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

Homework Set 2 Date :10/12/2021
Due :11/16/2021

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.

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

1. Problem 9 from HW1 with a ***Bayesian Hierarchical Spatial Model*** (see Verdin et al. 2015, for tips)
[In the Bayesian models, plot the posterior historgram/PDF of the parameters; spatial maps of posterior mean and standard error]

2. Compute the 75th percentile of the average winter 3-day maximum precipitation across all the stations and using this as the threshold categorize the annual precipitation at each station to a binary variable – 0 if annual precipitation less than the threshold, 1, otherwise.
i. Fit a 'best' GLM (i.e. logistic regression) with the appropriate link function using one of the objective functions. Test the model goodness using ANOVA
ii. Estimate the function on the grid and plot the surface. Also plot the standard error.
iii. Repeat (i) and (ii) with Local GLM

3. Develop a ***Bayesian Hierarchical Spatial Model*** for problem 2 above. 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.

PCA – Field Correlations
4. 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 (Dec - Mar) 3-day maximum precipitation and SST for the 1964 – 2018 period. 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.
(iv) Correlate the first four PCs of rainfall with the SSTs and vice-versa and, show correlation maps.

PCA + Multinomial Regression

5. 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 – Multivariate Forecasting

6. Fit a CART model to the leading 4 PCs of winter 3-day precipitation extreme, using the four leading PCs of winter SSTs as covariates.

-You will fit a CART model for each PC separately

-with the CART estimates of the PCs, estimate the 'model precipitation' at all the locations by multiplying the PCs estimates with Eigen Vectors

(a) Compare the CART model precipitation with historic. Also Perform a drop-10% cross validation and boxplot correlation and RMSE

(b) Repeat (a) above with a random forest model

Compare and summarize your findings.

Cluster Analysis – Kmeans and Extremes and SOM

7. Cluster the

(i) winter 3-day maximum precipitation average

Use latitude, longitude and elevation.

The clustering involves:

(a) 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.

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

(iii) Also employ any other clustering method – K-medoid clustering or Hierarchical Clustering

Compare and summarize your findings.

8. For the winter 3-day maximum precipitation over Southwest US 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 incomparison with those from problem 7. Above.

9. Apply self-organizing maps (SOM) to the winter 3-day maximum precipitation over Southwest US. Try 3 x 3 or 4 x 4 node configuration. For each node composite the rainfall and the SSTs and show the maps. This will provide insights into the relationship between the SSTs and their signature on the rainfall extremes.

Joint Spatial Analysis & Field Forecasting using CCA

10. One of the objectives of multivariate analysis is to enable multivariate forecasting. Here we wish to predict the winter 3-day maximum precipitation over Southwest U.S as a function of winter SSTs. We use CCA for this and the simplified steps are described below.
(i) Take the leading ~ 4 PCs of the winter SSTs [which you can obtain following the procedure in problem 4] and perform CCA with the leading ~ 4 PCs of 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) Do a leave-one-out cross validation
Compare with results from problem 6.

Bonus Extensions – if interested

Space-time model of winter precipitation extreme with winter SST PCs as covariates (based on Ossondon et al., 2021, WRR)