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

Homework Set 2 Date :3/1/2025
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 resources for the problems below will be at
http://civil.colorado.edu/~balajir/CVEN6833/HWs/HW-2

1. Problem 7 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. Use a precipitation threshold of 150 cm across all the stations and categorize the precipitation at each station to a binary variable – 0 if it is less than the threshold, 1, otherwise. Use the suite of predictors from HW1.
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 or STAN.

PCA – Field Correlations

4. Summer season (June-September total) rainfall over India exhibits significant spatial and temporal variability. You wish to identify the space-time patterns of variability of. To this end,
(i) Perform a PCA on the summer season rainfall (cm) on a 0.25^o x 0.25^o grid (i.e. the 'Rajeevan grid') over India is available for the period 1901 to 2016. Perform the analysis for the 1950 ~ 2016 period
-Plot the Eigen variance spectrum for the first 15 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 (uncorrelated) of rainfall with summer global tropical Sea Surface Temperatures (SSTs) and show correlation maps.

5. Repeat 5 by swapping summer rainfall with summer global tropical SSTs and vice-

PCA + Multinomial Regression

6. 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.

CART, Random Forest and Gradient Boosting

7. For the all India Summer Rainfall with the tropical SST PCs as covariates, fit these models
(a) CART
(b) Gradient Boosting Tree
(c) Random Forest
Infer the outputs and compare their relative performances on RMSE, plots of historical vs model rainfall.

CART and Random Forest – Multivariate Forecasting

8. Fit a CART model to the leading 4 PCs of summer season rainfall over India, using the 4 - 5 leading PCs of summer global tropical 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. Boxplot correlation and RMSE across the grids
(b) Repeat (a) above with a random forest model
Compare and summarize your findings.

Cluster Analysis – Kmeans and Extremes and SOM

9. Cluster the seasonal average precipitation using latitude, longitude and elevation (over average value at each location)
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.
(iv) Cluster the same using latitude, longitude, elevation, distance to Bay of Bengal and distance to Arabian Sea.
Compare and summarize your findings.

10. Perform clustering of the summer season precipitation 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 in comparison with those from problem 7. Above.

11. Apply self-organizing maps (SOM) to the summer season rainfall over India. 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.

Bonus - Joint Spatial Analysis & Field Forecasting using CCA

12. Repeat 8 using Canonical Correlation Analysis (CCA). The simplified steps are described below.

(i) Take the leading ~ 4 PCs of the summer global tropical SSTs and perform CCA with the leading ~ 4 PCs of summer rainfall over India.

(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.