9-2010

Grouped Variable Model Selection for Heterogeneous Medical Signals

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

Majid Sarrafzadeh
University of California, Los Angeles

Grouped Variable Model Selection for Heterogeneous Medical Signals

Jamie Macbeth
Department of Computer Science
University of California
Los Angeles, California, USA
macbeth@cs.ucla.edu

Majid Sarrafzadeh
Department of Computer Science
University of California
Los Angeles, California, USA
majid@cs.ucla.edu

ABSTRACT

We explore statistical regression techniques for use in medical monitoring and telehealth applications. Medical embedded systems of the present and future are recording vast sets of data related to medical conditions and physiology. In this paper, distributed time-lag linear models are proposed as a means to help explain relationships between two or more medical and physiological measurements. The issues associated with performing multiple regression with heterogeneous medical data are treated as problems in model selection. An automatic method of model selection is proposed to construct models for high sample rate data by grouping sets of predictor variables.

The grouped predictor variable model optimization problem is formalized. Once an initial regression is performed on all available variables, our approximate algorithm for finding the grouped variable model with the greatest validity runs in $O(n^2)$ time, where n is the number of available predictor variables. This is compared to the *all subsets* technique which requires $O(2^n)$ time for the same predictor set. In our experiments with medical signal data, we find that the method produces models with reasonable goodness of fit scores and high average confidence levels for grouped predictors.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Health; J.3 [**Life and Medical Sciences**]: Medical Information Systems; I.5.4 [**Pattern Recognition**]: ApplicationsSignal Processing

General Terms

Algorithms, Theory

Keywords

Medical Signals, Regression, Model Selection

BodyNets 2010 Corfu Island, Greece

1. INTRODUCTION

Medical monitoring has become a very fruitful and vital area of study for embedded sensing systems. Embedded systems in this domain are able to collect huge amounts of data at a very high resolution from multiple locations at very low costs [9] [12]. These systems are flexible in the way that scientists can reprogram or re-task them after deployment in the field. They also offer early detection of physical ailments and can enhance doctor and patient relationships by offering remote diagnoses. Systems also exist for emergency medical response to catastrophic events like earthquakes, typhoons or disease epidemics. These systems can help to enhance the expertise of trained medical professionals and provide tools for investigative efforts to cure chronic illnesses.

The volume of data that is to be collected by medical monitoring systems of the present and future is overwhelming. Data is collected using a variety of medical sensors and diagnostic devices in different environments. Collected data sets have different sampling rates, signal characteristics and signal-to-noise ratios, and can be single or multi-channeled. There is 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. *Regression* is a standard procedure for statistically fitting data observations to a model to discover correlations between the quantities. The system and algorithms under discussion perform efficient linear model regressions for correlation studies and for prediction to aid in clinical research and medical care environments.

2. 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 [2] perform activity recognition from acceleration data using several classification methods.

Motif finding attempts to find previously known or unknown patterns in time series databases [8] and probabilistic discovery of motifs is also possible [3]. Motifs are useful for activity detection in embedded sensing medical systems [13]. Oates, et al. study clustering of signals for robotics [10]. 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. [7] [1].

3. BACKGROUND

The relationships between an independent variable and one or more dependent variables is our primary interest in the current work. In this domain, experiments attempt to assess the effects of variations in the independent variable on the dependent variable as a response measure. Measurements are obtained on two or more variables through data collected either on a single subject or on multiple subjects and we are interested in knowing whether or not the variables move together or co-vary. Studies of this kind are *correlational* in that they attempt to determine whether or not two variables have a dependent tendency. We are also interested in measuring the strength and direction of these relationships.

In typical physiological studies involving exercise or movement, some variables may be sampled at a far higher rate than others. We propose efficient algorithms for developing descriptive grouped-variable models for use in physiological studies. 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.

4. FORMALIZATION

4.1 Multiple Regression

Let Y represent a dependent or criterion variable, and X_1, X_2, X_3, $\ldots X_n$ represent independent or predictor variables of Y. All variables are continuous. A single observation of the dependent and independent variables in this case consists of an observation of Y coupled with observations of the independent variables X_i. Usually we will consider cases where values for variables are chosen from a continuous, totally-ordered set.

Score values of these variables are observed from a *population*. We assume that any data set we use is a *sample* from the population as larger group. Multiple regression methods will attempt to derive or calculate β_1, β_2, β_3, $\ldots \beta_n$, for the predictor variables, and the constant β_0 such that

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n + \epsilon \qquad (1)$$

is a good predictor for 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.

4.2 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. This is because of the high costs involved in obtaining and storing information on a large number of predictor variables for all observations. Furthermore, 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 [4]. On the other hand, one needs 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.

Given p predictors to select from, any member of the power set of predictors 2^p could be used in the model. To perform a regression with every possible set of predictors is usually intractable. Thus when many predictors are available, schemes, heuristics, and algorithms are necessary to help find a good set of predictors.

Many procedures have been proposed for model selection. The *all subsets* algorithm actually does perform 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 of these methods have stopping criteria under which the process completes with a certain number of the available parameters. Draper and Smith [5] and Weisberg [14] provide useful surveys of the details of the inner workings of these methods.

4.3 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 $0 < t < T$. Y_t and X_t will represent the values of Y and X at time t if they exist. Causality is assumed, and if Y_t exists, X_t, X_{t-1}, X_{t-2}, X_{t-3}, $\ldots X_0$ 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 one variable is used to predict another, the model is more specifically known as 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.

4.4 Choosing Observations from the Predictor Time Series

We will restrict our set of predictor variables for Y_t to n values of the time series in X represented by X_t, X_{t-1}, X_{t-2}, X_{t-3}, $\ldots X_{t-n}$.

One typical objective in the use of linear models is to reduce the size of the set of predictor variables, which pursues the definition of a model with few explanatory factors. In many research and clinical applications, simple explanations and rules of thumb are desired to help understand parts of complex phenomena. Also, the number of explanatory variables used in the model needs to be kept small when compared to the number of observations made of the model, otherwise the number of degrees of freedom in fitting the model to the data is too small to guarantee confidence in the results.

4.5 Sample Groupings

We shall present a family of variable-grouped distributed lag models in the following way: we shall define new predictor variables $X_{i,j}$ subscripted by a closed interval representing the time interval of the group,

$$X_{i,j} = \frac{\sum_{k=i}^{j} X_k}{j-i} \tag{2}$$

which is the mean of the X_k during the time interval $[i, j]$. $\beta_{i,j}$ will represent the correlation coefficient for $X_{i,j}$.

4.6 Dyadic-Interval Group Time Series Models

We propose dyadic-interval group time series models where not all variable groups are the same size. We'll base these models on time lag interval windows τ_1, τ_2, τ_3, ..., τ_k with $\tau_i = [a_i, b_i]$, corresponding variables, $X_{[t-a_i,t-b_i]}$, and coefficients $\beta_{[t-a_i,t-b_i]}$. As a shorthand, we will describe variables solely with the lag interval endpoints and elide t in the interval specification of X and in the specification of Y:

$$Y_t = Y; \qquad X_{[t-a,t-b]} = X_{a,b}. \tag{3}$$

This way, the interval group time series model can be written:

$$Y = \beta_0 + \beta_{a_1,b_1} X_{a_1,b_1} + \beta_{a_2,b_2} X_{a_2,b_2} + \beta_{a_3,b_3} X_{a_3,b_3} \\ \ldots + \beta_{a_n,b_n} X_{a_n,b_n}$$

We require that the time lag interval windows are nonintersecting (except for endpoints). We will also require that for any interval $[a_i, b_i]$ there are non negative integers k and ν such that $a_i = k2^\nu$ and $b_i = (k+1)2^\nu - 1$. With this constraint, the set of possible time lag interval windows is dyadic and forms a binary tree with these properties:

1. The length of an interval is always an integer power of two.
2. Every interval is contained in exactly one "parent" interval of twice the length.
3. Every interval is spanned by two "child" intervals of half the length.
4. If two intervals overlap, then one of them must be a subset of the other.

5. GROUPED VARIABLE MODEL SELECTION

5.1 Analysis of Variance

The standard deviation for any observed variable A is written

$$s_A = E[(A - E[A])^2] = E[A^2] - (E[A])^2, \tag{4}$$

while the covariance between two variables A and B is

$$s_{AB} = E[(A - E[A])(B - E[B])] = E[AB] - E[A]E[B]. \tag{5}$$

The Pearson correlation coefficient, or, more simply, the correlation, between any pair of variables A and B is written as

$$r_{AB} = \frac{s_{AB}}{\sqrt{s_A s_B}}. \tag{6}$$

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. Written in terms of expectation values of the model prediction, $\hat{Y}$, and the dependent variable observations, Y, it is

$$R^2 = \frac{E[(\hat{Y} - E[Y])^2]}{E[(Y - E[Y])^2]}. \tag{7}$$

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.

5.2 Problem Definition

The heterogeneously-grained grouped variable time series regression optimization problem can also be stated as follows: given the time series' Y and X as discussed above, we can find a set of dyadic lag intervals spanning the time region $0 < t < k$. A set of grouped variables can be generated from this set of intervals. While the algorithm does not directly attempt to generate a model to maximize the value of R^2, it assumes that the correlations between variables will be relatively small given an ideal grouping.

Our algorithm works to find a set of dyadic interval variables that satisfy certain conditions. Firstly, the set does not contain a direct or indirect parent node of any other node in the set. Secondly, the average of r_{XY}s for nodes in the set is maximized over all other sets satisfying the above condition.

More formally, we would like to find a set of dyadic intervals S such that

$$\arg\max_S \frac{1}{|S|} \sum_{X \in S} r_{XY} \tag{8}$$

with the condition that none of the intervals overlap.

5.3 Algorithmic Solution

We require the use of a lemma in the following discussion of algorithms to solve the problem. First we would like to show that if two dyadic lag window variables are merged it is easy to recalculate the value of Pearson's r_{XY} for the resulting grouped variable. In this case, r_{XY} gives us an easy-to-calculate local metric for the algorithm.

Assume that $a = k2^\nu$, $b = (k+1)2^\nu - 1$, $c = (k+1)2^\nu$, and $d = (k+2)2^\nu - 1$ for some k and ν. If we are given two adjacent dyadic lag window variables $X_{a,b}$ and $X_{c,d}$, and $X_{a,d}$ is a parent dyadic lag window variable of $X_{a,b}$ and $X_{c,d}$, we try to calculate

$$r_{X_{a,d}Y} = \frac{s_{X_{a,d}Y}}{s_{X_{a,d}} s_Y} \tag{9}$$

given $s_{X_{a,b}Y}$ and $s_{X_{c,d}Y}$. In the numerator:

$$s_{X_{a,d}Y} = 2^{-\nu} \left(E\left[\sum_{i=a}^{d} X_i Y\right] - E\left[\sum_{i=a}^{d} X_i\right] E[Y] \right) \tag{10}$$

$$= 2^{-\nu} \sum_{i=a}^{d} s_{X_i Y} = \frac{s_{X_{a,b}Y} + s_{X_{c,d}Y}}{2} \tag{11}$$

so that effectively the covariance of a group variable that is the average of its children is the average of the covariances of the children.

In the denominator:

$$s_{X_{a,d}} = E\left[\left(2^{-\nu}\sum X_i\right)^2\right] - \left(E\left[2^{-\nu}\sum X_i\right]\right)^2 \quad (12)$$

$$= 2^{-2\nu}\sum_{i=a}^{d}\sum_{j=a}^{d}(E[X_iX_j] - E[X_i]E[X_j]) \quad (13)$$

$$= 2^{-2}(s_{X_{a,b}} + s_{X_{c,d}}) + 2^{-2\nu}\sum_{i=a}^{b}\sum_{j=c}^{d} s_{X_iX_j} \quad (14)$$

while s_Y is independent of the variable grouping. Thus we are provided with an efficient method for computing the correlation for a grouped variable when the correlation for its constituent variables has been calculated. When the covariances used in the expression for calculating $s_{X_{a,d}}$ are known, the correlation of the parent group variable can be found in $(2^{\nu-1})^2 = 2^{2\nu-2}$ steps, up to a constant.

The algorithm that we propose is similar in spirit to many stepwise regression methods. However, typically, stepwise regression methods require that one or more regressions are performed at each stage when a variable is added or removed. Our method sacrifices the accuracy of calculating and comparing the partial correlations of variables to be added or removed from the model. It makes gains in efficiency by avoiding the extra regression calculations.

To achieve this, the algorithm will use the dyadic interval tree as its main data structure. At each node x in the tree we shall keep a list of the set of nodes in the subtree rooted at x for which the average of the r_{XY}s is the greatest, and for which the time lag variable grouping conditions given above are satisfied. We shall also store at each node the maximum average of the r_{XY}s, called ρ, to make for easy comparisons.

The algorithm proceeds as follows:

1. We begin with the full set of ungrouped predictor variables for which $\nu = 0$ and $0 \leq k \leq N-1$. These variables are the leaves of a dyadic lag interval tree.

2. Beginning with the parent nodes of the leaves, the correlation r_{XY} of each variable is calculated.

3. For each parent, its correlation is calculated.

4. Pairs of variables are grouped and removed from the model and replaced with their parent lag time window variable if the average of the correlations of the children's subtrees is less than that of the parent.

5. The process is repeated until the root of the tree is reached.

In order to conveniently track the sets of variables to be included in the model as the tree is traversed, we add the following step

1. For each node x that has children y and z, ρ_x is set to be the greater of $(\rho_y + \rho_z)/2$ and r_x. if r_x is greater, the set s_x is set to the singleton x. If the sum $\rho_y + \rho_z$ is greater, then s_x is set to the union $s_y \cup s_z$.

Code for the algorithm is given in Figure 1. Once the correlation r_{XY} is calculated for individual lag times in the time series, it can be calculated for groups of these parameters without recalculating sums of squares for the groups.

Once r_{XY} is calculated for various group sizes, a tree of Pearson's r_{XY}s is generated. At the root of the tree is r_{XY} calculated using the median of the predictor time series over all time samples. The leaves of the tree are r_{XY} values for each sample.

At the root level of the tree $(n/2)^2 = n^2/4$ operations are required to calculate $r_{X_{1,n}Y}$. At the next level $2(n/4)^2 = n^2/8$ are required, and so on. This comes out to

$$\frac{n^2}{4} + \frac{n^2}{8} + \frac{n^2}{16} + \ldots + \frac{n^2}{2n} \in O(n^2) \quad (15)$$

This is compared to the complexity of a standard all subsets model selection technique. We calculate the complexity if all dyadic predictor groups are regressed over. Let S_i be the number of subsets for a dyadic tree of height i. Consider the two subtrees of height $i-1$. Each has S_{i-1} dyadic interval subsets, so, in combination, a covering subset can be generated from any pairing of a subset in the left subtree and a subset in the right subtree. The singleton subset consisting of the root node is also a covering subset. Therefore we can write the total number of possible covering subsets using the following recurrence:

$$S_i = (S_{i-1})^2 + 1 \quad (16)$$

so that

$$S_{\lg n} \in O(2^{2^{\lg n}}) = O(2^n). \quad (17)$$

The all subsets method would require that a full regression is performed for each subset. To perform these regressions, covariances between the new group variable and all other variables must be calculated. Additionally, even if we ignore the complexity of the multiple regression steps in stepwise regression methods, our scheme already provides major efficiency savings.

6. EXPERIMENTAL RESULTS

Our experiments use data from the PhysioNet project [6]. PhysioNet provides free access to medical and physiological signal datasets and open-source software for analyzing them. The associated PhysioBank database is a archive of physiological signals provided freely to the biomedical research community. PhysioBank has many multi-parameter datasets useful for correlation and regression studies. The datasets are from both healthy subjects and from subjects with various medical conditions. It contains cardiopulmonary data, neurological data, and even gait databases, and many datasets include professional annotations.

For our study we used a dataset from the MIMIC database, which contains medical signal data from ICU patients. It provides three-lead EKG, respiration, arterial blood pressure, pulmonary arterial pressure, central venous pressure, and fingertip plethysmograph data sampled at 125 Hz. In our experiments we used the respiration signal, RESP, as the independent predictor variable, and the arterial blood pressure signal, ABP, as the dependent variable.

10000 samples of a MIMIC dataset were used to construct a time series to be fit to a bivariate distributed lag linear model. 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 512 samples, or 4.1 seconds. Initially, if no variable grouping is performed, this generates a linear model with

Require: $e.r \leftarrow r_{X_{a,b}Y}$ for group variable $X_{a,b}$ represented by e. For simplicity $n = 2^\nu$ for some positive integer ν.
Ensure: $e.S$ is an non-overlapping subset of the dyadic interval group variables in the subtree rooted at e which maximizes the average of r_{XY}. $e.\rho$ is the maximum average of r_{XY} achieved by S.

```
1: T ← a complete binary tree with n leaves; each node e ∈ T represents a dyadic group time lag variable.
2: T.depth(d) returns all nodes at a given depth
3: for all e ∈ T.depth(log n) do
4:    e.ρ ← e.r
5:    e.S ← {e}
6: end for
7: for j = log n − 1 to 0 do
8:    for all e ∈ T.depth(j) do
9:       if e.r > (e.leftchild.ρ + e.rightchild.ρ)/2 then
10:         e.ρ ← e.r
11:         e.S ← {e}
12:      else
13:         e.ρ ← (e.leftchild.ρ + e.rightchild.ρ)/2
14:         e.S ← e.leftchild.S ∪ e.rightchild.S
15:      end if
16:   end for
17: end for
```

Figure 1: The model selection algorithm

512 variables. With the lag, this generates 9488 separate cases to be used in the multiple regression.

We begin by generating the absolute validity r_{XY} for each variable in this 512 variable model. For each adjacent pair of lag variables, we then compare the average validities of the two with the average validities of a variable that is the average of those two variables. The grouped variable is the parent of these two variables in the dyadic interval tree. If the validity of the parent is greater than the average validity of the two children, the parent group variable is chosen for the linear model over the two child variables. This process is performed for all variables in the 512 variable model. It is then repeated for all variables at the 256 variable level. This process is repeated at all levels up to the univariate model with a single group variable for all 512 lag samples.

The procedure was performed for the time series generated from the MIMIC data set. Figure 2a shows a plot of group variable lag position against group size for all variables in the model generated by our algorithm. Figure 2b shows the absolute validities for each grouped variable in the resultant model as a function of start lag of the group.

After the model selection process, a regression is performed using the selected model and the data used to generate it. We performed the regression using the freely available statistical computing environment, R [11].

For the typical metric for the goodness of fit of a regression, R^2, the model generated by the method had a score of 0.41, which indicates that 41% of the variability in the observations of R was represented by the model. For the intercept, β_0, and for variables representing the first 39 samples of respiration data taken before the blood pressure reading, the confidence in the model is high. However, for variables representing time lags longer than this, the t test confidence levels decline rather quickly. High confidence returns after a time lag of around one second, and continues for the most part until the time lags in the model end at the 512th sample.

The performance of the model produced by our grouping algorithm is compared to various models of equal group sizes and their performance on the same data. Figure 3a shows the R^2 values for regressions using models with equal group sizes for various group sizes. It is seen that R^2 peaks at 0.55 for the model of group size 1 with 512 predictor variables. The value of R^2 of 0.41 for the model generated by our grouping algorithm is competitive with that reached by the best of the equal group size models.

Figure 3b shows the average of $|t|$ over all predictor variables for each equal group size model tested. Larger values of $|t|$ indicate better significance of the predictor variables. Our algorithm produced a model for which the average of $|t|$ was 6.04. We note that for equal group sizes, many of the models with higher average $|t|$ have lesser values of R^2. Overall, it appears that the grouping method has struck a balance between finding a model that explains the variance of the data and one whose predictor variables are significant.

7. CONCLUSION AND FUTURE WORK

In this paper we have demonstrated an efficient method for model selection among sets of grouped parameters. The method is applied to distributed lag time series models and data, specifically biomedical and physiological time series data. We found that our method for grouping parameters achieved a reasonable goodness of fit according to standard measures. In the data set examined, the model achieved a multiple R^2 of 0.41. Many of the grouped variable sets produced by the algorithm sufficiently passed t tests for statistical significance.

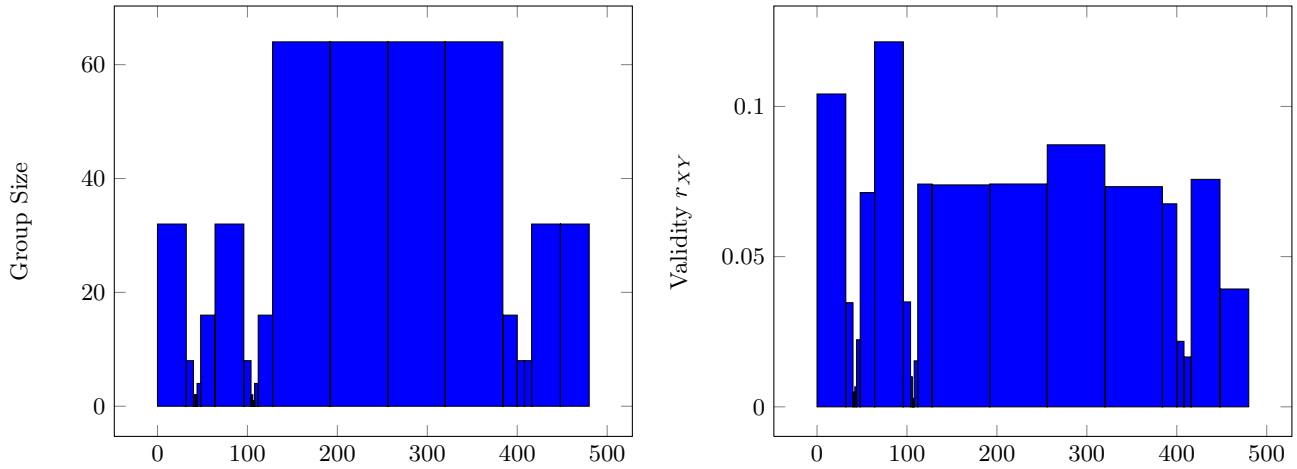
However, many improvements are possible in future work. We would like to consider problems related to multicollinearity both in the endogenous and the exogenous time series. Differencing will likely make an improvement in the quality of these regressions. As well, because of the regularity in certain types of physiological data such as blood pressure measurements and respiration, we would like to consider "seasonality" and periodicity issues in medical time series, and removing seasonalities before carrying out the regression. The grouping method may also be combined with standard methods for distributed-lag regression. Efforts to utilize polynomial-lag models in studies of medical and physiological signals are currently underway.

8. REFERENCES

[1] 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.

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

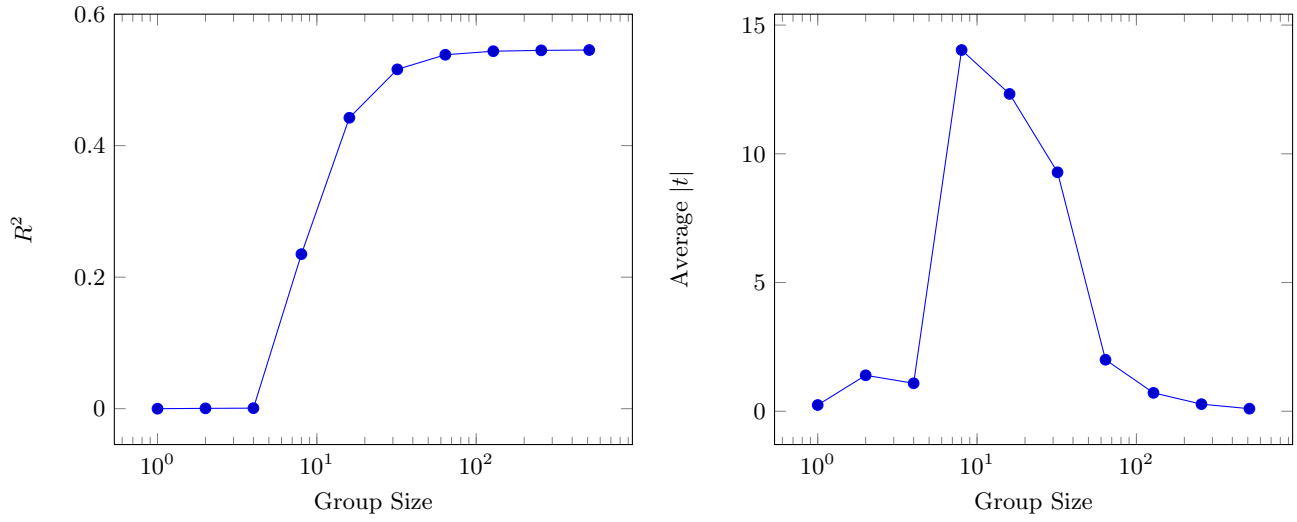
[3] B. Chiu, E. Keogh, and S. Lonardi. Probabilistic discovery of time series motifs. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–498, New York, NY, USA, 2003. ACM.



(a) Variable Groupings

(b) Grouped Variable Validities

Figure 2: Variable groups and grouped variable validities.



(a) R^2

(b) Average $|t|$ Values

Figure 3: R^2 values and Average $|t|$ Values for Various Group Sizes

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

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

[6] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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