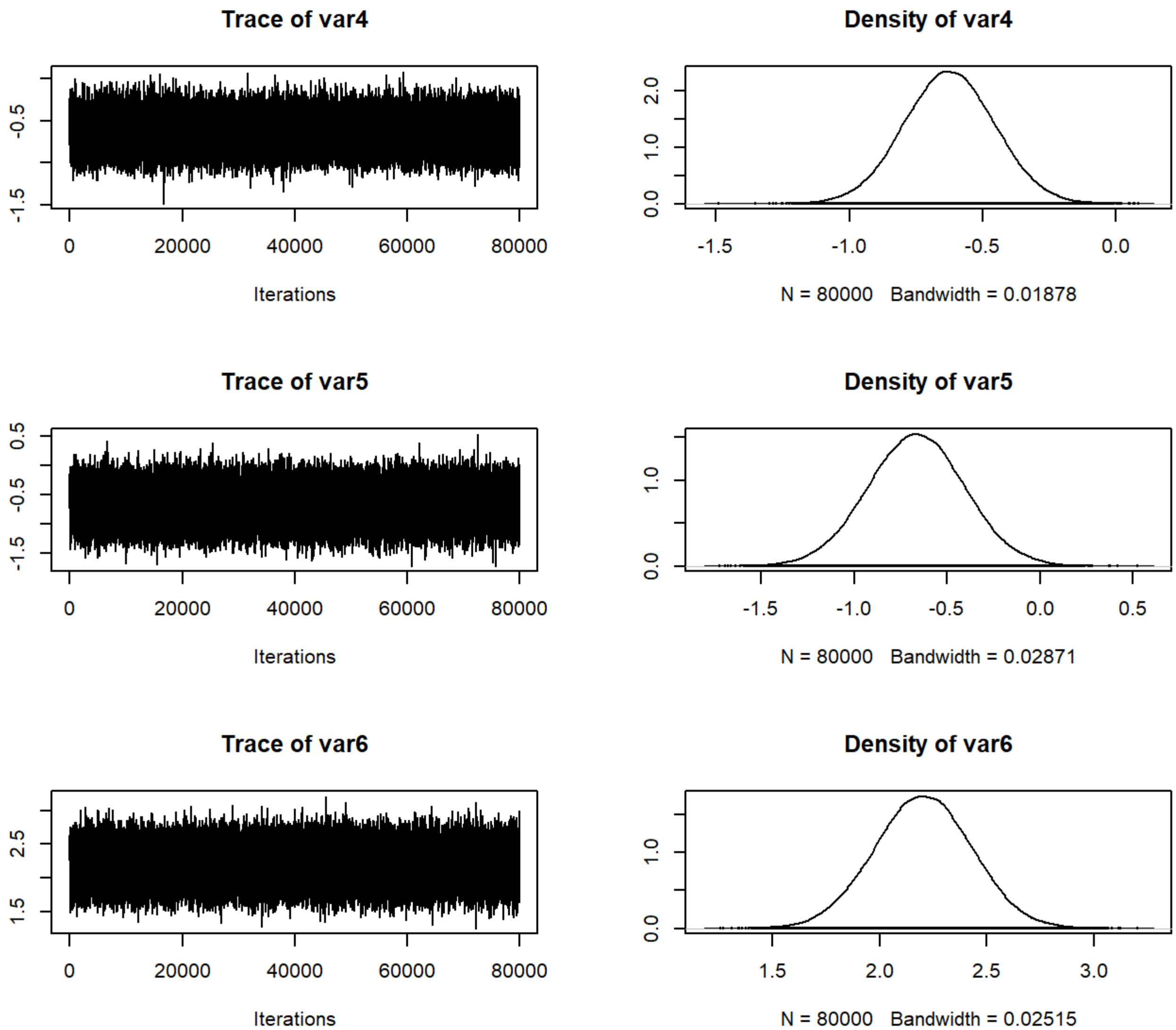


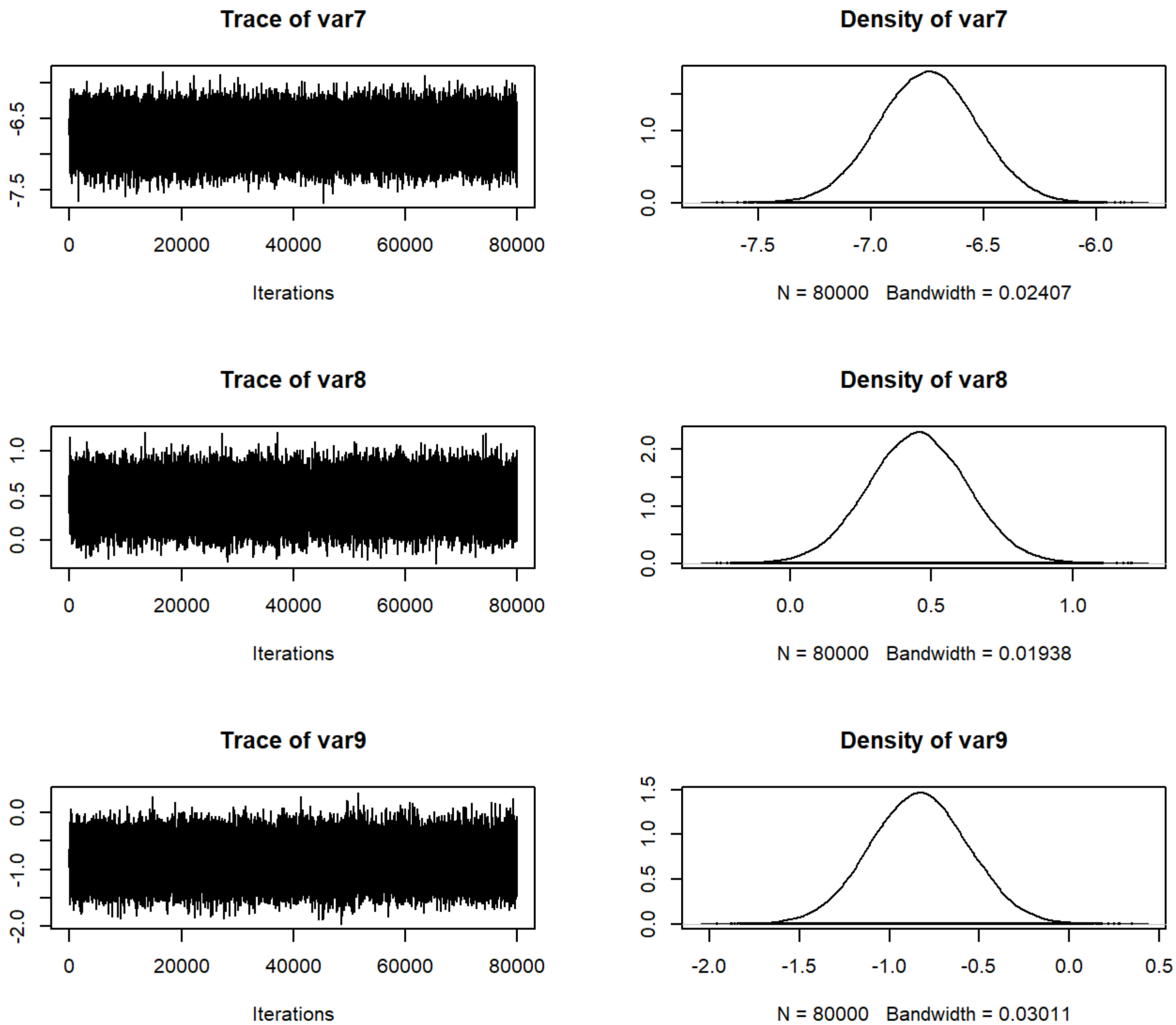


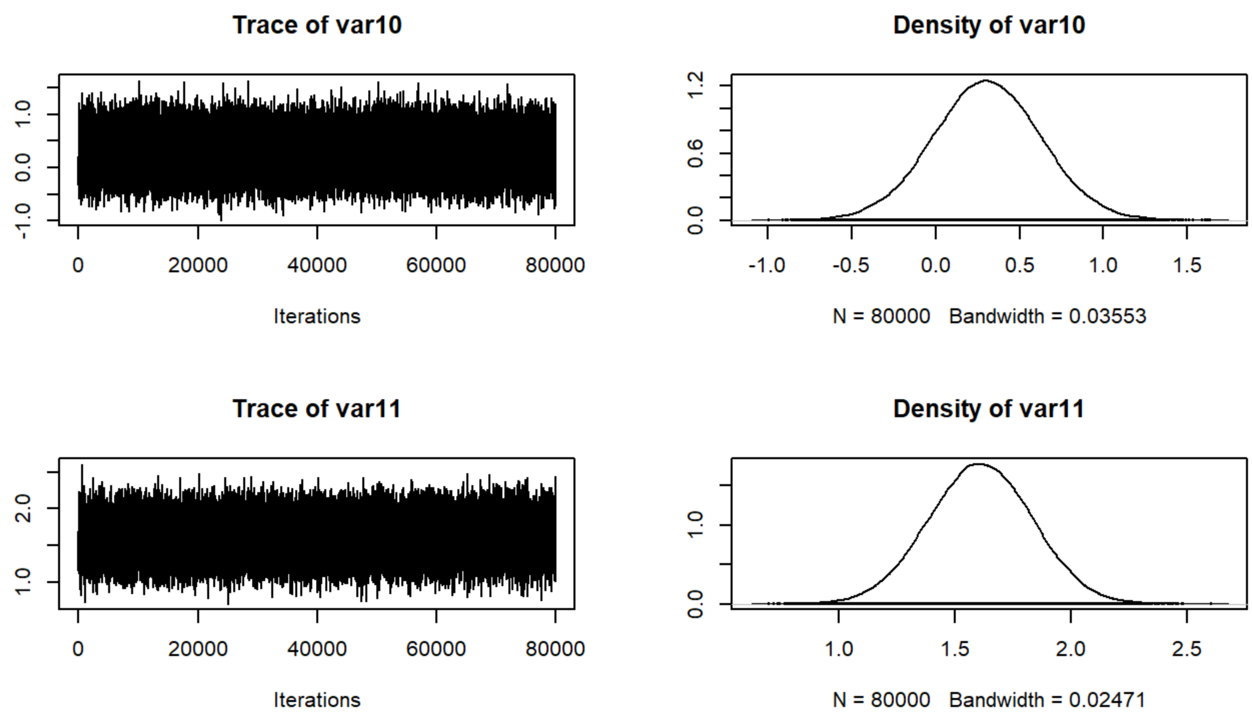


For $\beta_0, \beta_1, \ldots, \beta_{10}$:


Trace of var1
Iterations
Density of var1
N = 80000 Bandwidth = 0.007469
Trace of var2
Iterations
Density of var2
N = 80000 Bandwidth = 0.0172
Trace of var3
Iterations
Density of var3
N = 80000 Bandwidth = 0.02006


Trace of var4
Iterations
N = 80000 Bandwidth = 0.01878
Density of var4
Trace of var5
Iterations
Density of var5
N = 80000 Bandwidth = 0.02871
Trace of var6
Iterations
Density of var6
N = 80000 Bandwidth = 0.02515


Trace of var7
-6.5
-7.5
0
20000
40000
60000
80000
Iterations
Density of var7
1.0
0.0
-7.5
-7.0
-6.5
-6.0
N = 80000   Bandwidth = 0.02407
Trace of var8
1.0
0.5
0.0
0
20000
40000
60000
80000
Iterations
Density of var8
2.0
1.0
0.0
0.0
0.5
1.0
N = 80000   Bandwidth = 0.01938
Trace of var9
0.0
-1.0
-2.0
0
20000
40000
60000
80000
Iterations
Density of var9
1.5
1.0
0.5
0.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
N = 80000   Bandwidth = 0.03011



For all the parameters, the MCMC has converged well.

Comparison

Non-spatial:

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.268  -4.292  -0.703   3.870  52.981
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.088e+01  4.161e+00  -2.614  0.00898 **
## XPST120214  -2.693e-01  3.920e-02  -6.869 7.78e-12 ***
## XAGE775214   2.066e-01  4.508e-02   4.583 4.76e-06 ***
```

```
## XRHI225214  -1.095e-01  1.161e-02  -9.434  < 2e-16 ***
## XRHI725214  -1.526e-01  1.283e-02 -11.890  < 2e-16 ***
## XEDU635213   2.606e-01  3.697e-02   7.051 2.19e-12 ***
## XEDU685213  -7.169e-01  3.000e-02 -23.895  < 2e-16 ***
## XHSG445213  -1.572e-04  2.509e-02  -0.006  0.99500    
## XHSG495213  -1.748e-05  3.007e-06  -5.815 6.70e-09 ***
## XINC110213   1.541e-04  3.166e-05   4.866 1.19e-06 ***
## XPVY020213   2.275e-01  4.416e-02   5.151 2.75e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.482 on 3100 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.4929, Adjusted R-squared:  0.4912 
## F-statistic: 301.3 on 10 and 3100 DF,  p-value: < 2.2e-16
```

Spatial:

```
##                Median     2.5%    97.5% n.sample % accept n.effective
## (Intercept)   6.6794   6.5461   6.8113    80000    100.0     80000.0
## XPST120214   -0.4098  -0.7141  -0.1062    80000    100.0     66283.7
## XAGE775214    0.4813   0.1263   0.8373    80000    100.0     65951.3
## XRHI225214   -0.6262  -0.9573  -0.2931    80000    100.0     52149.4
## XRHI725214   -0.6657  -1.1763  -0.1567    80000    100.0     28365.6
## XEDU635213    2.2024   1.7563   2.6444    80000    100.0     63395.3
## XEDU685213   -6.7484  -7.1715  -6.3239    80000    100.0     73890.3
## XHSG445213    0.4525   0.1076   0.7944    80000    100.0     72957.7
## XHSG495213   -0.8339  -1.3657  -0.3025    80000    100.0     49195.7
## XINC110213    0.3043  -0.3287   0.9364    80000    100.0     74666.8
## XPVY020213    1.6112   1.1732   2.0443    80000    100.0     78033.6
## nu2          13.8452  11.8403  15.8597    80000    100.0     28176.6
## tau2        401.1524 346.1194 463.4129    80000    100.0     29635.9
## rho           0.9628   0.8993   0.9935    80000     45.2     63094.4
##             Geweke.diag
## (Intercept)        -2.0
## XPST120214          0.2
## XAGE775214         -0.6
## XRHI225214          0.4
## XRHI725214          0.3
## XEDU635213         -0.5
## XEDU685213         -1.4
## XHSG445213         -1.6
## XHSG495213          1.9
## XINC110213         -0.8
## XPVY020213         -1.7
## nu2                 1.5
## tau2               -1.4
## rho                 1.2
```

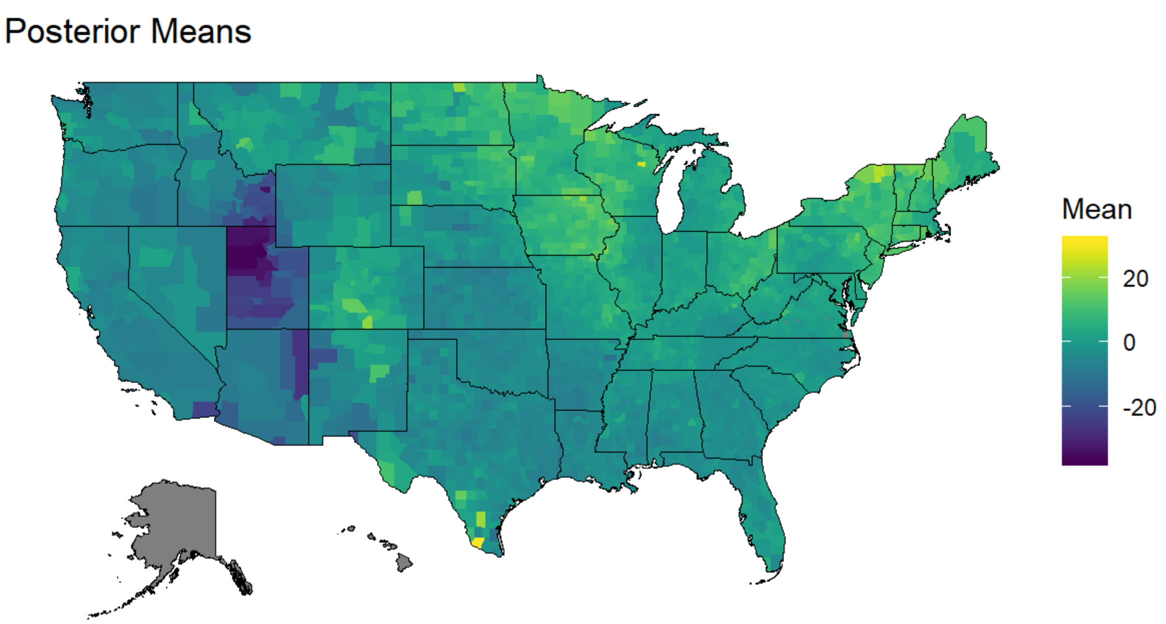
The posterior means and sd for the coefficient estimates:

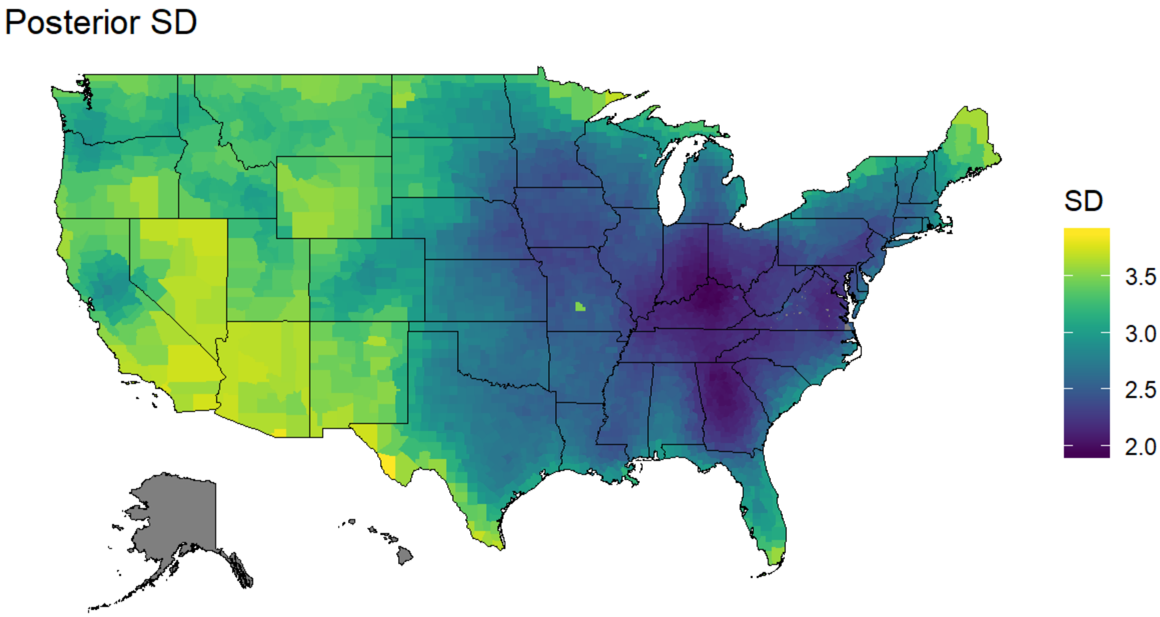
```
##       Post.Mean Post.SD
```

```
## var1      6.6791  0.0674
## var2     -0.4102  0.1552
## var3      0.4810  0.1810
## var4     -0.6256  0.1694
## var5     -0.6665  0.2590
## var6      2.2020  0.2269
## var7     -6.7486  0.2172
## var8      0.4520  0.1749
## var9     -0.8339  0.2716
## var10     0.3051  0.3208
## var11     1.6109  0.2230
```

There is some difference when we compare the estimates from non-spatial and spatial model fits. The standard errors are somewhat larger for spatial model.

Spatial random effects





From the plots, the random spatial effect estimates are higher along the US-Canada border, a few parts of central US and Texas and a little part on the upper part north eastern border. Estimates have low values in and around Utah. The standard deviations are high on the south western part, in some parts of the lower boundary of Texas. Overall, there is lower sd in the Eastern part. A missing covariate could be: Employment rate.

Codes

```
library(tidyverse)
library(housingData)
library(fields)
library(readr)
library(parallel)
library(choroplethr)
library(viridis)
library(CARBayes)


load("election_2008_2016.RData")
county_adj2010 <- read_csv("county_adjacency2010.csv")

#Removing missing values
miss <- is.na(Y)
Y1 <- Y[!miss]
X1<- X[!miss,]
fips1 <- fips[!miss]
all_dat1 <- all_dat[!miss,]


#Creating coordinates from fips
f <- as.numeric(paste(geoCounty[,1]))
```

```
s <- matrix(NA,nrow(all_dat),2)
for(i in 1:nrow(all_dat)){
  these <- which(fips[i]==f)
  if(length(these)==1){
    s[i,1] <- geoCounty[these,4]
    s[i,2] <- geoCounty[these,5]
  }
}

#Distance adjacency

d <- rdist.earth(s) # Distance in miles
d[is.na(d)] <- Inf  # Set Alaska and HI to be infinitely far away
diag(d)     <- Inf  # Make sure counties don't neighbor themselves
nearest     <- apply(d,1,min)
D           <- max(nearest[nearest<Inf])

ADJ2 <- ifelse(d <= D,1,0)
#rowSums(ADJ2)

#Linear regression
lreg <- lm(Y ~ X)

#Residuals
res <- lreg$residuals

#Fit
fit <- lreg$fitted.values


#Define a function to make county maps using viridis colour
 county_plot<- function(fips,Y,main="",units="",limits=NULL){
   library(choroplethr)
   library(ggplot2)
   library(viridis)

   temp  <- as.data.frame(list(region=fips,value=Y))
   choro <- CountyChoropleth$new(temp)
   choro$title = main
   choro$set_num_colors(1)
   choro$ggplot_polygon = geom_polygon(aes(fill = value), color = NA)
   choro$ggplot_scale = scale_fill_gradientn(name = units, colours = viridis(32))
   suppressWarnings(choro$render())
 }

 county_plot(fips1,res,main="Plot of residuals",units="OLS Res")
 county_plot(fips1,fit,main="Plot of fitted values",units="OLS Fit")


X           <- scale(X) # Scale covariates
X[is.na(X)] <- 0        # Fill in missing values with the mean
```

```
W <- ADJ2


# Remove counties with no neighbors
junk <- rowSums(W)==0
Y    <- Y[!junk]
X    <- X[!junk,]
W    <- W[!junk,]
W    <- W[,!junk]
fips <- fips[!junk]




tick <- proc.time()[3]
modelCAR  <- S.CARleroux(Y~X, family="gaussian", W=W, burnin=200000, n.sample=1000000,thin=10,
verbose=TRUE)
tock <- proc.time()[3]
(tock-tick)/60 # time in minutes

nu2 <- modelCAR$samples$nu2
tau2 <- modelCAR$samples$tau2
rho <- modelCAR$samples$rho
beta <- modelCAR$samples$beta
Zest <- colMeans(modelCAR$samples$phi)
Zsd <- apply(modelCAR$samples$phi,2,sd)
Yfit <- modelCAR$samples$fitted.values

#MCMC convergence

plot(model1$samples$nu2)
plot(model1$samples$tau2)
plot(model1$samples$rho)
plot(model1$samples$beta)

#Random spatial effects
county_plot(fips,modelCARsumm$Zest,main="Posterior Means",units="Mean")
county_plot(fips,modelCARsumm$Zsd,main="Posterior SD",units="SD")
```