Depth model selection outputs using LMM and wild bootstrap

July 21, 2015

1 Variables

Latitude, Longitude, Elevation, Tmax, Tmin, DelTT, DMI, Nino3.4;

u-wind at 200, 600 and 850;

v-wind at 200, 600, 850;

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 (EPNP), Pacific/North American (PNA), East Atlantic/Western Russia (EAWR), Scandinavia (SCA), Tropical/Northern Hemisphere (TNH), Polar/Eurasia (POL);

Solar Flux;

Land-Ocean Temperature Anomaly

Total 35 variables. Medians of precipitation and all these variables are taken across stations and year, and log of precipitation is modeled as LMM with all variables as fixed effects and yearwise random effect.

2 Bootstrap scheme

We are assuming the model $Y = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon}$. We do bootstrap on the random effect and residuals separately, i.e. bootstrap samples are calculated as:

$$Y_b = X\boldsymbol{\beta} + Z\boldsymbol{\gamma}_{b_1} + \boldsymbol{\epsilon}_{b_2} = X\boldsymbol{\beta} + ZU_{b_1}\boldsymbol{\gamma} + U_{b_2}\boldsymbol{\epsilon}$$

Where diagonals of U_{b_1} are mean 0 variance τ_k^2 and U_{b_2} are mean 0 variance τ_N^2: k and N being number of classes and number of samples, respectively. Number of bootstrap samples is 1000.

3 Model selection

The model selection procedure is as given below:

1. Compute depth-based model criterion C_n for the full model;
2. Drop a predictor, compute C_n for the reduced model: repeat for all predictors;
3. Collect the predictors dropping which causes a *decrease* in C_n. Build final model on these predictors.

Criterion values for all drop-1 models as well as full model (`<none>`) are given in the table below and sorted in increasing order. Also (experimental) since C_n is an expected value and we are comparing whether one mean is less than another, I thought of calculating p-values for each drop-1-vs.-full model comparison, by t-test with less than type alternative. They are given in third column of the table.

Scheme I: $\tau_n^2 = n^{0.1}$

```
             DroppedVar        Cn        pValue
1                - TMAX 0.1325144  0.000000e+00
2           - ELEVATION 0.1560887  0.000000e+00
3         - TempAnomaly 0.1745729 1.994038e-252
4  - del_TT_Deg_Celsius 0.2107055  6.206508e-84
```

5	- v_wind_850	0.2415606	1.059260e-07
6	- Nino34	0.2421164	5.947362e-07
7	- SolarFlux	0.2422948	7.880482e-07
8	- EPNP	0.2428719	1.553918e-06
9	- X120W	0.2432240	6.570307e-07
10	- POL	0.2442205	3.053574e-05
11	- u_wind_850	0.2442381	3.415376e-05
12	- LONGITUDE	0.2449414	1.107313e-04
13	- TNH	0.2464970	1.218154e-03
14	- LATITUDE	0.2496559	5.773439e-02
15	- EA	0.2498657	6.716721e-02
16	- u_wind_600	0.2498830	7.044640e-02
17	- DMI	0.2512373	1.888425e-01
18	- u_wind_200	0.2513900	2.072419e-01
19	- X20E	0.2525811	3.745168e-01
20	- NAO	0.2530687	4.614479e-01
21	- EAWR	0.2532869	4.987354e-01
22	<none>	0.2532943	1.000000e+00
23	- v_wind_600	0.2542720	6.610688e-01
24	- TMIN	0.2552696	8.003255e-01
25	- v_wind_200	0.2552784	7.984362e-01
26	- X140E	0.2556427	8.463850e-01
27	- PNA	0.2558863	8.674652e-01
28	- X160E	0.2562370	9.028804e-01
29	- WP	0.2568667	9.359407e-01
30	- SCA	0.2617446	9.998071e-01
31	- X120E	0.2618847	9.998383e-01
32	- X70E	0.2628806	9.999772e-01
33	- X40W	0.2632758	9.999829e-01
34	- X100E	0.2654437	9.999999e-01
35	- X10W	0.2672907	1.000000e+00
36	- X80E	0.2725583	1.000000e+00

Scheme II: $\tau_n^2 = n^{0.2}$

	DroppedVar	Cn	pValue
1	- TMAX	0.1594159	1.924099e-318
2	- ELEVATION	0.1831007	5.087412e-209
3	- TempAnomaly	0.1950830	1.610033e-149
4	- del_TT_Deg_Celsius	0.2272400	2.230674e-32
5	- v_wind_850	0.2445276	5.883643e-05
6	- LONGITUDE	0.2461506	7.222295e-04
7	- Nino34	0.2463974	1.399145e-03
8	- SolarFlux	0.2474990	6.114477e-03
9	- u_wind_850	0.2482293	1.402247e-02
10	- POL	0.2488754	2.575858e-02
11	- u_wind_600	0.2494258	4.762298e-02
12	- EPNP	0.2495537	4.843258e-02
13	- EAWR	0.2512024	1.698848e-01
14	- LATITUDE	0.2514350	2.025272e-01
15	- EA	0.2515075	2.070797e-01
16	- TMIN	0.2525158	3.550617e-01
17	- v_wind_200	0.2525735	3.644007e-01

```
18              - NAO 0.2526962  3.821308e-01
19       - v_wind_600 0.2527427  3.920411e-01
20            - X120W 0.2528025  3.941054e-01
21              - TNH 0.2530542  4.405466e-01
22       - u_wind_200 0.2530875  4.486618e-01
23             <none> 0.2533959  1.000000e+00
24              - PNA 0.2537365  5.583152e-01
25               - WP 0.2558643  8.517429e-01
26             - X20E 0.2562980  8.970100e-01
27              - DMI 0.2564950  9.070440e-01
28            - X160E 0.2595774  9.955072e-01
29              - SCA 0.2597392  9.960893e-01
30            - X140E 0.2601329  9.977884e-01
31            - X120E 0.2605941  9.986716e-01
32             - X70E 0.2618301  9.997547e-01
33             - X40W 0.2624374  9.999073e-01
34            - X100E 0.2633385  9.999840e-01
35             - X10W 0.2641683  9.999959e-01
36             - X80E 0.2733044  1.000000e+00
```

4 Discussion

- Top 5-6 variables are as expected;
- EPNP teleconnection and 120W MJO (X120W) are both selected in the model. Both deal with the same longitude region... are they related?
- Interesting variables: Solar Flux and Polar/Eurasia teleconnection (POL): an indicator of Eurasian snow cover;
- Temperature Anomaly has a huge effect. If we do not include this variable, some MJOs are selected, particularly 80E and 40W get selected consistently (Indian ocean and Atlantic oscillations?) across bootstrap schemes. But including temp anomaly makes MJOs almost expendable;
- A thing about p-values: don't really know how much they are useful here, but when I tried the method for linear model selection on a simulated dataset with 25 predictors from Charlie's webpage (`http://www.stat.umn.edu/geyer/5102/examp/select.html`) all variables with true non-zero coeffs had p-values < 0.05. Here is its output ($\tau_n^2 = n^{0.2}$):

```
  DroppedVar        Cn       pValue
1       - x2 0.2051023 1.698015e-42
2       - x3 0.2164894 1.192099e-10
3       - x4 0.2169777 1.211069e-09
4       - x1 0.2209519 1.723179e-04
5       - x5 0.2215058 6.844795e-04
6      - x20 0.2224559 4.012763e-03
7      - x21 0.2252643 1.730813e-01
8       - x9 0.2257273 2.498471e-01
9     <none> 0.2268667 1.000000e+00
```

```
10      - x22 0.2279793 7.439175e-01
11      - x17 0.2285479 8.406974e-01
12      - x10 0.2285719 8.426074e-01
13       - x6 0.2290502 8.971758e-01
14      - x13 0.2292558 9.147554e-01
15      - x19 0.2293505 9.295542e-01
16       - x8 0.2294200 9.298908e-01
17      - x16 0.2299119 9.591960e-01
18      - x24 0.2304502 9.795701e-01
19      - x23 0.2304744 9.806469e-01
20      - x25 0.2305981 9.846541e-01
21      - x18 0.2309955 9.905178e-01
22      - x14 0.2311176 9.935712e-01
23       - x7 0.2314548 9.958976e-01
24      - x15 0.2322827 9.991981e-01
25      - x11 0.2324040 9.993293e-01
26      - x12 0.2327577 9.996462e-01
```

In the truth, coeffs for `x1, x2, x3, x4, x5` are 1, others are 0.