7-2010

Shrinking Symbolic Regression Over Medical and Physiological Signals

Jamie Macbeth
University of California, Los Angeles, jmacbeth@smith.edu

Majid Sarrafzadeh
University of California, Los Angeles

Shrinking Symbolic Regression Over Medical and Physiological Signals

Jamie Macbeth
Department of Computer Science
University of California
Los Angeles California USA
e-mail: macbeth@cs.ucla.edu

Majid Sarrafzadeh
Department of Computer Science
University of California
Los Angeles California USA
e-mail: majid@cs.ucla.edu

***Abstract*—Medical embedded systems of the present and future are recording vast sets of data related to medical conditions and physiology. Linear modeling techniques are proposed as a means to help explain relationships between two or more medical or physiological signal measurements from the same human subject. In this paper a statistical regression algorithm is explored for use in medical monitoring, telehealth, and medical research applications.**

An essential element in applying linear modeling to physiological data is determining functional forms for the predictor signals. In this paper we demonstrate an efficient method for symbolic regression and model selection among possible transformation functions for the predictor variables. The three-stage method uses LASSO shrinkage regression to select a brief functional form and performs an polynomial lag regression with this form.

This method is applied to medical and physiological time series data exploring the link between respiration and blood oxygen saturation percentage in sleep apnea patients. We found that our method for selecting a functional transformation of the predictor variable dramatically improved the goodness of fit of the model according to standard analysis of variance measures. In the dataset examined, the model achieved a multiple R^2 of 0.3373, while a plain time-lagged model without transformation or polynomial lags had a R^2 of only 0.016. All of the variables in the model produced by the algorithm had high scores in t tests for validity.

***Index Terms*—Biomedical Modelling; Cardiovascular Modelling; Time Series Analysis; Respiratory Mechanics.**

I. Introduction

The volume of data that is to be collected by medical monitoring systems of the present and future is overwhelming. Data to be collected may be single or multi-channeled and different datasets may have different sampling rates, signal-to-noise ratios, and various signal characteristics. Additionally, data is collected using a variety of health sensors and diagnostic devices in different environments.

As a result there is a wide-ranging interest in systems for human-trained and automatic classification of physiological signals. Our interest in the current work concerns the establishment of a relationship between one medical signal or parameter and one or more others; These physiological quantities are modeled as variables in a linear model. The procedure for statistically fitting data observations to a model to discover correlations between the quantities is known as *regression*. The system and algorithms under discussion perform efficient linear model regressions for correlation studies and for prediction to aid in clinical research and health care environments. These systems perform pattern matching and learn signal patterns in data that may represent the onset or degree of the medical condition or phenomenon in question.

Embedded systems in the medical monitoring domain are able to collect huge amounts of data at a very high resolution from multiple locations at very low cost [15] [18]. It has become a very important and fruitful area of study for wireless embedded sensor networks. These systems offer early detection of physical ailments and can enhance the doctor and patient relationship by offering remote diagnoses. They can help to enhance the expertise of trained health care professionals, and provide tools for investigative efforts to cure chronic illnesses. Systems also exist for emergency medical response to catastrophic events like earthquakes, typhoons or disease epidemics. These systems are flexible in the way that scientists can reprogram or re-task them after deployment in the field.

II. Related Work

Similar work has been done on on processing, organizing and categorizing medical and physiological signals and time series. Activity detection studies attempt to classify physical activities that the subjects are performing purely through the physiological signals recorded. Bao and Intille [3] perform activity recognition from acceleration data using several classification methods. Motif finding attempts to find previously known or unknown patterns in time series databases [12] and motifs are useful for activity detection in embedded sensing medical systems [20]. Probabilistic discovery of motifs is also possible [4]. Oates, et al. study clustering of signals for robotics [16]. As another form of activity detection, the Smartfall system attempts to detect falls through the use of accelerometers and gyroscopic sensors embedded in a cane [11] [2].

Some studies focus on medical monitoring of vital health signs such as heart rate, blood pressure and EKG. In the field of body sensor networks, in-vivo monitoring, fitness, and athletics are studied. Mobile ad-hoc sensor networks have also been explored for medical emergency response and triage [8].

Some well known work in the statistical literature has used health and physiological data to test new and general multiple regression algorithms. Efron, Hastie, Johnston and Tibshirani use data from a diabetes study to generate a prediction model in their efforts to study least angle regression [7]. However, to our knowledge, our use of time-lagged regression to study physiological signal data is the first of its kind.

III. Background

Our interest in the current work concerns the relationship between an independent variable and one or more dependent variables; the purpose of experiments involving the variables being to assess the effects of variations in the independent variable on the dependent variable as a response measure. In medical monitoring studies or applications, one obtains measurements on two or more variables through data collected simultaneously on a single subject. We are interested in knowing whether or not the variables go together or covary. Studies of this kind are *correlational* in that they attempt to determine whether or not two variables influence each other, and regression measures and estimates the strength and direction of these relationships. Often times the methods are not used in fully controlled experiments where the independent variables are explicitly chosen, and random sampling is used to eliminate bias.

in medical embedded sensing systems many types of measurements may be treated as independent variables. For example, accelerometry or gyroscopic sensors may record data related to specific actions or motions of the subject, or may record the subject's general activity. The data can be studied with blood pressure, blood oximetry, or heart rate as the dependent variable for studies in exercise, training and physical fitness. This data may then be chosen for a correlational study with a nervous system or muscular system disorder.

In typical physiological studies, signals of interest may be sampled at a far higher rate than the rate in which they influence each other, and they may be sampled at different rates than each other. For example, typical range from 100 to 250 Hz or more while weight scale data for human subjects in a typical study may be in the micro-Hertz range at one or two samples per day. This creates a multiple orders-of-magnitude difference in sample rate between the data sources. Additionally, the time scales under which signals influence each other may not be known, and the functional form under which the relationship is modeled is important to the success of regression techniques. We propose efficient algorithms for dynamic time lag regression over model selection for use in physiological studies.

When data on variables is highly interrelated and observed over time, individuals, or space, econometrics models and methods are indispensable [14]. Relationships between measurements of physiological quantities would tend to be dynamic, in the sense that variations in an independent variable may take time to impact a dependent variable, and the impact may be long-lived. Techniques for dynamic time series models are well known in econometrics.

A. Model Selection

The availability of many possible predictors to choose from to perform a regression precipitates problems in linear model selection. Models can usually benefit from having less predictor variables—the estimated true validity of a sample multiple regression is very low when the number of predictor variables is large in relation to the number of observations [5].

Reducing the size of the set of predictor variables also pursues the definition of a model with fewer explanatory factors. In many research and clinical applications, simple explanations and rules of thumb are desired to help understand parts of complex phenomena. On the other hand, we need to choose enough predictor variables in order to get a reliable fit to the data. Including too few variables and making the model overly simplistic may ignore factors and predictors that are important to explaining the phenomena.

Many procedures have been proposed for model selection. The *all subsets* algorithm performs regressions with all 2^p possible linear models given p predictors to choose from. *Stepwise regression* adds parameters to the model one by one according to certain criteria. *Backward elimination* performs the opposite; it starts with a regression involving all available variables and selectively removes variables based on certain criteria. Both stepwise regression and backward elimination have stopping criteria under which the process completes with a certain number of the available parameters. Draper and Smith [6] and Weisberg [21] provide useful surveys of the details of the inner workings of these methods.

In linear models, transformations of predictor variables through functions such as log, square-root or polynomial functions are allowed as predictor variables in the model. "Dummy" encodings of categorical variables as quantitative variables are also allowed. The selection of appropriate transformations and representations of the various predictor variables comprise another form of model selection. Symbolic regression and system identification [13] techniques have been proposed and used for the purpose of discovering models to explain complicated financial data. Often symbolic regression techniques involve forms of genetic programming [10].

IV. Formalization

A. Multiple Regression

Let Y represent a dependent or criterion variable, and X_1, X_2, X_3 ... X_n represent independent or predictor variables of Y. We will consider cases where observations of values for any given variable form a continuous, totally-ordered set. An observation of Y coupled with observations of the independent variables X_i is a run of the experiment, often called a *case*.

In experimental runs, score values of these variables are observed from a *population*. We assume that any dataset we use is a *sample* from a population as larger group. Multiple regression methods will attempt to derive or calculate a constant β_0 and a set of weights, β_1, β_2, β_3, ... β_n for the predictor variables. In the equation

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n + \epsilon,$$

$\hat{Y}$ is then used to predict the observations of Y given the observations of the X_i

The β_i are called correlation coefficients, and ϵ is the uncorrelated error or disturbance. Regression fits the values from a set of observations to the model by estimating the correlation coefficients. Typically the coefficients are chosen

so that $\hat{Y}$ predicts Y with a minimum sum of squared errors for the sample. The model can be written as a summation

$$\hat{Y} = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon. \quad (1)$$

B. Time Series

Regression will be used to predict time series values of the dependent variable Y based on time series data of the independent variable X. Ideally, time series data for X will be sampled at regular intervals and will be represented by the X_i. Time series data for the dependent variable Y need not be sampled regularly. Observations of Y_i and X_i will be made over a time period $0 < t < T$. Causality is assumed, and if Y_t exists, X_t, X_{t-1}, X_{t-2}, X_{t-3}, ... X_0 can be used in a multiple regression to predict it.

The X_i predictor variables of Y used in the model represent observations made periodically during a continuous time period beginning at some time before Y was observed and ending at the time of observation of Y. Models of this kind are known as *distributed lag models*, and are useful when changes in the independent variable X have an effect on the value of Y over many samples of Y. Because two variables are involved, this is called a *bivariate distributed lag model*. Typically, if X and Y are observed at identical periods at the same frequency, T bivariate observations will be made of Y_t and X_t. We will restrict our set of predictor variables for Y_t to n values of the time series in X represented by X_{t-1}, X_{t-2}, X_{t-3}, ... X_{t-n}. The model can be succinctly written

$$\hat{Y_t} = \beta_0 + \sum_{i=1}^{n} \beta_i X_{t-i} + \epsilon. \quad (2)$$

C. Analysis of Variance

R^2, a scale-free measure representing the percentage of the variance in the data that is explained by the model, is a typical measure of the accuracy of the regression,

$$R^2 = \frac{E[(\hat{Y} - E[Y])^2]}{E[(Y - E[Y])^2]}.$$

The numerator is the "model" sum of squared differences between the value of Y predicted by the model and the value of Y actually seen in each observation. The denominator is the "total" sum of squared differences between observations of Y and the mean of Y. This is a biased estimator of the true value of R^2 in the population, but we assume that there are enough observations to overcome this bias.

The greater the value of R^2, the greater the goodness of fit of the model. As is typically done, we use R^2 as an objective in automated model selection problems and their respective algorithms.

D. Polynomial Distributed Lag Models

One of the main difficulties with regression using the equation above is that we cannot get reliable estimates of the parameters β_i due to the high correlations between values of the predictor signal close to each other in time. Almon [1] studied the case where the lagged coefficients β_i decline according to a polynomial of degree r in i. If the degree of the polynomial r is 2, for instance, we write the following equation for the β_i where $i > 0$,

$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2$$

If this expression for the β_i is substituted into the distributed lag model above we get

$$\hat{Y_t} = \beta_0 + \sum_{i=1}^{n} (\alpha_0 + \alpha_1 i + \alpha_2 i^2) X_{t-i} + \epsilon.$$

We then substitute new predictor variables Z_{0t}, Z_{1t}, Z_{2t} where

$$Z_{0t} = \sum_{i=1}^{n} X_{t-i}$$

$$Z_{1t} = \sum_{i=1}^{n} i X_{t-i}$$

$$Z_{2t} = \sum_{i=1}^{n} i^2 X_{t-i}$$

and the model is rewritten

$$\hat{Y_t} = \beta_0 + \alpha_0 Z_{0t} + \alpha_1 Z_{1t} + \alpha_2 Z_{2t} + \epsilon. \quad (3)$$

Instead of regressing Y_t on the X_{t-i} we use Z_{0t}, Z_{1t} and Z_{2t}. We indirectly estimate the β_i by obtaining estimates for α_0, α_1 and α_2.

V. Symbolic Model Selection

Another obstacle is choosing the symbolic representation of the predictor variables in the model. If a simple model with no transformation of the predictor variables (equation (2)) is poor in representing the variance in the data, a more sophisticated symbolic functional form for the predictor data may be needed. Later we shall exhibit an example with high levels of confidence for many of the predictor variables in the model coupled with very low values for R^2. We now explore some simple algorithms for seeking out and discovering a symbolic functional form that better captures the behavior present in the data.

We propose the use of the LASSO method for multiple regression as a means for symbolic regression-style selection of the functional form of the linear model. LASSO [19] is a "shrinkage" variable selection method for linear regression, meaning that the method shrinks or reduces to zero some coefficients for prediction accuracy and for interpretation purposes. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients:

$$\sum_{1>j>n} |\beta_j| \leq \tau$$

where $\tau \geq 0$ is a tuning parameter which controls the amount of shrinkage which is applied to the estimates. Note that the intercept β_0 is not included.

An efficient implementation of LASSO multiple regression is provided by the LARS Least Angle Regression package for the R software environment for statistical computing [17]. The LARS algorithm, which computes coefficients for LASSO regression with a minor modification, is described in [7]. The LARS package can generate coefficients for all values of the tuning parameter τ.

VI. A Shrinking Symbolic Regression Algorithm

We propose an algorithm for symbolic regression for dynamic time lagged models that works in the following stages. First, once a maximum predictor time lag n is chosen, a dataset is generated that averages the values of time lagged predictor samples. For an observation of the dependent variable at time t samples of the predictors are averaged in a window between t and $t-n$:

$$\bar{X}_t = \frac{1}{n}\sum_{i=1}^{n} X_{t-i}$$

Secondly, the lag-window averaged dataset above is multiplexed for m functional forms of interest in the symbolic domain. Then predictors representing the various function forms computed using the lag-window averaged dataset are added to the set of predictors of the dependent variable. This allows us to attempt to select the functional form of the model without Almon lags present. LASSO/LARS multiple regression is then used to shrink the number of functional parameters used.

In the third stage, a small number of functional parameters generated from the second, shrinkage stage are used in an Almon polynomial lag model for each functional form. If the functional forms are $f_j(x)$, then the Almon lag model is

$$\hat{Y}_t = \beta_0 + \sum_j (\alpha_{0j} Z_{0jt} + \alpha_{1j} Z_{1jt} + \alpha_{2j} Z_{2jt}) + \epsilon,$$

where for each $f_j(x)$

$$Z_{0jt} = \sum_{i=1}^{n} f_j(X_{t-i})$$

$$Z_{1jt} = \sum_{i=1}^{n} i f_j(X_{t-i})$$

$$Z_{2jt} = \sum_{i=1}^{n} i^2 f_j(X_{t-i}).$$

Then the model is more succinctly written

$$\hat{Y}_t = \beta_0 + \sum_{i=0}^{2}\sum_{j=1}^{m} \alpha_{ij} Z_{ijt} + \epsilon. \tag{4}$$

VII. Experimental Results

In our tests, we use the three-stage symbolic regression algorithm on data from the PhysioNet project. PhysioNet provides free access to large databases of physiological signal datasets via the web. Open-source software and libraries are also provided for mining and analysis. The associated PhysioBank database is a archive of physiological signals provided freely to the telehealth research community and its many multi-parameter datasets are useful to for correlation and regression studies. It contains cardiopulmonary and neurological data and even gait databases from both healthy subjects and subjects under treatment, and many datasets include professional annotations.

For our study we used a dataset from the MIT-BIH Polysomnographic Database [9], which contains a collection of recordings of multiple physiologic signals during sleep. The subjects were monitored for evaluation of chronic obstructive sleep apnea syndrome at Boston's Beth Israel Hospital Sleep Laboratory. Subjects were also monitored to test the effects of a standard therapeutic intervention to prevent or substantially reduce airway obstruction called *constant positive airway pressure* (CPAP). The database consists of four-, six-, and seven-channel polysomnographic recordings, and contains over 80 hours' worth of data.

The recording that we chose, SLP59, includes an ECG signal, an invasive blood pressure signal (measured using a catheter in the radial artery), an EEG signal, and two respiration signals—one signal from a nasal thermistor and the second being a respiratory effort signal derived by inductance plethysmography. The dataset also includes a cardiac stroke volume signal and an earlobe oximeter signal. All signals are sampled at a rate of 250 Hz. The dataset also contains annotation files; The ECG signal has beat-by-beat annotations, and the EEG and respiration signals annotated with respect to sleep stages and apnea.

In our experiments we've used the abdominal plethysmography respiration signal as the independent variable, and the oxygen saturation signal as the dependent signal. Example waveforms of RESP (nasal) and SO2 from the dataset are given in Figure 1. 3600 samples of a the dataset were used to construct a time series to be fit to a bivariate distributed lag linear model. The data was downsampled to a rate of 1 Hz in order to provide for longer lags. The use of a finite distributed lag model requires the selection of a lag cutoff point beyond which there are no lagged variables. For simplicity, in this case, we chose a lag cutoff of 30 samples, or, given the downsampling, 30 seconds.

At first we attempted a multiple time-lagged regression with out a functional form or polynomial lags. The intercept estimate had 95% confidence with a t value of 177.014. The coefficient estimates and t values are in Figure 2. About half of the time-lagged variables have t values at the 95% confidence level, with the t value curve peaking at a time lag of 9 seconds. However, this model achieves an R^2 value of 0.016, indicating that very little of the variability in the dependent variable was

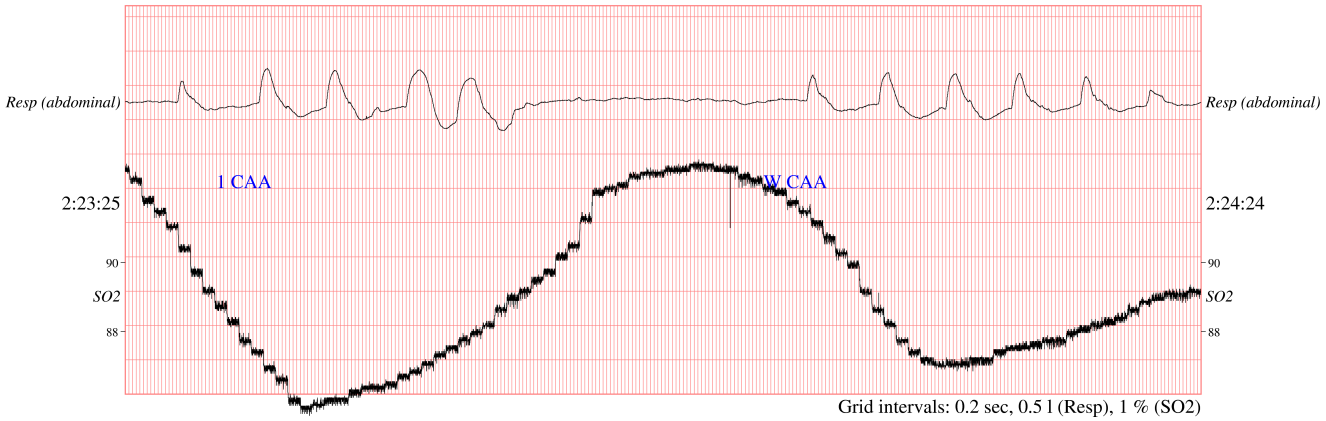
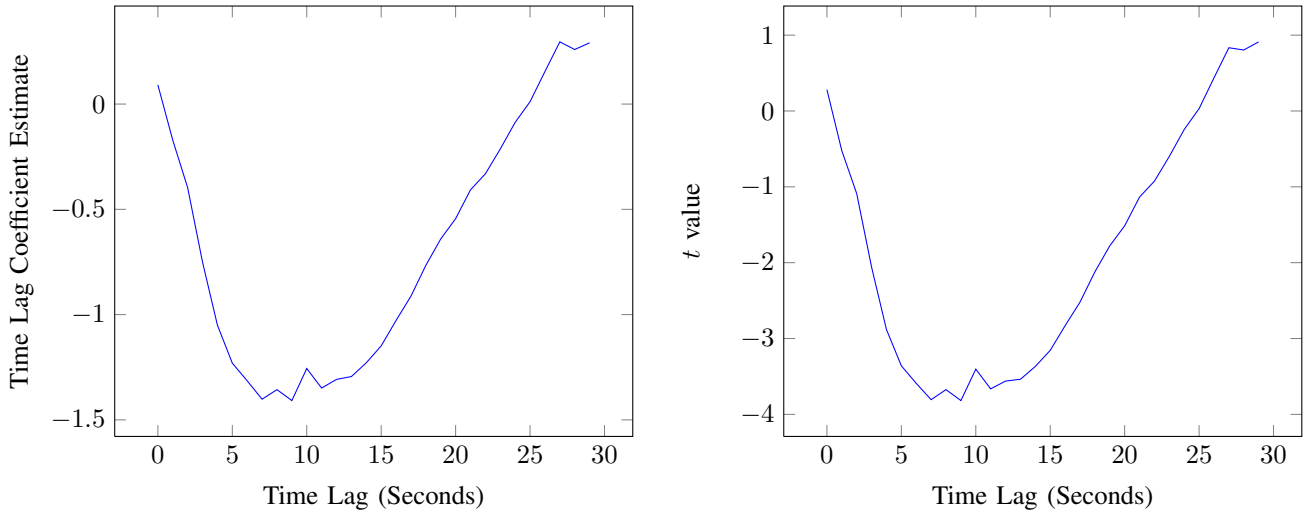


Fig. 1: Example abdominal respiration signal (in liters) and oxygen saturation signal (in percentage) from the MIT-BIH Polysomnographic Database dataset used. Also seen are the sleep stage annotations given at 30 second intervals. A sleep apnea episode occurs in the center of the chart, reducing the airflow through respiration. A corresponding decline can be observed in the oxygen saturation signal, which later increase when the sleep apnea episode subsides.



(a) Coefficient Estimates

(b) t values

Fig. 2: Coefficient estimates and t values for a time-lagged multiple regression with untransformed predictors. Data from the MIT-BIH Polysomnographic Database was used with a abdominal respiration as the time-lagged predictor signal and blood oxygen saturation as the dependent signal. Greater absolute values of t indicate greater statistical significance of the predictor for that time lag.

captured in the model.

Next we average data from 30 seconds and used it for the symbolic regression with the hope of later using Almon lag coefficients. The LASSO was used as a shrinkage method to determine a short functional form to use. We chose the following functional forms to use in shrinkage: $f_1(x) = x$, $f_2(x) = |x|$, $f_3(x) = x^2$, $f_4(x) = x^3$, $f_5(x) = |x^3|$, $f_6(x) = \log(|1 + x|)$, $f_7(x) = \sqrt{|x|}$, $f_8(x) = e^x$.

The results from running the LASSO are shown in Figures 3 and 5. The former is a plot (produced by the LARS package for R) of the progression of the LARS/LASSO algorithm as it adds coefficients and increases or decreases their values with each step. The progression is mapped as a function of the normalized L1 norm of the coefficient vector $|\beta|$. Figure 5 shows the step-by-step progression of the algorithm as it adds and removes parameters to and from the model. In LARS earlier steps represent smaller values of the shrinkage parameter t, while later steps represent larger values. The LARS regression had an R^2 value of 0.40.

The LARS/LASSO execution gave preference to the functional forms $|x|$ and $\log(|1 + x|)$ for small values of the shrinkage parameter. We chose to use these functional forms

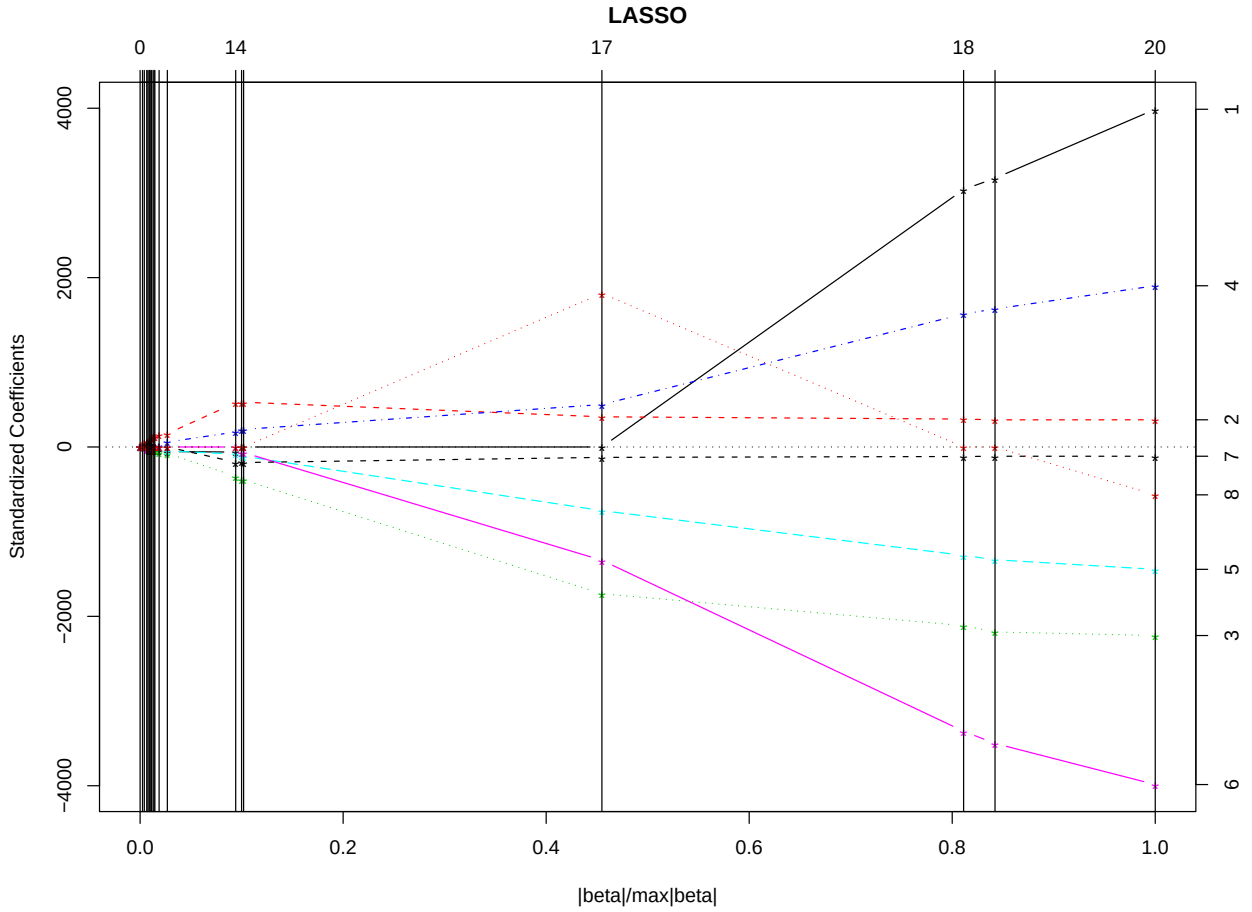


Fig. 3: A plot of the progression of the LARS algorithm on the symbolic regression model. Steps in the algorithm are represented by the vertical lines proceeding from left to right. Coefficients are plotted in relation to values of the normalized shrinkage parameter t as $|\beta|/\max|\beta|$. The predictors are represented in the graph by integer function index $f_1 = x$, $f_2 = |x|$, $f_3 = x^2$, $f_4 = x^3$, $f_5 = |x^3|$, $f_6 = \log(|1+x|)$, $f_7 = \sqrt{|x|}$, $f_8 = e^x$.

Function	Almon Predictor	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
	(Intercept)	90.56	0.4671	193.871	$< 2 \times 10^{-16}$
$\|x\|$	Z_{00t}	0.8942	0.1082	8.265	$< 2 \times 10^{-16}$
$\|x\|$	Z_{10t}	0.01543	1.708×10^{-3}	9.034	$< 2 \times 10^{-16}$
$\|x\|$	Z_{20t}	-9.245×10^{-5}	7.79×10^{-6}	-11.864	$< 2 \times 10^{-16}$
$\log(\|1+x\|)$	Z_{01t}	-0.7941	0.1506	-5.275	1.41×10^{-7}
$\log(\|1+x\|)$	Z_{11t}	-0.01734	2.010×10^{-3}	-8.627	$< 2 \times 10^{-16}$
$\log(\|1+x\|)$	Z_{21t}	9.975×10^{-5}	8.581×10^{-6}	11.625	$< 2 \times 10^{-16}$

Fig. 4: Regression results for the final stage of the shrinkage symbolic algorithm. The functions $f_2 = |x|$ and $f_6 = \log(|1+x|)$ were used with a 2nd-order Almon lag polynomial for a model with 6 predictors.

Step	Variable Index	$f(x)$	Action
LARS Step 1	Variable 2	$\|x\|$	added
LARS Step 2	Variable 6	$\log(\|1+x\|)$	added
LARS Step 3	Variable 7	$\sqrt{\|x\|}$	added
LARS Step 4	Variable 1	x	added
Lasso Step 5	Variable 6	$\log(\|1+x\|)$	dropped
LARS Step 6	Variable 5	$\|x^3\|$	added
LARS Step 7	Variable 8	e^x	added
Lasso Step 8	Variable 5	$\|x^3\|$	dropped
LARS Step 9	Variable 3	x^2	added
Lasso Step 10	Variable 7	$\sqrt{\|x\|}$	dropped
Lasso Step 11	Variable 8	e^x	dropped
LARS Step 12	Variable 4	x^3	added
LARS Step 13	Variable 5	$\|x^3\|$	added
LARS Step 14	Variable 7	$\sqrt{\|x\|}$	added
LARS Step 15	Variable 6	$\log(\|1+x\|)$	added
Lasso Step 16	Variable 1	x	dropped
LARS Step 17	Variable 8	e^x	added
LARS Step 18	Variable 1	x	added
Lasso Step 19	Variable 8	e^x	dropped
LARS Step 20	Variable 8	e^x	added

Fig. 5: The results of LARS algorithm used for shrinkage symbolic regression algorithm. The steps given correspond to the vertical lines in Figure 3, and increasing steps are the progression of the LARS algorithm and results of the LASSO regression for increasing values of the shrinkage parameter τ.

in the final stage of the algorithm. A regular regression was attempted with the window-averaged predictors using only the two functional forms $|x|$ and $\log(|1+x|)$ as separate predictors. This model achieved t test values of 29.12 and -14.77 for $|x|$ and $\log(|1+x|)$ respectively indicating a high level of significance. The model achieved a multiple regression R^2 value of 0.2685.

To perform the final stage of the algorithm, a regression was performed using Almon lagged predictors and using the functional forms $|x|$ and $\log(|1+x|)$. The results of this regression are listed in Figure 4. For this regression all variables enjoyed significance at the 99.9% level, and the model as a whole had an R^2 of 0.3373.

VIII. Conclusion and Future Work

In this paper we have demonstrated an efficient method for symbolic regression and model selection among possible transformation functions for the predictor variables. The three-stage method consists of averaging the time-lagged predictors over numerous functional forms, using the LASSO shrinkage regression method to select a small number of these forms, and performing a polynomial lag regression with these forms. It has been applied to medical and physiological time series data, specifically the link between respiration and blood oxygen saturation percentage in sleep apnea patients.

We found that our method for selecting a functional transformation of the predictor variable achieved a far higher goodness of fit according to standard analysis of variance measures. In the dataset examined, the model achieved a multiple R^2 of 0.3373, while a plain time-lagged model without transformation or polynomial lags had a R^2 of only 0.016. All of the variables in the model produced by the algorithm had high scores in t tests for validity.

More intelligent selection of functional forms for shrinkage may be possible in future work. A form of the LARS algorithm which takes possible functional forms explicitly into account is under investigation. A study of signal differencing will likely result in better quality of the regressions over such signals, since greater respiration should result in an increase in blood oxygen levels, not simply a higher absolute blood oxygen level. The methods may enjoy further success in different medical signal domains.

References

[1] S. Almon. The distributed lag between capital appropriations and net expenditures. *Econometrica*, 33(1):178–=196, 1965.

[2] L. K. Au, W. H. Wu, M. A. Batalin, T. Stathopoulos, and W. J. Kaiser. Demonstration of active guidance with smartcane. In *IPSN '08: Proceedings of the 7th international conference on Information processing in sensor networks*, pages 537–538, Washington, DC, USA, 2008. IEEE Computer Society.

[3] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. pages 1–17. Springer, 2004.

[4] B. Chiu, E. Keogh, and S. Lonardi. Probabilistic discovery of time series motifs. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–498, New York, NY, USA, 2003. ACM.

[5] R. B. Darlington. Multiple regression in psychological research and practice. *Psychological Bulletin*, 69(3):161–182, March 1968.

[6] N. R. Draper and H. Smith. *Applied Regression Analysis*. Wiley-Interscience, 1998.

[7] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2002.

[8] T. Gao, T. Massey, L. Selavo, D. Crawford, B. rong Chen, K. Lorincz, V. Shnayder, L. Hauenstein, F. Dabiri, J. Jeng, A. Chanmugam, D. White, M. Sarrafzadeh, and M. Welsh. The advanced health and disaster aid network: A light-weight wireless medical system for triage. *Biomedical Circuits and Systems, IEEE Transactions on*, 1(3):203–216, Sept. 2007.

[9] Y. Ichimaru and G. Moody. Development of the polysomnographic database on cd-rom. *Psychiatry and Clinical Neurosciences*, 53:175–177, April 1999.

[10] J. R. Koza. *Genetic programming: on the programming of computers by means of natural selection*. The MIT Press, Cambridge, MA, 1992.

[11] M. Lan, A. Nahapetian, A. Vahdatpour, L. Au, W. Kaiser, and M. Sarrafzadeh. Smartfall: An automatic fall detection system based on subsequence matching for the smartcane. In *BodyNets 2009: Proceedings of the ICST 3rd international conference on Body area networks*, ICST, Brussels, Belgium, Belgium, 2009. ICST.

[12] J. Lin, E. J. Keogh, S. Lonardi, and P. Patel. Finding motifs in time series. In *2nd Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 2002.

[13] L. Ljung. *System Identification: Theory for the User (2nd Edition)*. Prentice Hall PTR, 1998.

[14] G. Maddala. *Introduction to Econometrics, Third Edition*. John Wiley and Sons, LTD, 2001.

[15] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson. Wireless sensor networks for habitat monitoring. In *WSNA '02: Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, pages 88–97, New York, NY, USA, 2002. ACM.

[16] T. Oates, M. D. Schmill, and P. R. Cohen. A method for clustering the experiences of a mobile robot that accords with human judgments. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 846–851. AAAI Press / The MIT Press, 2000.

[17] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[18] R. Szewczyk, A. Mainwaring, J. Polastre, J. Anderson, and D. Culler. An analysis of a large scale habitat monitoring application. In *SenSys '04: Proceedings of the 2nd international conference on Embedded networked sensor systems*, pages 214–226, New York, NY, USA, 2004. ACM.

[19] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[20] A. Vahdatpour, N. Amini, and M. Sarrafzadeh. Toward unsupervised activity discovery using multi dimensional motif detection in time series. In *International Joint Conference on Artificial Intelligence*. ACM, 2009.

[21] S. Weisberg. *Applied Linear Regression*. Wiley-Interscience, 2005.