HEALTH ECONOMETRICS: RESPIRATION-OXYGENATION CORRELATION THROUGH SPECTRAL MODELS

Jamie Macbeth
Department of Computer Science
University of California
Los Angeles, California 90095
email: macbeth@cs.ucla.edu

Majid Sarrafzadeh
Department of Computer Science
University of California
Los Angeles, California 90095
email: majid@cs.ucla.edu

ABSTRACT
Medical embedded systems are capable of recording vast data sets for physiological and medical research. Linear modeling techniques are proposed as a means to explore relationships between two or more medical or physiological signal measurements where a causal relationship is believed to be present. Multiple regression is explored for use in medical monitoring, telehealth, and clinical applications.

Spectral regression methods for high-bandwidth medical and physiological signals are demonstrated. The two-stage method consists of performing an FFT over a time-lagged window of the predictor signal, and constructing a model based on the FFT coefficients. The output of the regression is used in a clustering to explore structure in the array of spectral predictors. 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.

Spectral predictors achieved a dramatically better goodness of fit than time-lagged predictors according to standard analysis of variance measures. In the dataset examined, the spectral model achieved a multiple R^2 of 0.90, indicating that 90% of the variation in the dependent signal was captured by the model, while an ordinary distributed lag model had a R^2 of only 0.016.

KEY WORDS
Biomedical Modelling; Cardiovascular Modelling; Time Series Analysis; Respiratory Mechanics.

1 Introduction

The volume of data that is projected to be collected by medical embedded systems is overwhelming [11] [17]. 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. Furthermore, data is collected using a variety of diagnostic devices and health sensors in various types of environments. As a result there is a wide-ranging interest in systems for human-trained and automatic classification and interpretation of physiologic 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. To discover correlations between the quantities we use *regression*, a well-known method for statistically fitting data observations to a model. 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. In signal data that may represent the onset or degree of the medical condition or phenomenon in question, these systems perform pattern matching and learn signal patterns.

Embedded systems in the medical monitoring domain offer early detection of physical ailments and can enhance the doctor and patient relationship by offering remote diagnoses. Additionally, they can help to enhance the expertise of trained health care professionals, and to search for cures to chronic illnesses. These systems are flexible in the way that scientists can reprogram or re-task them after deployment in the field. Systems also exist for emergency medical response to catastrophic events like earthquakes, typhoons or disease epidemics.

2 Related Work

Related work has been performed on categorizing, organizing, and processing medical and physiological time series and signals. Motif finding attempts to find previously known or unknown patterns in time series databases [8], and motifs are useful for activity detection in embedded sensing medical systems [18]. Probabilistic discovery of motifs is also possible [2]. Activity detection studies attempt to classify physical activities that the subjects are performing purely through the physiological signals recorded. Bao and Intille [1] perform activity recognition from acceleration data using several classification methods. Oates, et al. study clustering of signals for robotics [12]. To our knowledge, our use of time-lagged regression to study physiologic signal data is the first of its kind.

The data used in the current work was collected as part of a study by Garpestad, et al. [5] on cardiac function during sleep apnea cycles. In related work, Chon, Dash, and Ju [3] attempt to estimate the respiration rate through time-frequency spectra of the pulse oximetry signal. Lu et al. [9] explore algorithms for detecting peaks and valleys in respiratory signals. Shelley, Awad, Stout and Silverman [16] use spectral analysis of pulse oximetry signals as an alternative method for measuring respiration rates when compared to CO_2 detection. Studies on sleep apnea highlight its possible links to negative effects on cardiovascular physiology, hypertension, and cardiovascular disease [15] [13].

3 Background

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. 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. 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.

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. 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.

Econometrics models and methods are indispensable when data on variables is highly interrelated and observed over time, individuals, or space [10]. 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.

3.1 Model Selection

The availability of many possible predictors to choose from to perform a regression precipitates problems in linear model selection. Reducing the size of the set of predictor variables pursues the definition of a model with fewer explanatory factors, and in many research and clinical applications, simple explanations and rules of thumb are preferred to help understand parts of complex phenomena. On the other hand, one must 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. Additionally, models are usually more efficacious when they have 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.

Many procedures have been proposed for model selection. *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. The *all subsets* algorithm performs regressions with all 2^p possible linear models, given p predictors to choose from. Both stepwise regression and backward elimination have stopping criteria under which the process completes with a certain subset of the available parameters.

4 Formalization

4.1 Multiple Regression

Let Y represent a dependent or criterion variable, and let X_1, X_2, X_3, $\ldots X_n$ represent independent or predictor variables of Y. An observation of Y coupled with observations of the independent variables X_i is a a *case* or a *run* of an *experiment*. Observations of values for any given variable will form a continuous, totally-ordered set.

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 $\ldots \beta_n$ for the predictor variables. In the equation

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

$\hat{Y}$ 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

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

4.2 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}, $\ldots X_0$ 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 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 $X_{t-1}, X_{t-2}, X_{t-3}, \dots X_{t-n}$. The model can be succinctly written

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

4.3 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.

5 Oscillatory Analysis

As we perform distributed time-lagged regression over signals, where the time scales of the alleged correlations between the two waveforms may be much longer than their sampling frequencies, we seek out methods to manage the number of predictors. The predictors need to cover the time-lag region in which the suspected correlation is in place.

Because our study uses the respiration effort signal as a predictor signal, and because the respiration effort signal has so many periodic characteristics, we chose to use spectral characteristics of the signal in the regression. More specifically, Rather than simply perform multiple regression with time-lagged predictors, we propose multiple regression with coefficients from a Fourier transform of the predictor signal as predictors. In our study a fast Fourier transform of a segment of the predictor signal residing in a time lagged window is used to predict the exogenous signal.

6 Clustering of Spectral Predictors

We observe that the use of spectral information requires the use of many predictors in the model for the bandwidths of signals in use. However, multiple regression often benefits when less predictors can be used. The goal of reducing the independent variable set may be achieved when representative predictors are used, and when predictors can be placed in groups with similar characteristics.

We attempt the placement of predictors into similar groups in this study through the use of clustering algorithms. Clustering algorithms group sets of observations, usually according to a parameter k representing the desired number of clusters to be found by the algorithm. Hierarchical clustering algorithms solve the clustering problem for all values of k using bottom up and top down methods.

We use a hierarchical clustering algorithm called AGNES [7] to cluster the spectral predictors based on three criteria obtained from a multiple regression performed on the FFT coefficients. As measures of similarity used in clustering, these criteria are the FFT index, the regression coefficient estimates themselves, and the regression coefficient t values.

The AGNES algorithm constructs a hierarchy of clusterings. At first, each observation is a small cluster by itself. Clusters are merged until only one large cluster remains containing all of the observations. At each stage the two nearest clusters are combined to form one larger cluster. The AGNES algorithm also yields the agglomerative coefficient (a value between 0 and 1) which measures the amount of clustering structure found.

7 Experimental Results

In our tests we perform regression predictor clustering 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 [6], 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

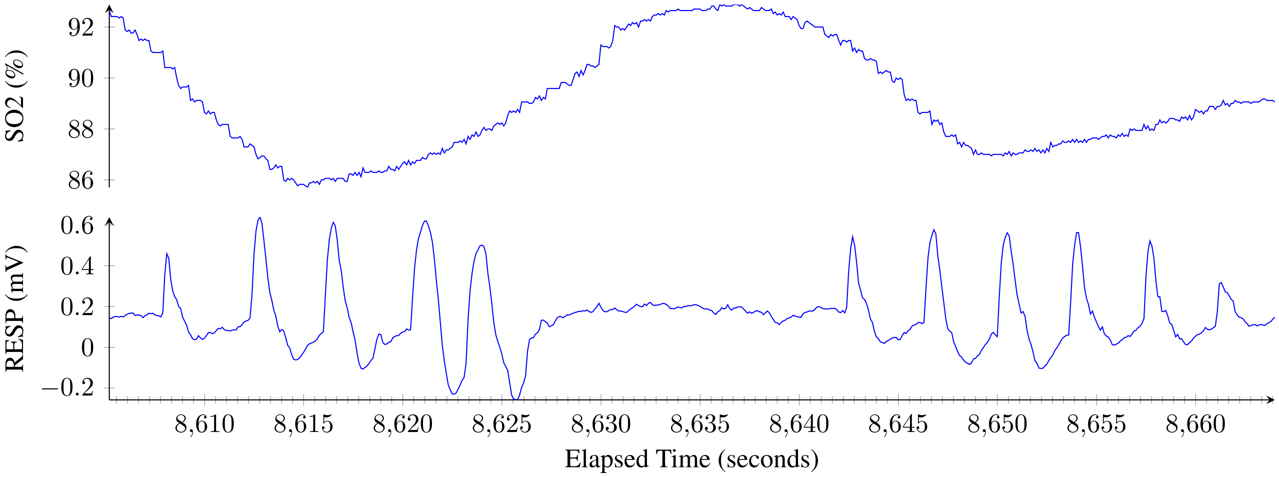

Figure 1: Example abdominal respiration signal (in millivolts) and oxygen saturation signal (in percentage) from the MIT-BIH Polysomnographic Database dataset used. A sleep apnea episode occurs in the center of the chart, reducing the airflow through respiration. A corresponding decline can be observed in the oxygen saturation signal, which later increases when the sleep apnea episode subsides.

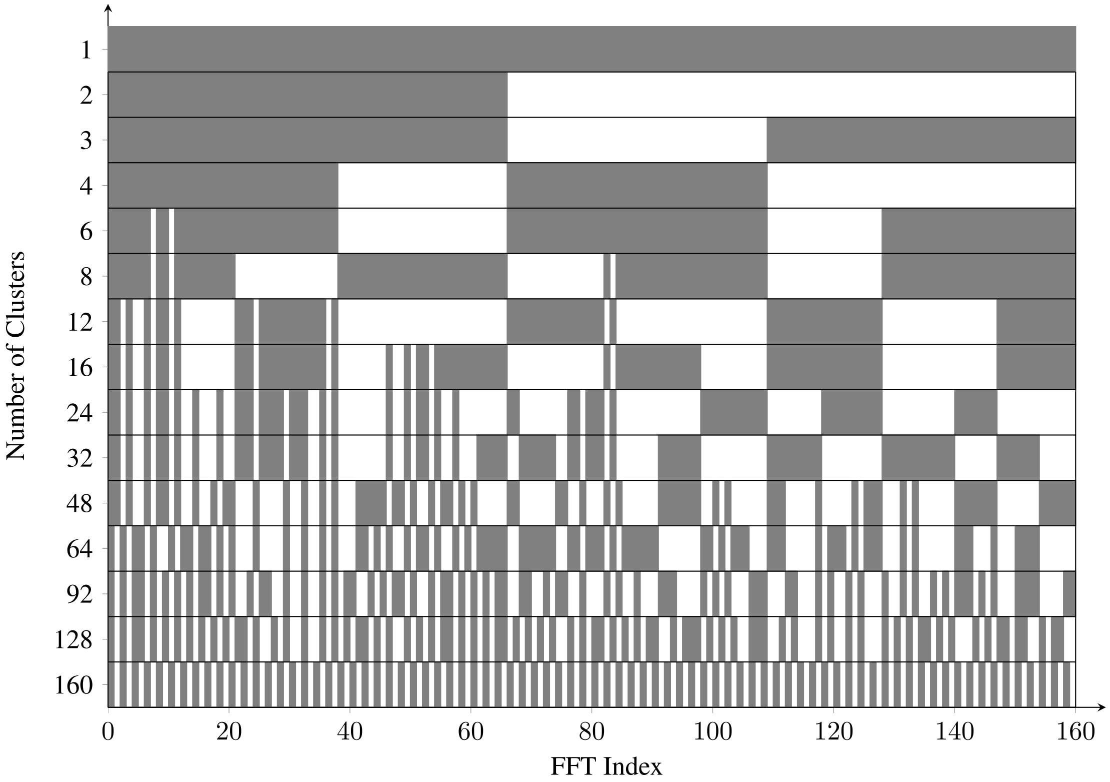

Figure 2: Clustered predictor groups for FFT coefficients 0-159, illustrating the execution of the AGNES clustering algorithm.

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 are annotated with respect to sleep stages and apnea.

7.1 Time-Lagged Multiple Regression

In our experiments we used the abdominal plethysmography respiration signal as the independent variable, and the oxygen saturation signal as the dependent variable. Example waveforms of RESP 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.

The R software environment for statistical computing [14] was used to perform the multiple regression. The intercept estimate had 95% confidence with a t value of 177.01. 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 captured in the model.

We had only moderate success using time-lagged multiple regression to predict blood oxygenation using the respiratory effort signal. As one can see from the figure, the plethysmographic waveform has a very periodic character as the patient inspires and expires air. Rather than simply perform multiple regression with time-lagged predictors, we propose multiple regression with coefficients from a Fourier transform of the predictor signal as predictors. In our study, a fast Fourier transform of a segment of the predictor signal residing in a time lagged window is used to predict the exogenous signal.

For the spectral regression algorithm, in total 90000 samples (360 seconds) of the dataset were used to construct a time series. Here the data was downsampled by a factor of 25 to a rate of 10 Hz. For each sample of the oximetry signal, a fast Fourier transform is performed on the segment of the predictor signal residing within a time-lagged window of 8000 samples (32 seconds). The first sample of the time-lagged window occurs at the same point in time as the dependent signal, and the last sample of the time-lagged window occurs at a point 8000 samples earlier.

Downsampling by a factor of 25X was performed. For accurate downsampling, rather than choose a single representative sample, the 10 samples for each signal were averaged. Smoothed samples were buffered and the fftw package [4] was used to perform FFTs. Under the assumption that little phase information would be useful in the prediction, the moduli of the of the FFT coefficients were utilized as predictors.

We are careful in the following text to distinguish between the FFT coefficients which are used as predictors in the regression, and the regression coefficients β which appear in front of the FFT coefficient values in the model. The multiple regression used only FFT coefficients indexed 0-159, representing the frequency band from 0 to 5Hz.

We observed that some of the lower-frequency FFT coefficients tend to have greater t values and thus greater validity. The regression resulted in a residual standard error of 0.7556 on 3118 degrees of freedom and a multiple R^2 of 0.90 indicating that 90% of the variability in the of the oximetry signal was captured by the respiration effort model.

The cluster package available for R was used to perform an AGNES clustering. A figure illustrating the fragmentation of the clustering groups is given in Figure 2. A change in color from one FFT index to another as a row is traversed from left to right indicates a boundary between two clusters. The groups are given for 160, 128, 92, 64, 32, 24, 16, 12, 8, 6, 4, 3, 2, and 1 clusters. The agglomerative coefficient, an indicator of the amount of clustering structure found, was 0.966.

8 Conclusion

In this paper we have demonstrated an efficient spectral regression method for high-bandwidth medical and physiologic signals. The two-stage method consists of performing an FFT over a time-lagged window of the predictor signal, and constructing a model based on the FFT coefficients. The output of the regression is used in a clustering to explore structure in the array of spectral predictors. 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 using spectral variables as predictors achieved a far higher goodness of fit than a plain distributed time-lag model according to standard analysis of variance measures. In the dataset examined, the spectral model achieved a multiple R^2 of 0.90, while a plain time-lagged model had a R^2 of only 0.016. Many of the variables in the model produced by the algorithm had high scores in t tests for validity.

In future work we will consider algorithms for selecting the length of the time-lagged window, as this will have a major effect on the spectral information available. The clustered spectral predictors may be used for subsequent regressions as well.

References

[1] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. *Pervasive Computing*, 2004, 1–17.

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

[3] K. Chon, S. Dash, and K. Ju. Estimation of respiratory rate from photoplethysmogram data using time-frequency spectral estimation. *IEEE Transactions on Bio-medical Engineering*, 56(8), 2009, 2054–2063.

[4] M. Frigo and S. G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2), 2005, 216–231.

[5] E. Garpestad, H. Katayama, J. A. Parker, J. Ringler, J. Lilly, T. Yasuda, R. H. Moore, H. W. Strauss, and J. W. Weiss. Stroke volume and cardiac output decrease at termination of obstructive apneas. *J Appl Physiol*, 73(5), 1992, 1743–1748.

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

[7] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data, An Introduction to Cluster Analysis*. (Hoboken, NJ, Wiley-Interscience, 2005).

[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*, Edmonton, Alberta, Canada, 2002.

[9] W. Lu, M. Nystrom, P. Parikh, D. Fooshee, J. Hubenschmidt, J. Bradley, and D. Low. A semi-automatic method for peak and valley detection in free-breathing respiratory waveforms. *Medical Physics*, 33(10), 2006, 3634–3636.

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

[11] 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*, New York, NY, USA, 2002, 88–97.

[12] 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*, Seattle, Washington, 2000, 846–851.

[13] P. Peppard, T. Young, M. Palta, and J. Skatrud. Prospective study of the association between sleep-disordered breathing and hypertension. *New England Journal of Medicine*, 342(19), 2000, 1378–1384.

[14] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[15] E. Shahar, C. Whitney, S. Redline, E. Lee, A. Newman, F. Javler, N. George, T. O'Connor, L. Boland, J. Schwartz, and J. Samet. Sleep-disordered breathing and cardiovascular disease: Cross-sectional results of the sleep heart health study. *American Journal of Respiratory and Critical Care Medicine*, 163(1), 2001, 19–25.

[16] K. Shelley, A. Awad, R. Stout, and D. Silverman. The use of joint time frequency analysis to quantify the effect of ventilation on the pulse oximeter waveform. *Journal of Clinical Monitoring and Computing*, 20(2), 2006, 81–87.

[17] 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*, New York, NY, 2004, 214–226.

[18] 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*, Pasadena, CA, 2009, 1261-1266.