5th International Workshop on Climate Informatics
September 24-25, 2015
Hosted by the National Center for Atmospheric Research in Boulder, CO

IDENTIFYING DRIVING FACTORS BEHIND INDIAN MONSOON PRECIPITATION USING MODEL SELECTION BASED ON DATA DEPTH

Subhabrata Majumdar[1], Lindsey Dietz[1], Snigdhansu Chatterjee[1]

***Abstract*—We introduce a novel one-step model selection technique for general regression estimators, and implement it in a linear mixed model setup to identify important predictors affecting Indian Monsoon precipitation. Under very general assumptions, this technique correctly identifies the set of non-zero values in the true coefficient (of length p) by comparing only $p+1$ models. Here we use wild bootstrap to estimate the selection criterion. Mixed models built on predictors selected by our procedure are more stable and accurate than full models across testing years in predicting median daily rainfall at a station.**

I. MOTIVATION

Obtaining a meaningful statistical model of Indian summer monsoon precipitation is challenging from both physical and statistical perspective due to its erratic nature. This is an extremely important problem because monsoon precipitation is the major source of water for the mostly seasonal agricultural practice in the subcontinent. Dietz and Chatterjee show in [1] and [2] that in addition to several covariates and climate variables (found in references cited therein) there is a need to include random effects in modeling to quantify the uncertainty in the process.

Statistical model selection may guide the choice of covariates so that only relevant ones are included, thus improving both accuracy and precision of prediction. However, unlike traditional scenarios like linear regression, the problem involves both fixed and random effects, and potentially heteroskedastic error structure. Also, linearity or other regression assumptions are not guaranteed to hold and are hard to verify with in the current context. Consequently, traditional likelihood-based methods may suffer from lack of robustness, while ad hoc techniques like randomization imply strong hidden assumptions which are unlikely to hold with current data. Here we tackle all these issues by selecting covariates utilizing a general model selection criterion that depends on the behavior of coefficient estimates in the parameter space, and demonstrate efficacy of the resulting model in out-of-time predictions.

Corresponding author: S. Majumdar, University of Minnesota Twin Cities, Minneapolis MN; majum010@umn.edu[1]

II. METHODS

A. Data depth-based model selection

The depth of a point $\mathbf{x} \in \mathbb{R}^p$, is any scalar measure of its centrality with respect to a data cloud $\mathbf{X}$ (or equivalently the underlying distribution F [3], and is denoted by $D(\mathbf{x}, \mathbf{X})$ (or $D(\mathbf{x}, F)$). Consider now a regression setup where estimates of a coefficient vector $\boldsymbol{\beta}$, based on a sample of size n, follow sampling distributions that can be asymptotically approximated by elliptic distributions F_n centered at $\boldsymbol{\beta}$ that approach unit mass at $\boldsymbol{\beta}$ as $n \to \infty$. In this context, we define a model selection criterion for any candidate model, specified by α, the set of indices where $\boldsymbol{\beta}$ takes non-zero values:

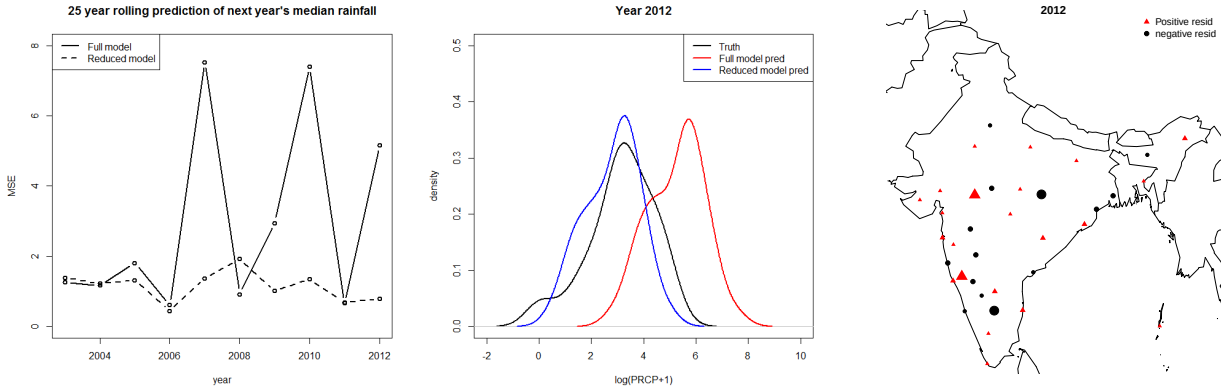
$$C_n(\alpha) = \mathbb{E}\left[D\left(\tilde{\boldsymbol{\beta}}_\alpha, F_n\right)\right] \tag{1}$$

Here $\tilde{\boldsymbol{\beta}}_\alpha$ is the estimate of $\boldsymbol{\beta}_\alpha$ obtained from data concatenated with 0 at indices not in α, and D is any depth function. When α does not contain all non-zero indices in the true model, we have $C_n(\alpha) \to 0$ as $n \to \infty$ [4]. Otherwise for any n, the criterion maximizes at the smallest correct model, say α_0, and decreases monotonically as zero indices are added to α_0 one-by-one. In a sample setup, the unknown distribution F_n and the expectation in 1 are estimated by bootstrap.

For large enough n, we can obtain the most parsimonious correct model from true C_n values of only $p+1$ models where p is the dimension of $\boldsymbol{\beta}$. We use the following scheme:

1) Calculate C_n for full model;
2) Drop a predictor, calculate C_n for the reduced model;
3) Repeat for all p predictors;
4) Collect predictors dropping which causes C_n to decrease. These are the predictors in the smallest correct model.

5th International Workshop on Climate Informatics
September 24-25, 2015
Hosted by the National Center for Atmospheric Research in Boulder, CO



(a) (b) (c)

Fig. 1: (a) Comparison of MSE for full and reduced model predictions across years, (b) Density plot for actual log rainfall and predictions in year 2012, (c) Station-wise reduced model residuals for 2012

B. Linear Mixed Models

Linear mixed models add an extra layer of complexity above the standard linear model setup by assuming latent unobservable random effects. We define this model as:

$$\mathbf{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \epsilon \tag{2}$$

where $\mathbf{Y}_{n\times 1}$ is the vector of responses, $X_{n\times p}$ is the matrix of predictors and $\boldsymbol{\beta}_{p\times 1}$ is the vector of coefficients, which are referred as *fixed effects* here. The latent layer comes in the form of the *random effect* vector $\boldsymbol{\gamma}_{k\times 1}$ ($k < n$), and the random effects design matrix $Z_{n\times k}$. The vector $\boldsymbol{\gamma}$ is modeled as a random draw from $\mathcal{N}_k(\mathbf{0}, \Sigma)$, with $\Sigma_{k\times k}$ being positive definite.

III. Data and Implementation

We use data from 36 weather stations across India for 1978-2012 to model daily median rainfall at a station within a year. In addition to station-specific variables of latitude, longitude, and elevation, we use yearly medians of local variables including maximum and minimum temperature, tropospheric temperature difference (ΔTT), u- and v- winds at 200, 600 and 850 mb, Niño 3.4 anomaly and Indian Dipole Mode Index (DMI), as well as of global variables that have known connections with the Indian monsoon pattern. These include 10 indices of the Madden-Julian Oscillation (MJO), 9 northern hemisphere teleconnection indices, solar flux levels, and land-ocean Temperature Anomaly (TA).

We implement the model in 2 taking all the variables mentioned above as fixed effects, and year as a single random effect (i.e. $k = 1$). We use separate wild bootstraps on estimated random effects and residuals to obtain resampled observations. Among 35 predictors considered, 21 are selected by our procedure- all of which have been proposed in literature. TA seems to have a large influence. We also note several MJO indices are selected when starting from a full model with everything but TA, but are dropped in favor of TA when it is included in the full model.

We use a 25 year rolling validation scheme to compare prediction performances of full and reduced models. For each of the years 2003–2012, we use the past 25 year's data as training data. Figure 1 summarizes some of the results. Predictions from the reduced model are generally more stable across testing years (less MSE in panel (a)) than those from full model. Also, as demonstrated by panels (b) and (c) for year 2012, the reduced model provides a less biased estimate of the true values. We observe this reduction in bias for all 10 testing years.

Future work includes investigating spatio-temporal dependencies, detailed studies into algorithmic efficiency issues and further development of theoretical properties of the proposed model selection tool.

References

[1] L. R. Dietz and S. Chatterjee, "Logit-normal mixed model for Indian monsoon precipitation," *Nonlin. Processes Geophys.*, vol. 21, pp. 939–953, sep 2014.

[2] L. R. Dietz and S. Chatterjee, "Investigation of Precipitation Thresholds in the Indian Monsoon Using Logit-Normal Mixed Models," in *Machine Learning and Data Mining Approaches to Climate Science: Proceedings of the Fourth International Workshop on Climate Informatics* (V. Lakshmanan, E. Gilleland, A. McGovern, and M. Tingley, eds.), pp. 239–246, Springer, 2015.

[3] Y. Zuo and R. Serfling, "General notions of statistical depth function," *Ann. of Statist.*, vol. 28-2, pp. 461–482, 2000.

[4] S. Majumdar and S. Chatterjee, "A model selection criterion for regression estimators based on data depth." Working paper, 2015+.