PREDICTING INDIAN SUMMER MONSOON RAINFALL (ISMR) USING A MIXTURE OF REGRESSION MODEL

Vidyashankar Sivakumar[1], Moumita Saha[2], Pabitra Mitra[2], Arindam Banerjee[1]

***Abstract*—We consider the problem of predicting total Indian summer monsoon rainfall (ISMR). A popular approach in prior literature [1], [2] has been to fit a regression model with the precipitation as predictand and various climatological indices and parameters as predictors. The predictor climatological indices and parameters are detected through an analysis of their linear correlations with the Indian monsoon precipitation. Due to limited success of such prior work based on a fixed regression model, in this work we investigate ISMR prediction based on the hypothesis that Indian monsoon operates in a few different regimes, where different predictors become relevant and influential. We model such a multi-regime setting as a finite mixture of linear regressions (MLR) model [3], with a ridge regression model for each regime of operation. The parameters of the model are determined using the Expectation Maximization (EM) algorithm. The prediction procedure consists of identifying the regime of operation and then applying the corresponding regression model. The MLR model seems to improve overall prediction accuracy compared to a single fixed regression model (SLR).**

I. MOTIVATION

India receives a major portion of annual rainfall during the months June-September from the southwest monsoon winds. With a large population and a major agrarian economy, it becomes imperative to accurately predict and characterize the Indian summer monsoon rainfall (ISMR). To get a sense of the prediction task, the mean ISMR (based on 1941-1990) is 890 mm with a coefficient of variation of about 10% [1]. ISMR in the range 90% to 110% of the mean is considered normal, whereas anything below or above 90 or 110% is considered deficient or excessive respectively, and can have devastating effects on the agricultural sector in India [4].

A popular approach in past work has been to predict ISMR using regression, with various climatological indices and parameters as the predictors [2], [1], [4]. The predictors are detected based on their high linear correlations with ISMR. In a nutshell, all of them are variations of the following basic model, which we call the single linear regression (SLR) model:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \langle X_i, \beta \rangle)^2 + \lambda R(\beta) \tag{1}$$

Here n is the number of years used for training the model, y_i is the ISMR in year i, X_i is the vector of predictors, β is weight on the predictors, λ is the regularization parameter and $R(\cdot)$ is any suitable regularizer. Existing regression based models do not predict ISMR with acceptable accuracy.

In this work, we work on the hypothesis that ISMR operates in two different regimes. Mathematically, this can be modeled as a mixture of regression (MLR) [3] model as follows:

$$\min_{z_i \in \{0,1\}, \beta_1, \beta_2} \sum_{i=1}^{n} (y_i - \langle X_i, z_i\beta_1 + (1 - z_i)\beta_2 \rangle)^2 + \lambda_n R(z_1\beta_1 + z_2\beta_2)$$

Basically the years are separated into two separate clusters, with the behavior in each cluster modeled with a different regression parameter. The parameter z_i identifies the cluster to which the year belongs. Due to limited training data we operate with the two regime hypothesis, although it remains to be thoroughly evaluated if there are more than two regimes of operation.

II. METHOD

We compare the performance of the single linear regression (SLR) and mixture linear regression (MLR) [3] models in predicting ISMR. We use l_2 regularization. For MLR, during the training phase the $z_i's$ are not known apriori and hence we use an Expectation Maximization (EM) style alternating minimization algorithm as below to fit the parameters. Note that, we work on the

Corresponding author: Vidyashankar Sivakumar, University of Minnesota, sivak017@umn.edu [1] Computer Science Department, University of Minnesota, Twin Cities [2] Indian Institute of Technology, Kharagpur

```
Input: $\beta_1^{(0)}, \beta_2^{(0)}$,no. of iterations $t_0$, samples
       $\{(y_i, X_i), i = 1, 2, \dots, n\}$
Output: $\beta_1^{(t_0)}, \beta_2^{(t_0)}$
for $t \leftarrow 0$ to $t_0 - 1$ do
    (Assign labels)
    $J_1, J_2 \leftarrow \phi$
    for $i \leftarrow 1$ to $n$ do
        if $\left|y_i - \langle X_i, \beta_1^{(t)} \rangle\right| < \left|y_i - \langle X_i, \beta_2^{(t)} \rangle\right|$ then
            $J_1 \leftarrow J_1 \cup \{i\}$
        else
            $J_2 \leftarrow J_2 \cup \{i\}$
        end
    end
    (Solve least squares)
    $\beta_1^{(t+1)} \leftarrow \min_{\beta_1} \sum_{i \in J_1} (y_i - X_i\beta)^2$
    $\beta_2^{(t+1)} \leftarrow \min_{\beta_2} \sum_{i \in J_2} (y_i - X_i\beta)^2$
end
```

Algorithm 1: EM algorithm

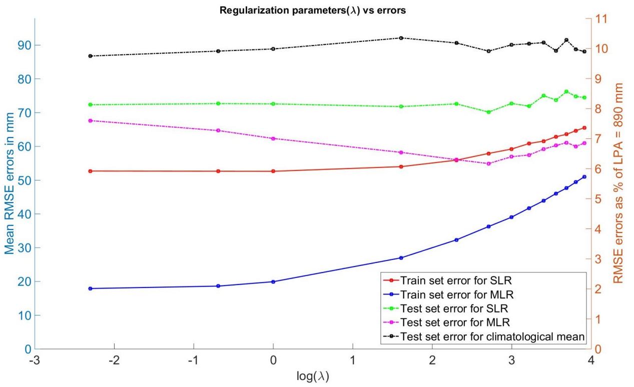

Fig. 1. Behaviour with different λ.

assumption that each year belongs to a single cluster and hence do a hard assignment of years to clusters.

III. EVALUATION

We consider ISMR during the period 1948-2013. There are 15 predictor parameters whose details are available in [5]. Due to limited space, we do not list them here. We generate 200 datasets by randomly splitting the 66 years into train and test sets in the ratio 85:15. For each dataset, we run (i) SLR model and (ii) MLR model on the training set and use it on the test set. For MLR, for the results on the test set, we assign years to the cluster which give the best result. Figure 1 plots the train and test set errors averaged over 200 runs, run for each regularization parameter λ. For very low values of λ the training error for MLR is very low due to overfitting. For the results on the test set, we see no particular pattern for SLR whereas for MLR the error decreases reaching an optimum value at around $\lambda = 10$ and then increases again. In particular, in all cases MLR performs better on the test set than SLR and the climatological mean. Figure 2 compares the average RMSE values of SLR, MLR and climatological mean for one particular 200 times run with $\lambda_{SLR} = 0.5$ and $\lambda_{MLR} = 10$. MLR(54.89 ± 11.53) performs much better than SLR(72.16 ± 15.12) and the climatological mean(88.67 ± 16.43). In particular in the comparison between SLR and MLR, on an average, MLR performs better than SLR in 175/200 runs. A t-test of significance on the (MLR - SLR) mean errors rejects the null hypothesis with a p-value of 3.53×10^{-39}. An analysis of the clusters shows that two clusters are consistently formed with each containing 22 and 27 years respectively. We intend to do a more thorough analysis of the clusters and its relationship with climatological phenomena like El Nino/La Nina etc. to get a better mechanistic understanding of ISMR.

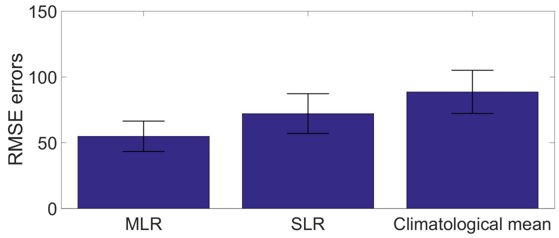

Fig. 2. Average test set errors for MLR, SLR and climatological mean.

ACKNOWLEDGMENTS

The research was supported by NSF Expeditions in Computing grant - IIS-1029711. The work was also supported by NSF grants IIS-0916750, CNS-1314560, and by NASA grant NNX12AQ39A. Arindam Banerjee acknowledges support from IBM and Yahoo. Access to computing facilities by University of Minnesota Supercomputing Institute.

REFERENCES

[1] M. Rajeevan, D. Pai, R. Kumar, and B. Lal, "New statistical models for long-range forecasting of southwest monsoon rainfall over India," *Climate Dynamics*, vol. 28, no. 7-8, pp. 813–828, 2007.

[2] T. DelSole and J. Shukla, "Linear Prediction of Indian monsoon rainfall," *Journal of Climate*, vol. 15, pp. 3645–3658, 2002.

[3] G. McLachlan and D. Peel, *Finite Mixture Models*. Wiley Series in Probability and Statistics, 2005.

[4] M. Rajeevan, D. Pai, S. Dikshit, and R. R. Kelkar, "IMD's new operational models for long-range forecast of southwest monsoon rainfall over Indian and their verification for 2003," *Current Science*, vol. 86, no. 3, pp. 422–431, 2004.

[5] M. Saha, P. Mitra, and A. Chakraborty, "Predictor-year subspace clustering-based ensemble prediction of the Indian monsoon," *To appear*, 2015.