Conflict Prediction with Spike and Slab prior

Anh Le
Political Science Department
anh.le@duke.edu

Introduction

Being able to predict conflict is very useful for governments and international organizations to deploy their limited resources effectively. Therefore, this project will try to predict the (binary) presence of four events (i.e. insurgency, rebellion, dpc–domestic crisis, and erv–ethnic violence). I use monthly data of 167 countries from 2001 to present (dimension = $27,000 \times 550$). The feature variables include each country's politics, economic performance, financial status, and their 2-month lags.

The project will use logistic regression with spike-and-slab prior to choose a small subset of variables. This will isolate important predictive factors, facilitating model interpretation. Given high correlations between temporally lagged variables, sparse regression will also avoid over-fitting. I will compare the predictive accuracy of this model against others currently used in our lab.

Analysis

I pre-process the data by removing extraneous variables (country names, time ID, etc.) and, for each country in the dataset, add a binary variable that indicates whether an observation belongs to a country. This allows the model to have different intercepts for each country, essentially adding country fixed effect.

I fit the logistic model with spike-and-slab prior using the package `BoomSpikeSlab`. To simultaneously fit four models for my four response variables, I use packages `doMC` and `foreach`.

The model is trained on March2001-June 2013 data and tested on July2013-Sept2014 data. The results below come from a MCMC chain with iterations $= 1000$ and burn-in $= 100$. Even though the MCMC chain is not that long, due to the size of the dataset the computational time is already over 1.5 hour.

Results

Predictive performance

- Table 1 and Table 2 show in-sample and out-sample predictive performance. As expected, in-sample fit is better than out-sample fit across all four events of interest.
- The model can predict *Insurgency* is predicted very well, but performs poorly in predicting *DPC (domestic crisis)*. This is reflected in the ROC-curve shown in Figure 1 and Figure 2.

	insurgency	rebellion	dpc	erv
brier	0.005	0.006	0.042	0.008
auc.C	0.996	0.998	0.918	0.989
precision	0.977	0.959	0.790	0.960
recall	0.763	0.772	0.392	0.681

Table 1: In sample performance

	insurgency	rebellion	dpc	erv
brier	0.009	0.019	0.088	0.034
auc.C	0.997	0.903	0.860	0.977
precision	0.976	0.906	0.642	0.905
recall	0.946	0.784	0.460	0.480

Table 2: Out sample performance

Compare with old models

Table 3 shows the model currently used in our lab. It is a ensemble model of multiple themes (politics, economics, infrastructure, etc.). My new model performs better than this ensemble model across the board.

	insurgency	rebellion	dpc	erv
brier	0.06	0.03	0.12	0.03
auc.C	0.94	0.97	0.78	0.93

Table 3: Out sample ebma (old) performance

Table 4 shows performance benchmark from other labs. Again, my new model performs better than this benchmark, with the exception of recall on *erv (Ethnic Violence)*

	insurgency	rebellion	dpc	erv
precision	0.88	0.84	0.65	0.98
recall	0.78	0.86	0.47	0.71

Table 4: Out sample benchmark (old) performance

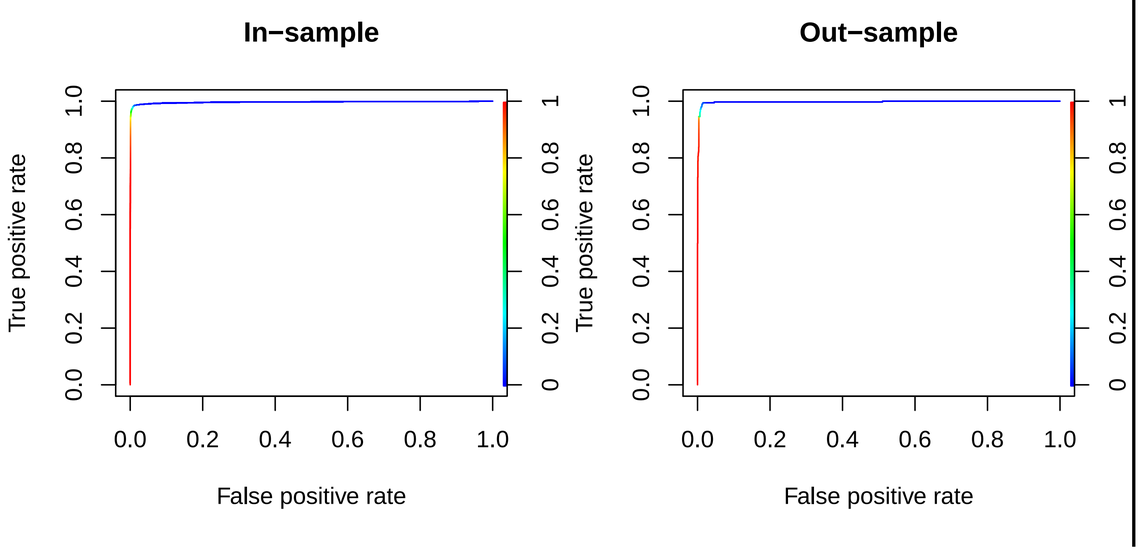

Figure 1: ROC-curve of Insurgency

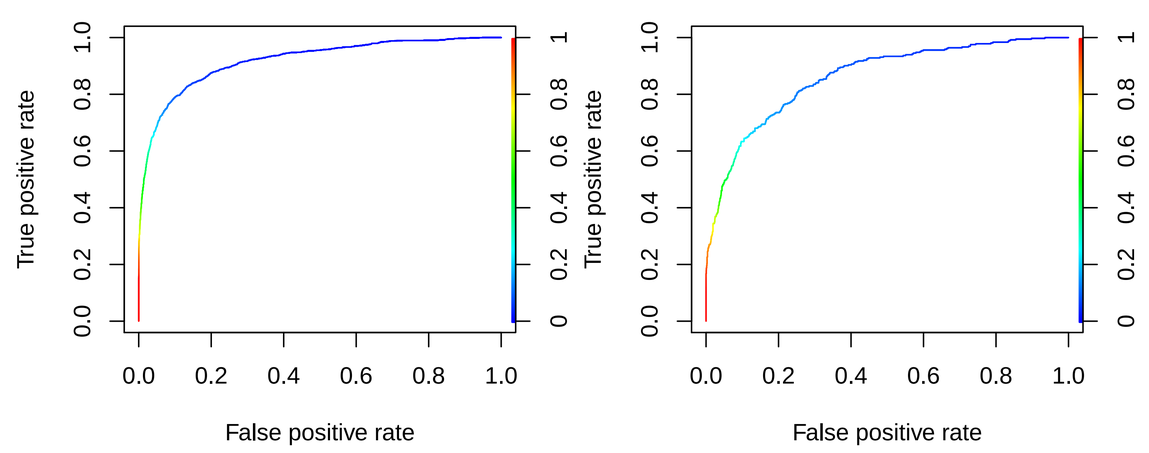

Figure 2: ROC-curve of DPC (Domestic Crisis)

Variable selection

Figure 3 shows the list of variables with positive inclusion probability. From 594 variables in the original dataset, there are only fewer than 40 variables left, achieving model sparsity.

However, we also recognize that most of the included variables are the binary country variable. This means that certain countries almost always (never) experience violence, crisis, etc. Therefore, it is a safe bet to predict that such countries will (not) have those events again.

Besides the country dummies, there are several factors crucial to predicting *insurgency*:

- W.gower.events.rebellion.l1: Spatial-temporal 1 month lag of variable rebellion based on Gower distances (-)
- AG.LND.TOTL.K2.l1: Land area (sq. km), lagged by 1 month (-)
- State.Dept..l1: State Department Political Terror Scale, 1-month lag (+)
- MS.MIL.TOTL.P1.l36: Total military expenditures in 2009 US dollars, lagged by 3 years (-)

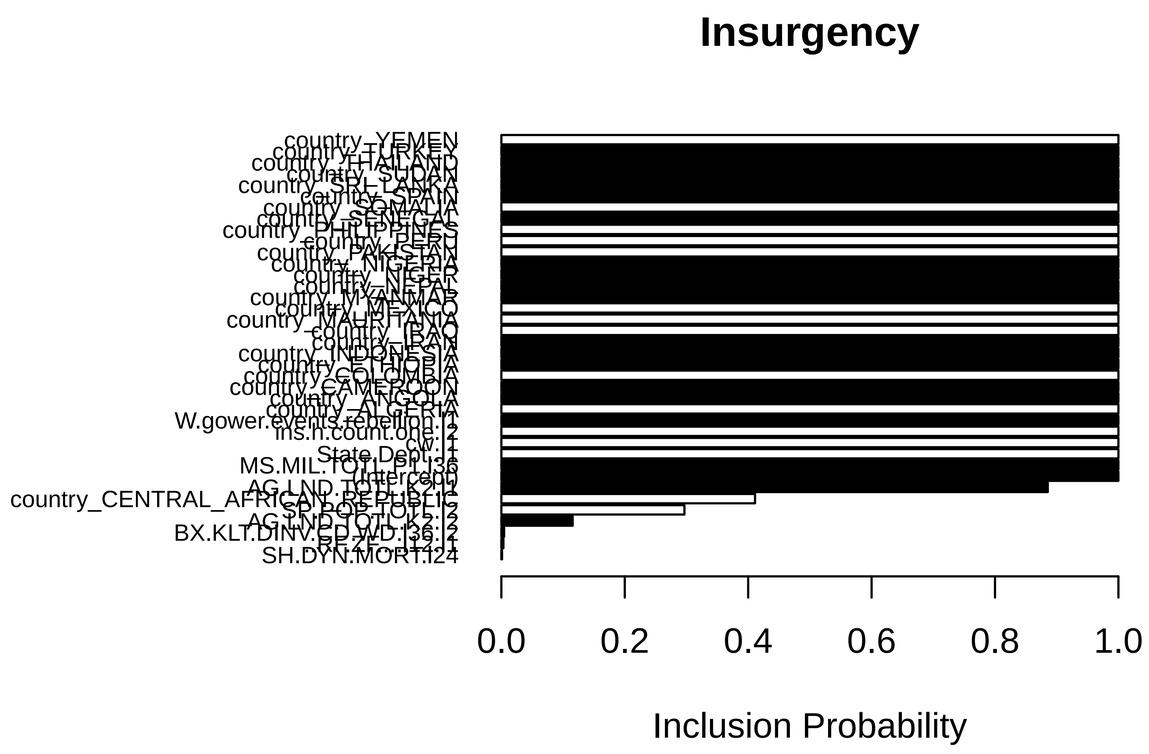

Figure 3: Variables with Positive Inclusion Probability

Conclusions

- Logistic regression with spike-and-slab prior results in a sparse model with good performance. Even though the dataset has many lagged variables that are highly correlated, the spike-and-slab prior does a good job of selecting only one among them.
- Most of the predictive performance comes from country fixed effects. This leads to better results than existing models, but does not give insights into which substantive factors leads to the events of interest. This probably means that the model won't be able to predict drastic change in a country.

Forthcoming Research

- Model without country dummies
- Focus on predicting moments of change