Stats 201B: Final Report

Thomas Merkh, tmerkh@ucla.edu; Juan Diego Vera, jdverajones@ucla.edu

March 21st, 2019

1 Introduction

The forecasting of mass atrocities based on social, political, geographical, and economical predictors has persisted as a challenge to both statisticians and political scientists alike. This investigation's purpose was to improve both the prediction methods for forecasting state sponsored mass killings (SSMKs) in the upcoming year as well as the interpretability of the resulting model. Several challenges present themselves in this type of prediction, ranging from the low base rate of occurrence to the asymmetry of making time sensitive predictions. Since the forecasts are purposed to provide early warning for countries, false predictions warning of a SSMK a year in advance should be treated as more useful than warnings one year late. Furthermore, model interpretation is essential here since the secondary purpose of these forecasts are to identify strong indicators which could warn the world's nations regardless of the model's predictions. In the following, these challenges are addressed by a two-step modeling effort. In the first step, a set of stronger predictors are determined using a sparsity promoting technique to sift through a nonlinear feature space of the original predictors. In the second step, a variety of generalized linear models are used to fit the selected features, and make predictions using the best fitted models.

This paper is organized as follows. Section 2 introduces the working definition of mass atrocities and any relevant background. Section 2.1 briefly overviews the data set, provides a partial overview of the predictors used, and covers data preprocessing. Section 2.2 overviews the most recent work from which this study was motivated, and outlines the model used for comparison against the new feature-selected models. Section 3 covers the underlying mathematics behind the two step method, provides an illustrative example, and briefly covers methods which did not have success. Section 4 applies the aforementioned techniques and addresses their successes and shortcomings. Comparison with previous works and the popular kernel regularized least squares method is provided. Last, Section 5 offers our suggestions for further investigation.

2 Background

In the following, a mass killing is considered having occurred within a country during a given year if the deliberate actions of armed groups, including but not limited to state security forces, rebel armies, and other militias, result in the deaths of at least 1,000 noncombatant civilians targeted as part of a specific group. Such fatalities are deemed deliberate if the perpetrators could have reasonably expected that their premeditated actions would result in widespread death among the targeted group. Examples of such mass killings in the past have been as the result of forced relocation, intentional destruction of health care supplies, forced starvation, forced labor, and more. `The Early Warning Project` has been tracked such events since the conclusion of World War II, and has also made public the relevant geographical, social, political, and economic data measurements which could be used for prediction purposes.

2.1 The Data

A data set consisted of 33 predictors reported on an annual basis, with about 10,000 samples initially available. Each data sample was a broad snapshot of one country's characteristics, such as its population size, geographical region, GDP, government structure, social freedoms, presence of an ongoing mass killing, and so one. Here, a few predictors that were of particular interest are defined for illustration,

- `anymk.ongoing` - A binary variable which is 0 if there is no ongoing SSMK, and 1 otherwise.
- `anymk.ever` - A binary variable which is 0 if the country has never had a SSMK, and 1 otherwise.
- `minorityrule` - A binary variable which is 0 if there is majority rule politically, and 1 otherwise.
- `elf.ethnic` - A real valued variable in $[0, 1]$ called ethnic fractionalization, which measures ethnic heterogeneity.
- `judicialreform` - A categorical variable which is 1 if there was no change to the judiciary's ability to control arbitrary power in a given year, 0 if it was decreased, and 2 if it was increased.
- `religiousfreedom` - A categorical variable in $\{0, 1, 2, 3, 4\}$ where 0 is "essentially no religous freedom", and 4 is full religous freedom.
- `pol_killing_approved` - A binary variable which is 1 if the agents of a state perform political killings without due process, and 0 otherwise.

As one can see, the scaling of each variable is not identical. For this reason, direct interpretation of a fitted model's coefficients is difficult, although not entirely impossible. Although approximately 10,000 samples were initially available, about 30% of them were incomplete

in some way. Due to the limited time constraint of this project, such entries were omitted entirely. Whether or not useful information can be reasonably obtained from partially complete samples has been considered in the past [7], however with little success. After omitting incomplete entries, the preprocessed data set consisted of 7020 samples ranging over 162 countries and 71 years. Last, it is noteworthy to mention that it there are likely spatiotemporal dependencies between the data samples. These dependencies intrinsically contain useful information, but have not been utilized in any study known the the authors to date.

2.2 Previous Works

The idea of forecasting mass atrocities based on a mix country characteristics is not new, but it is relatively contemporary. Perhaps one of the earliest works with this focus was Harff (2003) [2], where a logistic regression on six thought-to-be important country characteristics was performed. The author identified predictors such as political upheaval, prior genocide, current political structure, general economic state, and more as the most relevant predictors based on a mostly heuristic argument. Not all of these suggestions have held up under scrutiny [3]. In 2009, Heldt [4] helped refine the difference between types of prediction, specifically he discussed the differences between risk assessment models and early warning models. The takeaway from this distinction was that early warning models focus more heavily on dynamic short-term factors that increase the likelihood of violence, whereas risk assessment models focus on the big picture and are not necessarily useful for predicting when genocidal violence will take place. On this note, the model discussed here are primarily for early warning purposes and not for general risk assessment purposes. Looking ahead to the current decade, several works arose that improved upon Harff's model, and otherwise drew inspiration from methods used to study political instability. These include but are not limited to Hazlett (2011) [3], Ulfelder (2012, 2013) [6, 7]. Most recently, the Early Warning Project and its collaborators have taken up the task of developing useful models for SSMK forecasting. The latest predictions made were done through the use of a logistic elastic net model performed by C. Hazlett, and these predictions served as the baseline for comparison.

3 Methods

The method attempt at early warning prediction was done using a two step process consisting of *feature selection* (FS) and subsequently fitting the selected data features. This idea was inspired by the work of S. Brunton, *et al.* (2016) [1], where the authors used LASSO to identify the dynamical equations underlying a data set of time-ordered observations. The general idea developed here was that if LASSO is performed on a large *feature library* containing many nonlinear interactions of the original predictors, it will be able to identify the most predictive data features. This technique inspired the authors to utilize the sparsity promoting property of LASSO to select the most predictive data features from a pre-specified library, and then fit an array of models to the selected features. The intuitive appeal of including nonlinear data features can be seen by the following example. Two predictors considered in the study are `reg.SA` - "Country is in South America" and `pol_killing_approved`. It

is possible that when considering only the linear effects of these predictors, the average marginal effects of either one is negligible for predicting SSMKs. However, the presence of both variables being true may hold significant predictive power, which is something that the linear model could not accurately capture. From this, it is clear that the FS method is different from generalized linear models as nonlinear interaction terms could be present in the fitted model, and also differs from the common nonlinear method "kernel regularized least squares" (KRLS) since only a finite set of user specified data features are considered.

For illustration, consider the feature library containing all in linear, pairwise, and third-order data interactions, shown below.

$$\mathcal{X} = \left(\begin{array}{cccc|cccc|cc} \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ X_1 & X_2 & \cdots & X_{33} & X_1X_2 & X_1X_3 & \cdots & X_{32}X_{33} & X_1X_2X_3 & \cdots & X_{31}X_{32}X_{33} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \end{array}\right) \tag{1}$$

In the data set previously discussed, the column vector X_i is length 7020 and consists of one predictor's measurements for all countries over all years. The notation X_iX_j means the element-wise multiplication of the entries of X_i and X_j. The term *3rd order interaction* is occasionally used instead of *cubic nonlinear feature* since for binary variables, the k-th element of $X_iX_jX_l$ is 1 if and only if $X_{ik} = X_{jk} = X_{lk} = 1,$ and is zero otherwise.

Once the user specifies the nonlinear features of the data that he or she believes may act as strong indicators of SSMKs, the next step is to apply LASSO. Recall that the LASSO problem can be written as the optimization problem,

$$\beta^* = argmin_\beta \ \frac{1}{N}||Y - \mathcal{X}\beta||^2 + \lambda||\beta||_1. \tag{2}$$

Alternatively, one may also wish to use the logit LASSO if the task is binary prediction. After fitting the LASSO model, the non-zero entries of the optimal coefficient vector β^* will indicate a subset of the feature library corresponding to the most useful data features. Denote this subset as $\mathcal{X}_{\beta^*}$. The size of $\mathcal{X}_{\beta^*}$ can be made as small as desired by increasing the penalty weight λ in the LASSO step. Alternatively, one may choose λ by doing cross validation, and then live with the resulting size of $\mathcal{X}_{\beta^*}$. Treating $\mathcal{X}_{\beta^*}$ as the only data going forward, one can fit an array of models to $\mathcal{X}_{\beta^*}$. The feature selected models discussed in the following section are the FS-Ridge and FS-Elastic-Net, where both the ridge and elastic net used the logit link function. Additionally for comparison, the KRLS model that was fit to the original data set (non-feature-selected), since this model is both common in social science prediction and is intrinsically nonlinear.

3.1 Failed attempts

Several techniques were attempted with no success. The first method considered was a cluster elastic net proposed by Price, *et al.* (2018) [5]. This method has the advantage of simultaneously determining a natural clustering among the data and performing elastic net regression. This method was primarily unsuccessful because it was too complicated to program in R under the time constraint. The other method initially attempted was to first use a standard clustering algorithm on the data, and incorporate each sample's discovered class as an additional categorical predictor when fitting generalized linear models. Unfortunately, it was found that the computation resources available were unable to perform this clustering in a reasonable amount of time, and this attempt was given up.

4 Results

The predicted one year risks of the FS-Ridge, FS-Elastic-Net, and ordinary KRLS are show in the left table below. For comparison, the predictions of the logit elastic net previously mentioned are included separately in the right table. The feature library used for the presented results contained only the data and its pairwise interaction terms.

2017 Risk Predictions			
Country	FS-Ridge	FS-Elastic-Net	KRLS
DRC	0.2724	0.3050	0.05905
South Sudan	0.1667	0.1562	0.05605
Afghanistan	0.142	0.1362	0.05590
Somalia	0.1134	0.1126	0.09162
Egypt	0.0804	0.1023	0.04824
Chad	0.0729	0.0520	0.05195
Pakistan	0.0709	0.0827	0.05235
Yemen	0.0659	0.0718	0.04855
Angola	0.0639	0.0744	0.07012
Sudan	0.0503	0.0687	0.07603

Logit-EN 2017 Risk Predictions	
Country	Elastic Net
DRC	0.1377
Afghanistan	0.134
Egypt	0.08709
South Sudan	0.08947
Yemen	0.07602
Pakistan	0.07403
Somalia	0.0702
Turkey	0.0702
Angola	0.05646
Sudan	0.05609

The feature selection step for the FS-models was done using cross validation to choose an optimal λ, and in doing so reduced the feature library down to approximately 25 nonlinear features to base predictions on. The KRLS model was essentially unchanged from its default parameters in R. As one can see, KRLS appears to predict somewhat inconsistently in comparison to the generalized linear models. One notable aspect of the feature selected risk predictions is that for the highest predicted at-risk countries, the risk estimates were over twice that of the KRLS and logit elastic net models. The authors have interpreted this occurrence as the feature selected models having discovered hybrid indicators in the data which are stronger at predicting SSMKs, and places like the DRC exhibit such indicators in 2016. This interpretation was motivated by observing the coefficients of the model and the data features they correspond to. The table below shows the largest positive and negative coefficients from the best fitted FS-Elastic-Net model.

Fitted FS-Elastic-Net Coefficients (abbr.)	
Feature	Coefficient (10^{-2})
Political killings & ethnic fractionalization	0.95713
Political killings occur & non SSMK	0.86560
Ever SSMK & South Central Asia	0.70470
Judicial power was enhanced & ongoing MK	0.70048
Minority in control & successful coup	0.38293
In Africa & log battle related deaths	0.32372
SSMK ever & ethnic fractionalization	0.27303
Trade openness & South Central Asia	-0.2319
Freedom of men to move	-0.1824
Even civil rights	-0.0311

As previously stated, it appears that the model is basing its predictions primarily off of pairwise interaction terms, although the magnitude of the coefficients are not fully descriptive due to the various predictor scales, see Section 2.1. The interpretation of the selected features and the sign of their fitted coefficients respectively is significant however. Specifically, customarily beneficial characteristics for reducing the risk of mass atrocities, such as equal civil rights, freedom for the civilian population to move, and trade openness, all have negative coefficients. On the contrary, possibly significant indicators such as a successful coup in the previous 5 years and a minority in control of the government, exhibit positive coefficients and therefore increase the risk of SSMKs. The authors believe that the LASSO step for selecting relevant predictor interactions is useful in itself for finding interpretable indicators for SSMKs, and should be subject to future investigation.

One desirable feature to work into early warning models is to have a model which penalizes certain types of errors more strongly than others. For example, a model which assigns small risks where it should assign large risks (e.g. false negatives) should be considered less useful than a model which assigns higher risks when little risk is truly present (e.g. false positives). Another example could be a model which predicts a high risk a year too early should be considered useful, whereas a model that predicts a high risk a year too late would need improvement. These considerations were the next topic of focus for the project. Although the authors were unable to fully attempt to incorporate such considerations into the trained model, the number of false negatives and false positives were measured for each of the fitted generalized linear models. Here, a false negative is consider to be when a risk of $< 5\%$ was assigned to a country during a year when a mass atrocity occurred, and a false positive was when over 5% risk was assigned an no mass killing occurred. For the best fitted logit-elastic-net, it had 66 false negatives and 202 false positives on the training set. The best fitted FS-Ridge has 67 false negatives and 229 false positives, and the FS-elastic-net has 64 false negatives and 265 false positives. Overall, these numbers fared well in comparison to KRLS which had 142 false negatives and 302 false positives.

Last, as suggested during the presentation, the ROC curves were plotted for each model to the best of our ability and the AUC was measured for each. See Fig. 1. Here, the KRLS-logit model found in the R package `KRLS2` was used in place of the previous KRLS model. The AUCs for each model were: FS-Ridge 0.72, FS-Elastic-Net 0.71, Logit-Elastic-Net 0.76, and KRLS-Logit 0.49.

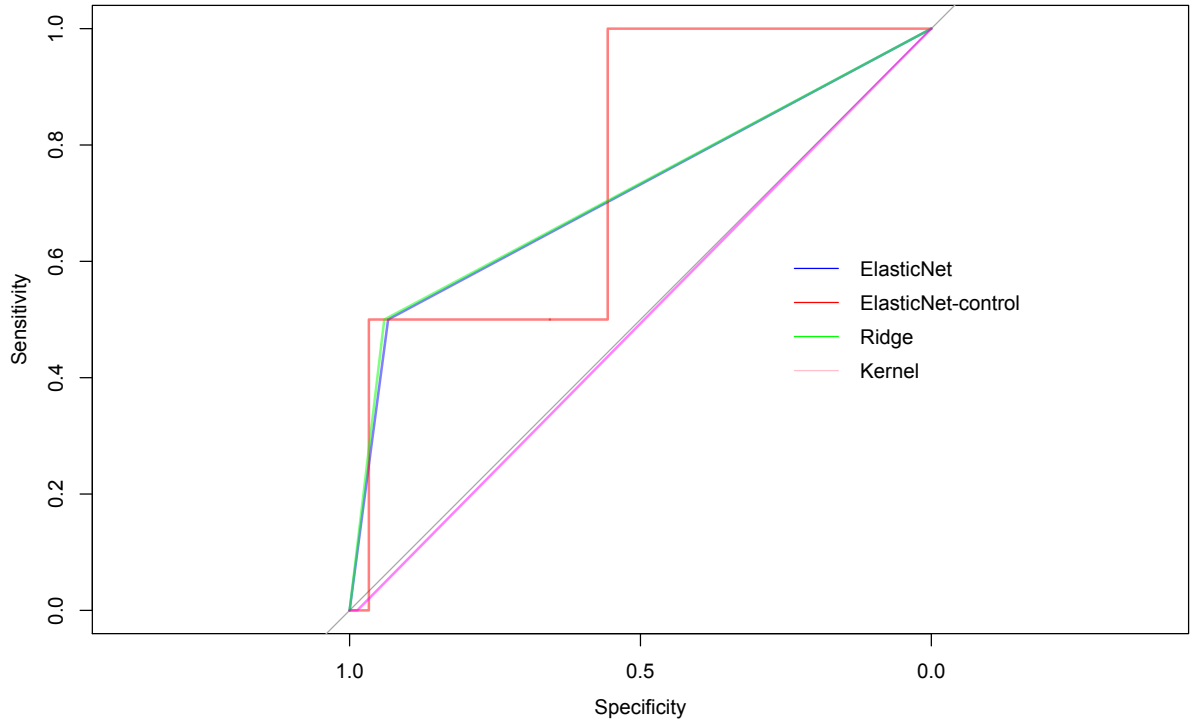

Figure 1: The ROC curve for each of the models. The KRLS model here underperformed randomly guessing by a slight margin.

5 Future work

This project has only scratched the surface on many interesting and important challenges. Perhaps the most immediately feasible addition to this work could to use an improved loss function when training the models, such as one that penalizes false negatives more harshly than false positives. Another area of this work that is in need to attention is in quantifying model performance. Originally, a less-than-useful performance measure (0-1-100) was suggested in class. This performance measure had obvious flaws which would have ranked an entirely nonsense model similar to the observed performance of the best FS-Elastic-Net model. Instead of this, we believe that measuring the area under the ROC curve could serve as a reasonable method for comparing models. Last out of the low hanging fruit, one could obtain a larger computing resource and perform clustering beforehand, using the assigned clusters as new predictors in the generalized linear models. This method could possibly lead to interesting interpretations depending on how the clustered algorithm converges for the data set. In addition to these, there exist possibilities which require a wider expertise and larger time commitment. One idea for example is to utilize state-of-the-art data completion methods from the deep learning community to fill in the partial data samples that were omitted from our study. Another fun idea would be to use the bootstrap PCA idea presented by the other students to generate additional data, and compare the performance of neural networks to the performance of generalized linear models.

References

[1] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

[2] Barbara Harff. No lessons learned from the holocaust? assessing risks of genocide and political mass murder since 1955. *American Political Science Review*, 97(1):57–73, 2003.

[3] Chad Hazlett. New lessons learned? improving genocide and politicide forecasting. *United States Holocaust Memorial Museum*, 2011.

[4] Birger Heldt. Risks, early warning and management of mass atrocities and genocide: Insights from research. 2009.

[5] Bradley S Price and Ben Sherwood. A cluster elastic net for multivariate regression. *Journal of Machine Learning Research*, 18(232):1–39, 2018.

[6] Jay Ulfelder. Forecasting onsets of mass killing. 2012.

[7] Jay Ulfelder. A multimodel ensemble for forecasting onsets of state-sponsored mass killing. 2013.