STA6241 - STDA
Homework 3
DUE Friday, May 29

Feel free to work together, but your answers/code should be your own. You must write up your solutions using `LaTeX`. You should submit one pdf file containing solutions/codes. In this assignment you will analyze data on rental rates in Munich. The city uses this data to produce an official rental guide. The dataset we will use has 2035 observations from the year 2003. For each of the problems below, write `R` code to answer the questions. You do not need to turn in written answers to these questions individually. Just turn in your code as well as the items requested at the end of the assignment.

1. Load the file `munichrents.RData` and look at `head(rents)`. The variables are

   - RentPerM2: Monthly rent per square meter in Euros
   - Year: Year the apartment was built
   - Location: District index
   - NoHotWater: Indicator variable for no hot water
   - NoCentralHeat: Indicator variable for no central heat
   - NoBathTiles: Indicator variable for no tiles in the bath
   - SpecialBathroom: Indicator variable for a special bathroom
   - SpecialKitchen: Indicator variable for a special kitchen
   - Room1-Room6: Indicator variables for corresponding number of rooms

   Fit a linear model relating rent per square meter to the covariates using least squares, and extract the coefficient estimates. You can ignore the Location variable for now since we will later treat this as a spatial random effect. Also note that the room indicator variables include one that is redundant, so treat a single room as the baseline (i.e., leave Room1 out of the model, so the intercept corresponds to a single room and coefficients for the others represent adjustments to the intercept for a different number of rooms).

2. There are two `SpatialPolygons` objects associated with this dataset, `districts.sp` and `parks.sp`. The first corresponds to city districts in which apartments may be located. The second corresponds to districts with no possible apartments, such as parks or fields.

   Create an `nb` object with neighbors for the districts, defining neighbors as districts that share a common boundary. Make a plot showing the districts, then add the parks shaded a different color.

3. There are 380 districts in `districts.sp`, and the corresponding district numbers are indicated by the `Location` variable in `rents`.

   I've included a matrix `H` that provides a mapping between the districts as they're ordered in `districts.sp` and as they appear in the `rents` dataframe. Use `H` to create a new vector containing the number of observations in each district, and make a color or grayscale plot to illustrate this.

4. We will now create a Gibbs sampler to sample from the posterior distribution under the following Bayesian model. Let X be the matrix of covariates, including the intercept term. Let n be the number of data points in Y and m be the number of spatial locations in η.

   Data model:

$$Y|\beta, \eta, \sigma^2 \sim MVN(X\beta + H\eta, \sigma^2 I)$$

   Process model:

$$p(\eta|\tau^2) \propto (\tau^2)^{-(m-1)/2} \exp\Big\{ -\frac{1}{2\tau^2}\eta'(D_w - W)\eta \Big\}$$

   where W is the matrix of 0 and 1 indicating the neighborhood structure from problem 2, and D_w is a diagonal matrix with diagonal entries $\sum_j W_{1j}, \cdots, \sum_j W_{nj}$. That is η follows an (improper) intrinsic autoregressive model.

   Prior model: Specify independent priors for β, σ^2, and τ^2 with

$$p(\beta) \propto 1, \quad \sigma^2, \tau^2 \sim InverseGamma(0.001, 0.001)$$

   The full conditional distributions for β, η, σ^2 , and τ^2 are given at the end of this assignment. Construct a Gibbs sampler that cycles through each of the full conditionals and stores the results for $B = 10,000$ iterations. The full conditionals are given below. A few notes to keep in mind when constructing the sampler:

   - The matrix W can be computed from your `nb` object in problem 2; see `help(nb2mat)`. I also included objects X and y with the data file.
   - The function `rinvgamma` is in the library `MCMCpack`.
   - IMPORTANT: The intrinsic autoregressive model is an example of pairwise difference prior. It defines proper distributions for the differences $\eta_i - \eta_j$, but it implicitly contains a distribution for $\frac{1}{m}\sum_{i=1}^m \eta_i$ that has infinite variance. In practice, since there is also an intercept term in $X\beta$, we impose the constraint $\sum_{i=1}^m \eta_i = 0$ when we sample from the full conditional for η. Do this numerically by subtracting the mean $\frac{1}{m}\sum_{i=1}^m \eta_i^{(j)}$ from $\eta^{(j)}$ in each iteration j.

   Turn in the following:

   - Your map with the neighbors from problem 2
   - Your map of the apartment counts for each district

- Trace plots and ACF plots for σ^2 and τ^2
- A table with posterior means of the β and 95% credible intervals constructed using the 0.025 and 0.975 quantiles of the posterior samples
- A color or grayscale map of the posterior means for the vector η
- A color or grayscale map of the posterior standard deviations for the vector η

Full conditionals:

$$\beta|Rest \sim MVN((X'X)^{-1}X'(Y - H\eta), \sigma^2(X'X)^{-1})$$

$$\eta|Rest \sim MVN([H'H/\sigma^2+(D_w-W)/\tau^2]^{-1}H'(Y-X\beta)/\sigma^2, [H'H/\sigma^2+(D_w-W)/\tau^2]^{-1})$$

$$\sigma^2|Rest \sim InverseGamma(0.001 + n/2, 0.001 + (Y - X\beta - H\eta)'(Y - X\beta - H\eta)/2)$$

$$\tau^2|Rest \sim InverseGamma(0.001 + (m-1)/2, 0.001 + \eta'(D_w - W)\eta/2)$$