University of Colorado
Department of Civil, Environmental and Architectural Engineering
Advanced Data Analysis Techniques
CVEN 6833

Homework Set 2 Date :10/12/2018
Due :11/13/2018

Topics: Bayesian Hierarchical Modeling, CART, Clustering, PCA, Multinomial Regression, CCA and Multivariate field forecasting

Please present your work neatly. Organization of R-commands, functions will fetch 15% of points.

Bayesian Hierarchical

1. Problem 11 of HW1. Repeat 9. Of HW1 with a ***Bayesian Hierarchical Spatial Model*** (see Verdin et al. 2015, for tips)
In the Bayesian models, plot the posterior historgram/PDF of all the parameters; spatial maps of posterior mean and standard error
The first level can be a GLM with elevation and the second level the spatial model.

2. Problem 12 of HW1. Develop a ***Bayesian Hierarchical Spatial Model*** for problem 7 of HW1. For this the first level is a Binomial distribution (i.e. logistic regression)
You can use the steps/code from Andrew Verdin at http://civil.colorado.edu/~balajir/CVEN5454/R-sessions/bayes/spatial-bayes/
And/or write your own code in RJAGS.

Data and commands for the problems below will be at
http://civil.colorado.edu/~balajir/CVEN6833/HWs/HW-2-2018

PCA

3. The data library at IRI, Columbia University is a wealth of data http://iridl.ldeo.columbia.edu/ (feel free to explore it). Gridded monthly global sea surface temperature (SST) anomalies for the period 1856 - present, also known as 'Kaplan SST' can be obtained from http://iridl.ldeo.columbia.edu/SOURCES/.KAPLAN/.EXTENDED/.v2/.ssta/
(Follow the link to 'Data Files' and download this monthly data as a binary direct access file. You can also do the NETCDF format as well). The reference to the paper describing this data set (Kaplan et al., 1998) is at the bottom of the above link. R-commands to read the binary data file and other commands are in R-Session#2
You wish to identify the space-time patterns of variability of the winter season (Dec – Mar) average SST for the Nov 1900 – Mar 2018 period (118 yrs). To this end,
(i) Perform a PCA on the winter global SST anomalies
-Plot the Eigen variance spectrum for the first 25 modes
-Plot the leading 4 spatial (Eigen vectors) and temporal (PCs) modes of variability
(ii) Perform a rotated PCA (rotate the first 6 PCs) and plot the leading 4 spatial and temporal modes of variability. Compare the results with (i) above.
(iii) Repeat (i) and (ii) on the western US winter precipitation (Nov1900 – Mar2013) 113 yrs
(iv) Perform (i) just for Pacific Ocean domain and correlate the first four PCs of winter SSTs with that of the western US precipitation and show correlation maps.

PCA + Multinomial Regression

4. Construction safety outcomes data have been compiled from accident reports and described in Tixier et al. (2016). The data, attribute information and relevant R-commands can be found at http://civil.colorado.edu/~balajir/CVEN6833/const-data For the safety outcome of *body part*, with five categories (see Table 2 of Tixier et al., 2016).

Twenty leading attributes, all binary (1 or 0 – i.e., present or absent) are provided for each accident record.

(a) Perform a PCA on the attribute data – show the Eigen spectrum and the leading four Eigen vectors

(b) Fit a best multinomial regression with the principal components as predictors to model the categorical probability of injuries to the five category of *body parts* and compute the RPSS

(c) Repeat (b) for the *Injury Severity*

Tixier et al. (2016) modeled these using Random Forests and Stochastic Boosting. They could not get good skill for *Injury Severity.* Compare your results with theirs.

(d) *Apply CART and compare the results*

CART

5. Pick a locations in Arizona and fit a CART model to its

(a) the seasonal pirecipitation along with 3 PCs of the winter SSTs over Pacific. Compare the CART estimtes from observed. Perform a drop-10% cross validation and boxplot correlation and RMSE

(b) Fit a random forest model and compare the performance with (a) above.

Cluster Analysis – Kmeans and Extremes and SOM

6. Cluster the January and June precipitation over Colorado – use latitude, longitude and elevation. The clustering involves:

(i) identify the best number of clusters, say, Kbest

Select a desired number of clusters, say, *j;* cluster the data and computer the WSS; repeat for *j=1,10;* plot *j versus WSS*; and select the best number of clusters Kbest, where the WSS starts to saturate.

(ii) cluster the data into Kbest clusters and display them.

Try with and without elevation in the clusters and summarize the results.

7. For the winter seasonal maximum precipitation perform clustering of the Extremes using the method proposed in Bernard et al. (2013) and used in Bracken et al. (2015) – these papers are linked on the class page. Comment on the cluster patterns of the average and maxima. [**data at http://civil.colorado.edu/~bracken/western-us-extremes-data/]**

8. Apply self-organizing maps (SOM) to the Western US winter seasonal precipitation data. Try 3 x3 or 4 x 4 node configuration. For each node composite the winter SSTs. This will provide insights into the relationship between the SSTs and their signature on western US winter precipitation.

Joint Spatial Analysis & Field Forecasting using CCA

9. One of the objectives of multivariate analysis is to enable multivariate forecasting. Here we wish to predict the winter precipitation over Arizona as a function of winter Pacific SSTs. We use CCA for this and the simplified steps are described below.

(i) Take the leading ~ 5 PCs of the winter SSTs [which you can obtain following the procedure in problem 3] and perform CCA with the leading ~ 5 PCs of winter precipitation.

(ii) Fit a regression for each canonical variate – canonical variate of precipitation related to canonical variate of SSTs. Use these regressions to predict the flow variates and back transform them to the original space – i.e., the precipitation space.

(iii) Evaluate the performance by computing R^2 between the observed and predicted summer precipitation at each grid point.

(iv) Use the model to predict the maximum precipitation for the last two years.