FINDING THE RIGHT CUMULATIVE LAG FOR ENVIRONMENT EFFECTS ON INDIVIDUAL WELLBEING - An application of lasso-based feature extraction for linear mixed effect models

Final project report as part of course fulfillment for

STAT 574 - Statistical Machine Learning

Instructor: Hao Helen Zhang

By:

Karthik Srinivasan

MIS department, Eller College of Management
University of Arizona

December 10, 2016

1 INTRODUCTION

Regularization techniques such as the least absolute shrinkage and selection operator (LASSO) have embedded feature selection properties. Coefficients of features that are less relevant are penalized and set to zero to reduce the variance and improve the prediction accuracy of statistical models. In this study, we apply the LASSO approach to propose a new method for feature extraction in large-scale multilevel longitudinal data.

Feature extraction involves reducing the amount of resources required to describe a large set of data or representing features in an optimal manner. During explanatory modeling of longitudinal data, expert inputs or prior theory is normally used to choose the best alternative among range of feature transformations (such as logarithm, exponential, moving average, absolute lags, cumulative lags, etc). Lack of theory or prior knowledge for selecting one transformation over another can be resolved using statistical methods. The underlying criteria for such methods are metrics related to model fit and prediction accuracy. We propose a novel approach to extract features for multilevel longitudinal models. We use a real dataset created in an ongoing research experiment to demonstrate the utility of our proposed approach.

1.1 A brief summary of the ongoing research experiment

It has been observed that environment has an impact on people's emotional and physiological well-being. A multi-disciplinary study carried out in an office building to understand the relationship between indoor environment conditions and an individual's wellbeing. The experimental setup consists of participants wearing two sensors on their chest for three days while carrying out their day to day activities: (a) A heart rate monitor and (b) An environment quality sensor. More than 200 participants across multiple locations participated in the experiment and the data granularity is 'per minute' interval. The data for this study is hierarchical, where participants are the secondary level of data abstraction. Multilevel explanatory models (such as linear random effects model) can be fit for such data to model correlational evidence between the several inputs and outcome, which is instantaneous physiological stress response. SDNN, RMSSD, normalized HF and LF/HF ratio are the four heart rate variability (HRV) measures [1] used as outcomes for four independent univariate models. These outcomes characterize different aspects of instantaneous physiological stress response in this study. One of the research problem in this study is to understand the correlational relationship of these physiological outcome variables with five indoor environment factors, namely CO2, Sound, Temperature, Relative-humidity and Light intensity.

From a statistical modeling point of view, we are posed with several challenges while addressing the above research question. One among them is how to characterize the functional relationship of the environment factors and instantaneous stress response. That is, do these inputs have an instantaneous effect on stress response? Does their variability modulate human stress over time? These are impending questions which we try to answer in this study. We explore the regularization approach to address the above question. The research problem is formulated as a feature extraction problem where we have a set of inputs and want to extract features that optimally represent input-output functional relationship pairs.

1.2 Relevant literature

Feature extraction has different connotations for different problems, such as dimension reduction for problems with high-dimension or sparse data setting to feature engineering for enhanced predictive modeling. Our problem is related to the latter aspect where we have a feasible dimensional setup ($p << n$) and a one to many mapping of input to feature transformations as discussed in the previous section. Without starting with initial set of features, one could use streamlined procedures such as lagged effect analysis [2] , Box-Cox transformations [3]. However, their application is very specific (e.g. Box-Cox transformation over inputs is used to transform individual inputs to get normalized model residuals and lagged effect models identify set of absolute lags of input that are optimally related to the outcome. There are no existing approaches for identifying optimal cumulative lags for inputs, other than supervised feature selection to the best of our knowledge.

The traditional stepwise feature selection procedures [4] are ridden with challenges such as sensitivity to changes in data and low external validity [5]. Embedded feature selection such as the lasso approach cannot be directly used for multilevel models due to the hierarchically clustered nature of the data. Specialized embedded feature selection methods have been proposed for multilevel statistical models [6]. However, our feature extraction problem entails selection one feature among features that are functional representations of an input instead of determining the optimal feature set in the input space. That is, we do not have control over the feature selection procedure, as to select at least one out of k different representations of an input. This is similar to the logic of grouped lasso where groups are features are intended to be retained. But here we want to do feature selection within each group of features, as a feature extraction procedure.

2 Formulation

Given a mixed model of the form

$$y = X\beta + Zu + \epsilon$$

where, Y is the outcome vector,
β is fixed effects coefficients matrix for input matrix X,
u is random effects for input matrix Z, and
ϵ is the residual distribution

We define the problem as selecting a single optimal variable x_i from set of k transformations of an input variable $\{x_1, x_2, x_3, ..., x_k\} \in X$. The set of k transformation as features representing the same input phenomenon are assumed to be given in the problem. For example, the input variable *temperature* can be represented as x_1 = Instantaneous temperature level, x_2 = previous 30 minutes averaged temperature, x_3 = previous 60 minutes averaged temperature, and so on.

We propose a three step procedure to determine the optimal feature representing each input in a multilevel problem. First, fit lasso-regularized linear regression model to each of the n groups in the multilevel data. For now, we assume a 2-level dataset, but our approach can be extend to scenarios with more than 2 levels. Secondly, we combine the co-efficients using a weighted pooling strategy to get overall importance score of each feature for feature subset of each input. We then select the feature with highest score as an optimal representation for each input.

Mathematically, we can represent the steps as follows:

Step I: Fit lasso models for each of the M group data

$$L_m : \min_{\beta}(\sum_{i=1}^{n}(y_i^{(m)} - \sum_{j=1}^{d} x_{ij}^{(m)}\beta_j)^2 + \lambda \sum_{i=1}^{d} |\beta_j| \tag{2.1}$$

$$(m \in 1, .., M)$$

The coefficients for each element in the feature subsets $\{x_1, x_2, ..., x_k\}^r \subseteq X$ for inputs $r \in R$ is therefore determined for each group.

Step II: Generate the feature importance score as a weighted sum of each lasso coefficient

$$\beta_j = \frac{1}{n} \sum_{m=1}^{M} \beta_j^{(m)} W^{(m)}$$

where,

$$W^{(m)} = \frac{S(m)}{L(\hat{Y}(m))^{0.5}} \tag{2.2}$$

$S(m)$ and $L(\hat{Y}(m))$ are size and loss function for linear model fit for group m respectively.

Step III: Determine the optimal feature corresponding to each input r as:

$$\max_{x_j^r \subset X^r} \beta_j^r \tag{2.3}$$

Step I is fitting standard lasso models for M groups or clusters in the multilevel data. Here, we circumvent the need to account for within group dependence [7]. For multilevel data with more than 2 levels, we propose a MxN factoring approach (e.g. with four groups in level 2 and 10 groups in level 3, we can implement step one for 40 factored groups). There is scope for improvement over this proposed strategy and we do not delve deeper into the multilevel generalization in this current study. In Step II, we propose the weights $W^{(m)}$ as a ratio of cardinality of each group $S(m)$ and square root of loss function $L(\hat{Y}(m))$ of the model for corresponding group. This ensures that the $beta^(m)$ for model corresponding to group m is penalized for smaller representation as well as inferior fit, compared to coefficients of models for other groups in the dataset. Step III simply selects the best feature from subset of functional representations of each input X^r, as the final step of the feature extraction process.

3 Analysis

We use the environment wellbeing dataset for analysis and model comparison. Data collection, integration, preprocessing and cleaning has been done and the final dataset contains around 200,000 minutes of heart rate monitor and environment quality data streams. The dataset is randomly split into training and test dataset (75:25 split) for analysis.

After considering different values of cumulative lags for inputs temperature, sound, CO2, Relative humidity and light, we consider three possible transformations (instantaneous value, 30 minutes cumulative lag and 60 minutes cumulative lag) as three functional transformations for these inputs or feature candidate sets. We then apply the proposed feature extraction approach to select the optimal feature corresponding to each input.

We compare the performance of model with proposed feature set *(denoted as Mixed lasso)* to models having following feature sets: (a) Instantaneous inputs (b) 30 minutes cumulative lagged inputs, (c) 60 minutes cumulative lagged inputs, (d) Instantaneous, 30 minutes and 60 minutes cumulative lagged versions of inputs, (e) Supervised stepwise feature selection approach using AIC. Covariates such as *Time of day, Day of week, Age/BMI of participant* are included in each model. Model fit compared using Akaike Information Criteria (AIC) and pseudo R Squared for random effects model. Predictive performance compared using Root Mean Squared error (RMSE) and Mean Absolute Error (MAE) on a test dataset.

4 Results

Using our proposed approach, 60 minutes cumulative lagged versions of for Temperature, CO2, Pressure, Relative humidity and Light and instantaneous version of Sound are identified as optimal feature representations for the environment factors across all four HRV outcomes (SDNN, RMSSD, norm HF and LF/HF).

The model fit and prediction accuracy comparison results across models described in previous section are shown in tables 4.1,4.2 [8], 4.3 and 4.4. The model fit and error estimates for best performing models are highlighted for reader's convenience.

Table 4.1: Model fit comparison using AIC

Models	RMSSD	SDNN	norm. HF	LF/HF
Fixed effects only (baseline)	173992.9	211033	175931.4	139254.5
Instantaneous	170094.9	210157.5	174400.1	138544.7
30 mins cumulative lag	169917.3	210348.7	174668.6	138695.9
60 mins cumulative lag	169562.9	210222.4	174622.2	138704.1
All cumulative lag	170092.5	**210112.3**	**174335.1**	138538.2
MinAIC	**169461.7**	210155.7	174568	**138454.9**
Mixed lasso (cumulative lag)	169764.4	210171.1	174400.8	138700.4

Better model fit and prediction accuracy is determined by lower value in tables 4.1 [8], 4.3 and 4.4 and higher value in Table 4.2 showing pseudo R-squared [8]. We see that even though model fit is not always optimal for the proposed mixed lasso approach, the performance is at least second best in terms of prediction accuracy.

Table 4.2: Model fit comparison using Pseudo R-squared

Models	**RMSSD**	**SDNN**	**norm. HF**	**LF/HF**
Fixed effects only (baseline)	0.6724	0.6643	0.5152	0.4544
Instantaneous	0.7735	**0.6988**	0.5968	0.4994
30 mins cumulative lag	0.7801	0.6936	0.5896	0.4971
60 mins cumulative lag	**0.7882**	0.6969	0.5938	0.498
All cumulative lag	0.7734	0.6975	0.5986	0.5007
MinAIC	0.786	0.697	0.5952	0.5042
Mixed lasso (cumulative lag)	0.7856	0.6987	**0.5995**	**0.5011**

Table 4.3: Prediction accuracy comparison using RMSE

Models	**RMSSD**	**SDNN**	**norm. HF**	**LF/HF**
Fixed effects only (baseline)	8.5419	17.9597	9.1037	6.5198
Instantaneous	7.6825	17.4308	8.6555	6.4416
30 mins cumulative lag	7.5697	17.4366	8.7344	6.4454
60 mins cumulative lag	7.5225	17.3871	8.7064	6.4508
All cumulative lag	7.6769	17.4306	8.6441	6.4413
MinAIC	7.5293	17.4307	8.7071	**6.4372**
Mixed lasso (cumulative lag)	**7.5131**	**17.3448**	**8.6363**	6.4419

Table 4.4: Prediction accuracy comparison using MAE

Models	**RMSSD**	**SDNN**	**norm. HF**	**LF/HF**
Fixed effects only (baseline)	5.8292	12.4376	6.3598	2.4438
Instantaneous	5.1534	11.9571	6.0013	2.344
30 mins cumulative lag	5.0853	11.9575	6.063	2.3474
60 mins cumulative lag	5.0664	11.9048	6.0488	2.3437
All cumulative lag	5.1533	11.9535	**5.9953**	2.3423
MinAIC	**5.0610**	11.957	6.0386	**2.3341**
Mixed lasso (cumulative lag)	5.0636	**11.8837**	6.0007	2.3347

5 DISCUSSION

We propose an novel method for identifying best feature for each input in a multilevel models setting. Our method uses a pooled regularization strategy, where we demonstrate its utility for identifying best cumulative lags for inputs of a problem with multilevel longitudinal data as an example. However, our method address the general set of problems of selecting one out of k transformations of inputs of multilevel models, when there is no prior theory for the input to output functional relationship.

The proposed lasso based approach in this study for feature extraction for multilevel longitudinal models has better performance in terms of prediction accuracy. It is not sensitive to noise in data or manual error as stepwise/manual approaches. This study therefore contributes to statistical modeling research community in a modest way with the proposed three step feature extraction approach. The application to environment wellbeing problem contributes to the domain literature, by suggesting a 60 minute cumulative lag effect for four out of the five inputs of interest.

In addition to the linear multilevel regression models, we also tested the feature extraction method using the RE-EM tree, a tree-based multilevel model fitting approach [9]. The models with all positive feature representations of inputs (full model) performed much better than any other model, included the model fit using proposed approach. Hence, we conclude that our approach is more suitable to explanatory statistical modeling than tree-based models.

6 Conclusion

This study was a result of data analysis for ongoing experiment investigating effect of indoor environment on individual wellbeing. We posed the question of which cumulative effect optimally characterizes each of the five environment factors - temperature, sound, CO2, relative humidity and light in a multilevel longitudinal model capturing environment relationship with instantaneous stress. A lasso based approach was used for the above feature extraction problem to determine best set of features. Our approach not only introduces a robust and efficient approach to solve the feature extraction problem, but also shows better prediction accuracy for the given data.

Our approach has a direct application in feature extraction for multilevel longitudinal analysis where prior theory for functional transformation of inputs to features is non-existent. With the rising number of applications with multilevel longitudinal setup such as that involving heterogeneous streaming sensors, our approach can be useful to improve prediction performance of explanatory models as well as enhance the literature pertaining to the domain about functional relationships.

References

[1] B. Xhyheri, O. Manfrini, M. Mazzolini, C. Pizzi, and R. Bugiardini, “Heart Rate Variability Today,” *Progress in Cardiovascular Diseases*, vol. 55, no. 3, pp. 321–331, 2012.

[2] A. Gasparrini, “Distributed Lag Linear and Non-Linear Models in R: The Package dlnm,” *Journal of statistical software*, vol. 43, no. 8, pp. 1–20, 2011.

[3] R. M. Sakia, “The Box-Cox Transformation Technique: A Review,” *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 41, no. 2, pp. 169–178, 1992.

[4] P. MacNaughton, J. Spengler, J. Vallarino, S. Santanam, U. Satish, and J. Allen, “Environmental perceptions and health before and after relocation to a green building,” *Building and Environment*, vol. 104, pp. 138–144, 2016.

[5] J. Hastie, Trevor, Tibshirani, Robert, Friedman, *The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition.* 2009.

[6] S. Müller, J. L. Scealy, and A. H. Welsh, “Model Selection in Linear Mixed Models,” *Statistical Science*, vol. 28, no. 2, pp. 135–167, 2013.

[7] J. D. Singer and J. B. Willett, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* 2009.

[8] S. Nakagawa and H. Schielzeth, “A general and simple method for obtaining Rˆ2 from generalized linear mixed-effects models,” *Methods in Ecology and Evolution*, vol. 4, pp. 133–142, 2012.

[9] R. J. Sela and J. S. Simonoff, “RE-EM trees: A data mining approach for longitudinal and clustered data,” *Machine Learning*, vol. 86, no. 2, pp. 169–207, 2012.