**INTRODUCTION**

**Objective**: Selection of important predictors behind Indian Monsoon rainfall, and using them to build a predictive model.

**Challenges for covariate selection**:
- Several sources of variability, e.g. variation across years and weather station;
- Potentially heteroskedastic error structure;
- Linearity or other regression assumptions are not guaranteed to hold and are hard to verify;
- Huge number of possible models ($2^p$) for even moderate number of predictors ($p$).

**Our solution**: Use a novel model selection criterion based on data depth that works on a wide range of models, and selects important predictors by comparing only $p + 1$ models.

**DATA AND MODELLING**

Annual median observations for 1978-2012;

**Fixed covariates** ($X_{i,p} = 1, p = 35$)

(A) **Station-specific** (from 36 weather stations across India) Latitude, longitude, elevation, maximum and minimum temperature, tropospheric temperature difference ($\Delta T$), Indian Dipole Mode Index (DMI), Nilio 3.4 anomaly;

(B) **Global**:
- $u$-wind and $v$-wind at 200, 600 and 850 mb;
- 10 indices of Madden-Julian Oscillations: 20E, 70E, 80E, 100E, 120E, 140E, 160E, 120W, 40W, 10W;
- Teleconnections: North Atlantic Oscillation (NAO), East Atlantic (EA), West Pacific (WP), East Pacific-North Pacific (EPN), Pacific/North American (PNA), East Atlantic/Western Russia (EAWR), Scandinavia (SCA), Tropical/Northern Hemisphere (THN), Polar/Eurasia (POL);
- Solar Flux;
- Land-Ocean Temperature Anomaly (TA).

**Random effects (R)**: Random intercept by year;

**Linear Mixed Model (LMM)**: $Y$: log of annual median rainfall at a weather station (WS);
- **Model 1**: $\gamma_{WS, year} = \gamma \sim N(\theta_{WS, year}, \sigma^2)$;
- $\theta_{WS, year} = \gamma_{WS, year} + \beta + \gamma_{year}$;
- **Model 2**: $\beta_{WS, year} = N(\gamma, \tau^2)$.

**DEPTH-BASED MODEL SELECTION**

![Figure 1: Samples from bivariate normal and their depths: points away from center have less depth while those close to center have more depth.](image)

**Data depth**

A nonparametric, scalar measure of centrality for a point $x$ in sample space with respect to a data cloud $X$ or probability distribution $F$: denoted by $D(x, X)$ or $D(x, F)$, respectively [Zuo and Serfling, 2000].

**The selection criterion**

In any regression setup, consider estimators of the coefficient $\beta$ based on a sample of size $n$ having elliptical sampling distributions $F_0$ centered at $\beta$ that approach unit mass at $\beta = n \rightarrow \infty$.

For a candidate model, uniquely specified by its non-zero index set $\alpha$, define

$$C_n(\alpha) = E[D(\beta_n, F_n)]$$

where $\beta_n$ is estimate of truncated coefficient vector $\beta_\alpha$, concatenated with 0 at indices not in $\alpha$.

Suppose $\alpha_0$ is the smallest correct model. Then:
- For any correct model, i.e. when $\alpha \supseteq \alpha_0$, we have $C_n(\alpha) = C(\alpha)$, i.e. depends only on $\alpha$;
- For any wrong model, $C_n(\alpha) \rightarrow 0$ as $n \rightarrow \infty$;
- Among correct models, $C(\alpha)$ maximizes at $\alpha = \alpha_0$, and decreases monotonically as superfluous variables are added;
- In a sample setup, we use bootstrap to estimate $\hat{\beta}_n$ and $F_n$ [Majumdar and Chatterjee, 2015+].

**THE ONE-STEP ALGORITHM**

1. For large enough $n$, 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.

**IMPLEMENTATION**

**Bootstrap scheme**
- Wild bootstrap [Mammen, 1993];
- Say $n$ is total number of observations, $k$ is number of years;
- Start with estimators from initial LMM: $\hat{\beta}_1, \gamma_1$;
- Generate $X_{WS, year} \sim N(0, \sigma_2)$, $\gamma_{WS, year} \sim N(0, \tau^2)$;
- Get ‘new’ observations: $Y_{WS, year} = X_{WS, year} + \gamma_{WS, year}$;
- For large enough $n$, calculate $C_n$ for full model;
- Drop a predictor, calculate $C_n$ for the reduced model;
- Repeat for all $p$ predictors;
- Collect predictors dropping which causes $C_n$ to decrease. These are the predictors in the smallest correct model.

**Table 1**: Ordered values of $C_n$ from bootstrap.

**Figure 1**: Samples from bivariate normal and their depths: points away from center have less depth while those close to center have more depth.

**Figure 2**: Bias and MSE of rolling predictions.

**Figure 3**: Density plots of 2012 predictions and truth.

**Figure 4**: Stationwise residuals for 2012.

**DISCUSSION**

- All selected variables (colored blue in Table 1) have documented effects on Indian monsoon;
- EPN teleconnection and 120W MJO are both selected: both deal with same longitudinal region;
- Interesting variables: Solar Flux and Polar/Eurasia teleconnection (POL): an indicator of Eurasian snow cover;
- TA has a large influence. Several MJO indices, particularly 80E and 40W, 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;
- Reduced model predictions have consistently less bias and are more stable across testing years (Figs. 2 and 3). Also there are no spatial patterns in residuals (Fig. 4).

**References**


**Acknowledgement**

NSF grant IIS-1029711.