PROBLEM SET 5

AAEC 6984
INSTRUCTOR: KLAUS MOELTNER

General Instructions

Please type everything in LaTeX (including all Math) and hand in a pdf file. For problems involving Matlab, answer questions in LaTeX, and attach your script, log file, and any graphs to your main pdf file.

Question 1: Bayesian Model Search via MC^3

Consider the data set "waterdays500". It contains observations for 500 Reno, NV, households' weekly water consumption. For each household are minimum of 5 and a maximum of 9 weeks are available, for a total sample size of 2839 observations. You can ignore the panel structure for this exercise and treat all observations as independent.

A main focus of this research is on the impact of weekly outdoor watering patterns on weekly consumption. Specifically, are households that use both of their assigned watering days (Wed, Sat for even address, Thu, Sun for odd address) using less water? If it is windy in an assigned day there could be substantial losses. The data show that many households do NOT use both of their assigned days in a given week. Do they use more or less water in those weeks?

These effects are captured by two sets of variables: Indicators for weekly frequency (wfeq1-wfreq567), and interactions of these indicators with an indicator if both assigned days are used (SB2 - SB567). "SB" stands for "schedule-based". So the coefficient for SB2 is the *differential* effect of using the assigned days for a week with an overall frequency of two, the coefficient for SB3 shows the differential effect of using the assigned days for a week with overall frequency of three, and so on.

The implicit baseline is zero frequency, so you can use all other frequency and SB variables in your model without running into rank violations (dummy traps).

Perform a Bayesian model search via MC^3, regressing **logged** weekly use on a column of ones and the remaining variables in the data (to be specific: bathrooms to SB567, 24 regressors). All of the included variables (except the column of ones, of course) are subject to model scrutiny. Use the same priors as in class and de-mean your data. Use $g = k^2$. Use the same approach for starting draw of $\boldsymbol{\gamma}$ based on OLS t-values as in `mod8_MC3_growth`. Use 200,000 burn-ins and 500,000 keepers. This takes about 20 minutes on my PC, so run time shouldn't be too bad on your end either. Allow for about an hour or so.

The following should get you started:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% PS5 Q1
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

rand('state',37); % set arbitrary seed for uniform draws
randn('state',37); % set arbitrary seed for normal draws

tic;  % start stop watch
%%
[fid]=fopen('c:\klaus\AAEC6984\mlab\logs\ps5Q1.txt','w');
if fid==-1;
   warning('File could not be opened');
   return
else;
   disp('File opened successfully');
end;

% load c:\klaus\AAEC6984\mlab\worksp\waterdays500.txt;
% data = waterdays500;
% clear waterdays500;
% save c:\klaus\AAEC6984\mlab\worksp\waterdays data;
% 'ok'
% break

load c:\klaus\AAEC6984\mlab\worksp\waterdays;
%500 Households with between 5 and 9 weeks of observations on weekly water
%use. 2839 observations in total. You can ignore the panel structure for
%this exercise and simply treat all observations as independent.

%Variables in data:
%%%%%%%%%%%%%%%%%%%%%%%%
%
% 1     cusid           sorted customer id (1 to 1000)
% 2     weekid          id for sampling week (1-9)
% 3     wkuse           weekly use in 1000 gallons
% 4     wkpeak          weekly peak in 1000 gallons (highest daily use in a given week)
% 5     bathrooms       number of bathrooms
% 6     fixtures        number of fixtures (faucets, showers, etc)
% 7     bedrooms        numeber of bedrooms
% 8     age             age of home
% 9     age2            age squared
% 10    lnland          log of lot area (square feet)
% 11    lnsf            log of home's square footage
% 12    lnvalue         log of home's assessed value, in 2008 dollars
% 13    avgtemp         weekly avg. of avg. daily temp, F
% 14    mintemp         weekly avg. of min. daily temp, F
% 15    maxtemp         weekly avg. of max. daily temp, F
% 16    avgwind         weekly avg. of avg. daily wind, mph
% 17    maxwind         weekly avg. of max. daily sustained wind, mph
% 18    maxgust         weekly avg. of max. daily wind gust, mph
% 19    prcp            weekly sum of daily precipitation, inches
% 20    wfreq1          weekly watering frequency = 1 day
% 21    wfreq2          weekly watering frequency = 2 days
% 22    wfreq3          weekly watering frequency = 3 days
% 23    wfreq4          weekly watering frequency = 4 days
% 24    wfreq567         weekly watering frequency > 4days
% 25    SB2              1= frequency = 2 days AND HH uses all assigned watering days
% 26    SB3              1= frequency = 3 days AND HH uses all assigned watering days
% 27    SB4              1= frequency = 4 days AND HH uses all assigned watering days
% 28    SB567            1= frequency = 5 days AND HH uses all assigned watering days
```

```
y=log(data(:,3)*1000); %log of (weekly use in gallons)
n=length(y);
Z=[ones(n,1) data(:,5:28)];
```

Label everything as “ps5Q1”.

a Make a nice Excel table that shows all explanatory variables, and the following columns:

(a) variable name
(b) inclusion probabilities
(c) posterior mean
(d) posterior std
(e) pr(> 0)

b Based on inclusion probabilities, comment on the *relative importance* of the following three groups of regressors: (i) home and lot characteristics, (ii) climate characteristics, (iii) watering frequency and pattern.

c Is there evidence that water use increases with weekly frequency? Explain.

d Is there evidence that water use increases or decreases if all officially assigned days are used in a given week? Does this effect change over frequencies? Explain

Question 2

After some additional thought you decide that if a given model includes a specific frequency, it should also include the corresponding “SB” variables. In other words, the following pairs of variables should always show up together, IF they are included:

i wfreq2 and SB2
ii wfreq3 and SB3
iii wfreq4 and SB4
iv wfreq567 and SB567

a What is the new model space (i.e. how many possible models are there)?

b What is the prior model probability, assuming equal priors for all models?

c Open `gs_MC3` and save it as `gs_MC3ps5Q2`. Save your main script as `ps5Q2`. Make sure it calls the correct Gibbs Sampler. Keep all other settings as for the first model (g, burn-ins, keepers).

The easiest way to implement these restrictions is to let the model indicator vector $\boldsymbol{\gamma}$ be of dimension “number of fully flexible coefficients (call it kf) PLUS number of pairs (call it kp)”, in this case $16 + 4 = 20$. This is the relevant dimension for selecting a new γ for the MH part, and for collecting keepers.

Let $kg = kf + kp$, and make sure to send kg as input to your Gibbs Sampler. This is the new relevant dimension for drawing a new γ in the MH part.

Then, whenever you need to adjust the $\mathbf{X}$ matrix to fit a specific model, use something like

```
gamint=[gamdraw;gamdraw(17);gamdraw(18);gamdraw(19);gamdraw(20)];
Xg=X;
f=find(gamint==0);
Xg(:,f)=[];
```

Do the same when adjusting the dimension of $\mathbf{X}$ for `gamnew`. You also need to make a few adjustment for the draws of $\boldsymbol{\beta}$, when you collect them and need to decide where to put the zeros.

d Make a nice Excel table that shows all explanatory variables, and the following columns:

(a) variable name
(b) inclusion probabilities
(c) posterior mean
(d) posterior std
(e) pr(>0)

e How has this pairing-up restriction affected the posterior results for `SB2, SB3, SB4, SB567`?

f In summary, what can you say about the effectiveness of assigned watering days on water conservation?

Question 3

Using your (model-averaged) results from question 2, plot the *posterior predictive density* of weekly water use, in 1000 gallons, for a home with average settings for home characteristics and climate indicators. Do this for frequencies of 2,3, and 4 watering days per week. For each case plot 2 lines: One for a household that does NOT use all assigned days, and one for a houshold that uses all assigned days.

Be careful - because of the de-meaning of $\mathbf{X}$, including all dummy variables, you cannot use 0's and 1's to switch an indicator variable "on" or "off" when creating predictions. Instead of 0, use the mimimum of the corresponding de-meaned variable. Instead of 1, use the maximum.

For the kernel density evaluation (in preparation for plotting) use 1000 evaluation points to get a smooth figure.

Comment on these graphs - do they confirm your conclusion from the preceding question?

Hint: To compute the PPD for a specific case, do NOT loop over all keepers - this takes forever. Instead, use:

```
cmat=[amat;betamat];

m20=cmat'*x20;
m21=cmat'*x21;
m30=cmat'*x30;
m31=cmat'*x31;
m40=cmat'*x40;
m41=cmat'*x41;

yr20=exp(normrnd(m20,sqrt(sig2mat)'))/1000;
yr21=exp(normrnd(m21,sqrt(sig2mat)'))/1000;
yr30=exp(normrnd(m30,sqrt(sig2mat)'))/1000;
yr31=exp(normrnd(m31,sqrt(sig2mat)'))/1000;
yr40=exp(normrnd(m40,sqrt(sig2mat)'))/1000;
yr41=exp(normrnd(m41,sqrt(sig2mat)'))/1000;
```

where x20, x21, etc are k by 1 vectors of mean settings for the first 16 variables (use the minimum for wfreq1), plus 8 settings for wfreq2 through SB567, depending on the scenario you want to create.