APPLICATIONS OF SPARSE RECOVERY AND DICTIONARY LEARNING TOWARDS ANALYSIS OF ELECTRODERMAL ACTIVITY

by

Malia Kelsey

B.S in Bioengineering, University of Maryland, 2013

Submitted to the Graduate Faculty of

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

Malia Kelsey

It was defended on

October 28, 2016

and approved by

Murat Akcakaya Ph D., Assistant Professor, Electrical and Computer Engineering

Ervin Sejdić Ph D., Assistant Professor, Electrical and Computer Engineering

Zhi-Hong Mao Ph D., Associate Professor, Electrical and Computer Engineering

Thesis Advisor: Murat Akcakaya Ph D., Assistant Professor, Electrical and Computer Engineering

APPLICATIONS OF SPARSE REVOCERY AND DICTIONARY LEARNING TOWARDS ANALYSIS OF ELECTRODERMAL ACTIVITY

Malia Kelsey, M.S.

University of Pittsburgh, 2016

Electrodermal Activity (EDA) – a measure of sympathetic nervous system arousal – is one of the primary methods used in psychophysiology and has been used to investigate a variety of physiological topics. EDA signals can be separated into two distinct parts, skin conductance level (SCL) and skin conductance response (SCR). The SCL is a slowly fluctuating response that reflects general trends in activity and level of activation. In contrast SCRs are quick responses which can be seen as peaks in the signal and typically reflect responses to a specific stimulus from the nervous system. While many existing studies collected EDA data in short, laboratory-based experiments, recent developments in wireless biosensing have allowed for out-of-lab studies to become more common. The transition to ambulatory data has introduced challenges in SCR and artifact identification and may hinder analysis of ambulatory data. Therefore, the interest in developing automated systems that can facilitate the analysis of EDA signals has increased in the recent years. Ledalab, a current gold standard, is one such system which can be used to identify SCRs. However, Ledalab is computationally inefficient when applied to long, ambulatory data and does not have the ability to distinguish between SCRs and artifacts. This thesis presents a novel technique that can be used to accurately and efficiently identify SCRs using curve fitting and sparse recovery methods, namely orthogonal matching pursuit. Sparse recovery can also be easily extended to include identification of both SCRs and artifacts given seperability between the two. In this work we apply discriminant analysis to determine if seperability exists between SCRs and artifacts. We have shown that our novel approach was able to detect 69% of the SCRs in an EDA signal compared to the 45% detection of Ledalab. Our method also exponentially decreased the computational complexity compared to the Ledalab. Finally, our system begins addressing the issue of artifact detection by determining that the difference between artifacts and SCR shapes could be distinguished with an accuracy of 74%.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

First I would like to thank my advisor Dr. Murat Akcakaya for all of the support he has given me, both technically and personally, since starting my masters work. I am gratefully for all of the advice and suggestions he has made throughout this entire process. I would also like to thank Dr. Ervin Sejdić and Dr. Zhi-Hong Mao for being part of my committee to review my thesis work.

I would also like to thank all of our collaborators from Northeastern University who provided the data used for our study and have worked closely with us through our various projects. The data came from research supported by the National Institute of Mental Health post-doctoral award (F32MH096533) to I.R.K., and the National Institutes of Health Director's Pioneer Award (DP1OD003312) to L.F.B. Additionally, all our collaborators have given endless advice and suggestions and have helped to make this project successful. Without them these projects would not have been possible. I am also grateful for all the time they have spent reviewing and editing my papers. Thanks to Matthew Goodwin, Ian Kleckner, Richard Palumbo, Christophe Gerard, and Karen Quigley for all of the support and data you provided.

Finally, I would like to thank everyone I have met and worked with since arriving in Pittsburgh, my friends who have always supported me, and my family. Their support, advice, friendship, and love is invaluable.

1.0 INTRODUCTION

Electrodermal Activity (EDA) - a measure of sympathetic nervous system arousal - is one of the primary methods employed in psychophysiological research due to its sensitivity to the autonomic nervous system, the relative ease in which it can be collected, and the non-invasive nature of collection [1], [2]. EDA is widely accepted as an indicator of arousal as it has been shown to reveal when psychologically novel events occur [3], [4], [5]. For over a century, EDA has been used in many different studies to investigate a variety of physiological topics, including stress, depression, anxiety, attention and information processing [6] [7], [8], [1]. While EDA is highly sensitive to a variety of psychophysiological states, accurate modeling of those states can be complicated by non-psychological signal effects associated with increased physical activity, humidity, changes in air temperature, electrical noise, and wearer movement [4], [9]. The presence of these artifacts can obscure or confuse psychological inference if not carefully controlled. Formally, EDA is a measure of electrical conductance on the skin surface, which changes as eccrine sweat glands release and absorb salt water [10].

The skin is a vital organ that provides numerous essential bodily functions, including acting as a barrier between external exposure and internal viscera, passing material through the blood stream, maintaining water balance, and thermoregulation of body temperature [1]. The top layer of the skin is the epidermis followed by the dermis underneath and the subdermis as the lowest layer [1]. Sweat glands consist of a coiled body that rests in the subdermis with a relatively straight

duct that extends through the dermis and epidermis up to the surface [1]. There are two different types of sweat glands in the body, the apocrine, which has been less studied, and the eccrine, which is the gland of primary interest in the study of EDA [1]. The primary function of eccrine sweat glands is thermoregulation accomplished through evaporation of sweat from the skin. However, when using EDA to model psychological states, the primary area of interest is cognitive and affective induced sweat gland activity [1]. As mentioned previously, the eccrine sweat gland consists of a coiled body, which is the secretory portion of the gland, and a sweat duct which acts as the excretory portion of the gland [1] (see Figure 1). Previous studies conclude that changes in EDA can be modeled entirely based on the sweat gland activity [11]. It can easily be seen how EDA signals relate to sweat gland activity if one thinks of the sweat ducts as a set of variable resistors wired in parallel [1]. Proportional to the degree of physical, thermoregulatory, and psychological activation, columns of sweat rise, in varying amounts and through varying numbers of ducts. As the ducts slowly fill, a more conductive path is created through the relatively resistant epidermis and changes the value of the variable resistor which is seen as an observable change in the EDA response [1].

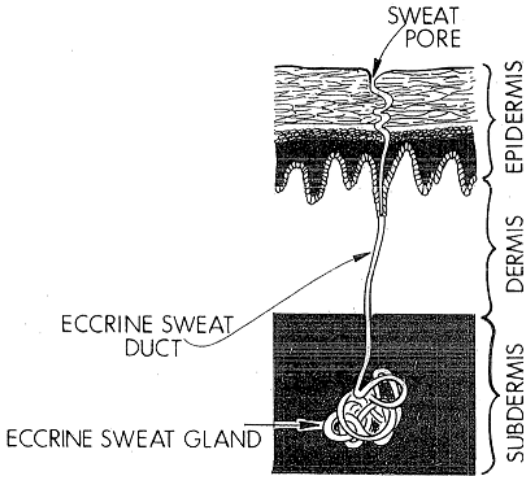


Figure 1: Make-up of skin and configuration of sweat glands.

1.1 ELECTRODERMAL ACTIVITY COMPONENTS

EDA signals are sensitive to physiological activity, however, due to EDA being produced by fluctuations in sweat gland activity, EDA signals are also highly susceptible to artifacts caused by changes in physical activity, temperatures, and pressure [12], [13].

EDA signals can be separated into two parts distinct parts, skin conductance level (SCL) and skin conductance response (SCR). SCL and SCR have been shown to contain information associated with different aspects of the brain and information processing [14]. Therefore, separating these components can be useful during analysis. The SCL, or tonic response, is a slowly fluctuating response that reflects general trends in activity and level of activation. For example, SCL will show a rise during periods of high activity and a decrease during periods of rest [3], [10], [1]. SCL baselines can also vary widely, ranging between 2 and 20 μS. Further due to factors, such as hydration status, recording site, and psychological state, baselines are not consistent within or between individuals [1]. Thus, the same person in the same state could have a different SCL baseline level. In contrast SCRs are quick responses, superimposed on the tonic response, which can be seen as peaks in the signal and typically reflect responses to a specific stimulus [10], [15]. Unlike SCL, SCRs typically have a predictable shape which can be characterized by rise time, amplitude, and half recovery time. In healthy adults, rise time is usually between 1 and 3 seconds, amplitude varies, but a minimum is commonly set between .01 and .05 μS, and half recovery time is typically between 2 and 10 seconds [1]. Figure 2 shows the typical shape and parameters used to describe an SCR. A complicating factor occurs when a second SCR is elicited before the previous SCR has fully recovered. This case, referred to as a compound SCR, indicates two separate stimuli or psychologically different events have occurred [13]. As compound SCRs may be caused by different stimuli, accurate identification of each SCR is important during analysis.

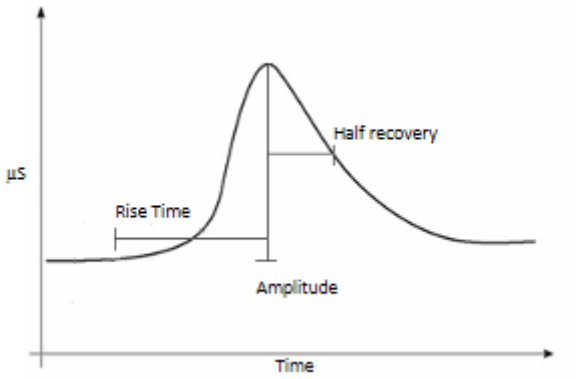


Figure 2: Parameters used to characterize SCRs

1.2 DATA COLLECTION METHODS

Until the early 2000's, most studies that employ EDA were conducted in a laboratory setting with few long term, out-of-lab or ambulatory signals being collected [8]. Ambulatory recordings can offer greater insight into whether findings from short lab recordings extend out to daily life providing ecological validity to studies [16]. For example, a study investigating panic disorders found that SCL trends in participants with panic disorders were normal when recorded in a lab setting but were significantly elevated during longer ambulatory recordings [8]. Additionally, ambulatory recordings allow for longer recording times which can provide more data for analysis and allow for longer trends to be seen [3]. Advances in wireless biosensing have allowed for more studies to be conducted in ambulatory settings but challenges with artifact identification and general signal analysis may have hindered full analysis of these signals [8]. As wireless biosensing has become more prevalent, many groups have started to develop automatic systems that can be used to identify SCRs and artifacts and provide an analysis of EDA signals. Many different systems have been developed using a variety of methods ranging from simple curve fitting to more complex systems that involve machine learning principals however, these systems were mainly developed

to analyze lab collected signals [17], [6], [18]. The shift towards ambulatory collected data has led to an increase in the need for efficient and accurate SCR and artifact detection that current systems still fall short of achieving.

1.3 ANALYSIS REQUIREMENTS

When analyzing EDA signals that are collected either in a lab based setting or in an ambulatory setting there are two main goals

1. Analysis methods should be able to robustly identify SCRs throughout the signal regardless of the SCR shape, amplitude or SCR length
2. Since SCRs are highly susceptible to artifacts analysis methods should be able to distinguish between an artifact and a true SCR. Artifacts include, but aren't limited to:
    - Patient movement or touching the sensor
    - Changes in temperature, humidity, or other environmental conditions
    - Physical activity
    - Electrical noise

Many different methods currently exist that address one or the other of these goals (see sections 1.0). However, to our knowledge, no currently available system or proposed method addresses both. Additionally, signals collected in ambulatory settings allow for data to be collected for long time periods but many of the current analysis methods that do address the first issue are slow, especially for longer signals. Finally, while there are a few proposed methods to address the second

goal, the similarity in the shape and phase of many artifacts to SCRs, makes this a challenging endeavor using current practices. To further the utility of EDA data in-vivo, the following manuscript presents a novel strategy for identifying SCRs and removing artifacts through sparse recovery methods.

1.3.1 Our Contributions

Improving the ability for analysis to be completed efficiently and accurately on Electrodermal Activity data will enhance the ability for longer EDA studies to be conducted and for the data to be fully analyzed. To be able to accomplish this we aim to develop a system using sparse recovery and dictionary learning with the following goals:

1) We aim to show that using orthogonal matching pursuit to analyze long ambulatory collected signals will reduce the computational complexity of analysis, compared to the current gold standard Ledalab, without sacrificing performance.
2) We aim to show that using a data driven dictionary improves the performance of SCR identification over the use of a knowledge driven dictionary.
3) In an attempt to generalize the dictionary, therefore limiting the dictionary size, we will investigate the removal of the tonic level using different techniques. We then compare our estimation to the estimation performed by Ledalab.
4) We provide a holistic view of our methodology to Ledalab by reporting on the performance measures of both systems, which, to our knowledge, has yet to be done.
5) Finally, we aim to investigate the classification ability between artifacts and SCRs, using the Bateman equation parameters as features, to determine the possibility of including artifact columns in the dictionary.

1.4 ORGANIZATION OF THIS THESIS

The outline of this thesis is as follows: In Chapter 2 we will present the current state of the art methodologies used to analysis EDA data. This chapter presents the different models the can be used to fit to the shape of a SCR, the current automatic systems that can be used to identify SCRs, and current methods being used to identify and remove artifacts. Chapter 3 will present and describe the novel methodology we propose for EDA analysis. Chapter 4 will present the analysis completed on experimental data including details about the data, procedure, and results and discussion relevant to each study. Chapter 4 is divided into 3 different subsections corresponding to the 3 different experiments run to evaluate our methodology. The first experiment was a proof of concept to evaluate the utility and computational complexity of our methodology with ambulatory data. The second study builds upon the methodology used in the first experiment and evaluates the performance using expert labeled lab data as our ground truth. Finally, our last study investigated the seperability of artifacts and SCRs using minimal parameters to model each. The last experiment again used expert labeled data. Finally, conclusions and future work are presented in Chapter 5.

2.0 STATE-OF-THE-ART EDA ANALYSIS

2.1 MODELING SCRS

Historically, EDA signals were analyzed by hand and, in fact, the Society for Psychophysiological Research (SPR) still recommends manual analysis to identify SCR locations and remove artifacts [13]. Manual analysis however, is time consuming and prone to human error and variation. As a first step towards automatic analysis, many groups have developed different models to represent the shape of an SCR. These models are typically based on the underlying physiology that creates SCRs and have varying levels of complexity [15].

2.1.1 LTI systems

Several studies investigating the models that can be used to fit SCRs throughout a system make assumptions that EDA is produced by a linear time-invariant system (LTI) [15], [19], [20]. Being able to model EDA as a LTI system is a standard concept in signal processing and typically is used to simplify analysis [15]. Previous work showed that using and LTI assumption holds for modeling SCRs and therefore several assumptions about EDA data can be made. Under the assumption of linearity SCRs can be deconvolved into individual responses even if the SCRs are overlapping [19]. Using the time-invariant assumption, a majority of SCR response can be modeled using a

single model. This model can then be scaled to fit the SCRs throughout a signal [15]. After showing the EDA can be modeled as an LTI system, SCRs were modeled by quantifying the area under the curve (AUC) of responses in a signal [19]. Using the area under the curve as the basis for identifying SCRs within a signal comes with two major limitations (1) finding the AUC requires either an integral or convolution of the data [19], [20] which will be difficult to scale for longer signals and (2) noise will have a large negative affect on the performance [19]. While finding AUC will not be scalable to long signals being able to model EDA as a LTI system will be useful in the general signal analysis.

2.1.2 Sigmoid – Exponential

A major concern of many different proposed models is the ability to successfully identify compound SCRs [2]. Many traditional lab base studies employ inter-stimulus interval paradigms which would elicit compound SCRs. Even moving to more ambulatory collected data the ability to accurately model compound SCR is important so that all responses can be successfully captured [2]. To be able to accurately capture a variety of different compound SCRs combinations of asymmetrical sigmoid and exponential functions were used to model an SCR [2]. These models range in complexity from a four-parameter model to an eight-parameter model. The different numbered parameter models were each developed to match to a different complexity of SCRs being modeled with more parameters being added as more compound SCRs were added to the analysis. The three-parameter model is shown below in equation (1)

$$f(t) = a_0 e^{-\frac{t}{t_d}} + f_{s1}(t) + f_{s2}(t) + c \quad (1)$$

Where t_d is the decay factor, f_{s1} and f_{s2} account for each individual SCR that makes up the total compound response (in this case 2) and c is used to represent the tonic level [2]. This model was successfully able to capture SCRs with little error but the complexity makes it more computationally complex than other models. Additionally, since each individual SCR requires a new parameter prior knowledge about the signal is need.

2.1.3 Bateman Equation

A popular model that has been used in several recent studies is the Bateman equation, shown below

$$b(t) = e^{\frac{-t}{\tau_1}} + e^{\frac{-t}{\tau_2}} \quad (2)$$

In (2), t is time and $\tau_1$and τ_2 are parameters that characterize the shape of the function. The Bateman function is characterized by a steep onset followed by a slow recovery period; the steepness and recovery of the function are controlled by the constants τ_1 and τ_2 respectively [14]. Because the Bateman equation relies on only two parameters, minimal computation complexity is required to estimate optimal parameters to fit to an SCR making it ideal for different SCR detection software [10], [21]. Several studies have also shown the Bateman equation to be a good fit for modeling SCRs. Using this model as the basis for an SCR shape, several groups have created software capable of determining SCR locations throughout a signal; however, most of these methods were developed for short, laboratory-based studies and have not been optimized for longer ambulatory recordings [17]. One such group is the Leipzig Electrodermal activity laboratory, which developed Ledalab for the analysis of EDA signals. Ledalab's, along with other currently available software's, methodology and shortcomings will be discussed next.

2.2 SCR DETECTION

Most of the existing software programs used to identify SCRs in a signal fall into two categories; 1) data collection plus analysis systems or 2) model based analysis. As an example Mindware and AcqKnowledge are two systems that fall into the first category [22] while SCRalyze and Ledalab are two model based systems [17], [18]. Additionally, other methods that do not fall into the above 2 categories have been proposed for detecting SCRs. Some of these methods include the use of peak detection with an automatic scoring system and machine learning principals [23], [24]. As these methods are not used as widely for the analysis of EDA data the focus will be more into the two main categories which are accepted as the gold standard analysis methods.

2.2.1 Data collection Systems

While the systems that fall into the first category allow for SCR detection, they were designed initially for data collection which means that the main focus is towards lab collected signals [22]. User interfaces to visualize the data and allow researchers to locate SCRs were then added to these systems as secondary function. Many of these programs will use the stimulus onset with peak detection to locate potential SCRs in the signal [22]. The use of the stimulus means that these types of programs cannot be used to analyze ambulatory data which does not present specific stimuli. One of the major benefits of these programs is that most of the systems allow for collection of multiple types of signals which allows for cross analysis of the signals to be performed [22]. The major limitations of these systems is that as data collection shifts to ambulatory collection the systems can no longer be used for analysis due to the lack of a specific stimuli [22].

2.2.2 Model based Systems

Model base systems will generally use similar methodologies to identify SCRs based on convolution of the raw data with a chosen function that can be used to model the shape of SCRs. Causal models are typically used to map the underlying processes (namely the sympathetic nerve activity) to the empirical observation made in the EDA data [25] so that inferences can be made relating the detected responses to activations of the nerves. After the selection of an appropriate model most model based systems use some sort of convolution to identify the locations of SCRs within a signal. Different systems employ different types of convolutions to better match the assumptions made about the process which causes conductance fluctuations to be seen [25]. However, as the focus in our work was a comparison with Ledalab we will focus on only the introduction of the methods employed by that system. As other systems do exist we wanted to present them but a few comparisons between the model based systems have previously been made [17] and it was shown that these methods have similar performance (and similar methodologies) making more extensive comparison unnecessary.

2.2.3 Ledalab

General Info

Ledalab is a MATLAB based program developed by the Leipzig Electrodermal activity laboratory to be used for the analysis of EDA data. Ledalab is currently considered a gold standard for computationally processing and analyzing EDA, therefore, we chose it as the basis of comparison for the proposed methodology introduced in this thesis. Ledalab returns optimal parameters, which describe SCR shape, number of SCRs present, and onset time for each detected

SCRs [10]. To identify possible SCR locations, Ledalab models SCR shapes based on a diffusion model previously introduced [21]. This method proposes that the underlying processes driving a SCRs shape is caused by the diffusion of sweat passing through the corneum, the top most layer of the skin, and distal parts of the sweat ducts [21]. This lead to modeling the shape of an SCR as a two compartment diffusion model with coupled first-order differential equations. With the assumption of forward only diffusion through the ducts, SCRs can then be modeled using the Bateman function [21] introduced in (2)

Recall from (2) $\tau_1$and τ_2 are parameters that help characterize the shape of the function, specifically, the Bateman function is characterized by a steep onset followed by a slow recovery controlled by the constants τ_1 and τ_2 [18]. Using the assumption that EDA data can be viewed as the result of a driver function that triggers an impulse response, deconvolution with a known impulse response can retrieve the original driver function [18]. In this assumption, the driver function represents the activity of sweat glands while the impulse response represents the increased conductivity of the skin due to sweat perfusion [21]. Using these definitions, the impulse response used in deconvolution is the Bateman function. To perform the overall analysis of a signal, Ledalab estimates the tonic component, uses nonnegative deconvolution of the phasic data (to determine a nonnegative driver function and remainder), segments the driver and remainders to identify single impulses, and finally reconstructs the EDA data [18]. This procedure is initially run with preset parameters, τ_1 = 0.75 and τ_2 = 20 [18] and then an optimal τ_1 and τ_2 are found using an optimization procedure to increase goodness of fit for individual data sets. Each of the individual pieces involved in Ledalab's analysis are described in more detail below.

Tonic Estimation

As previously discussed the tonic component of an EDA signal is removed by Ledalab leaving only the phasic component for analysis. To estimate the tonic component Ledalab first deconvolves the signal using the Bateman equation as the impulse function [18]. Deconvolution serves to remove phasic activity from the signal leaving only the driver function, or estimated tonic component. Peak detection is then carried out to identify additional possible impulses in the signal. Peak detection is achieved by finding zeros in the first time-derivative of the driver function. All time sections that are not part of impulse responses are considered driver functions and therefore the tonic component [18]. The tonic activity is then estimated using a 100-second time grid and averaging the driver function values within the impulse sections over each grid section [18]. Finally, a cubic spline fit is used to interpolate the tonic activity based on the created grid. Once the tonic component is estimated, it is subtracted from the original signal to yield the phasic estimate [21]. An important part of phasic estimation is that it must remain non-negative to match assumptions Ledalab makes about the physiological process being modeled. To keep the phasic estimate non-negative, a technique known as non-negative deconvolution is used during the initial deconvolution process [18].

Non Negative Deconvolution

As previously stated to estimate the tonic level the signal is first estimated by deconvolving the original signal to retrieve the driver function. Deconvolution was chosen to analyze EDA signals since skin conductance can be viewed as the result of the activity of sweat glands (driver function) triggering increased conductivity of the skin (impulse response). This definition of EDA data, however, is based on 2 assumptions being true. 1) The sudomotor neurons which cause the increase in sweat profusion are either active or inactive [18] and 2) a single SCR corresponds to a

discrete burst of sudomotor activity. Looking at the first assumption it can be seen that since the activity of sweat profusion is either active or inactive the driver function should either have positive deflections or should be zero. The second assumption implies that impulses should be compact in time having an explicit onset and offset which can be marked in time [18]. Using traditional deconvolution however doesn't match the positivity assumption. Traditional convolution, if thought of in the terms of long division, can be expressed using equation (3) below

$$f = g * h + r \tag{3}$$

In (3) f is the original signal, g is the result of the driver function, h is the impulse response applied, and r is the remainder which would be produced from long division. In traditional deconvolution the remainder is assumed to be zero meaning that g can be solved for by dividing the signal by the impulse response, $g(i) = f(i)/h(i)$ [18]. Assuming that r is zero yields negative responses which does not fit with the initial assumptions made. To produce a nonnegative driver function a convolution, termed nonnegative deconvolution, is instead used. Nonnegative deconvolution relaxes the assumption that the remainder is zero allowing $r(i) \geq 0$, which means the deconvolution equation becomes

$$g(i) = \min(\frac{f(j)}{h(j)}) \tag{4}$$

Where $j = i\ to\ i + k - 1$ and k is the length of h. Using nonnegative deconvolution ensures that a nonnegative driver is produced which matches with the initial assumptions made about EDA data [18].

Optimization

After the signal is initially decomposed and then reconstructed, parameters used in the initial run are optimized through the use of the gradient descent method. For optimization, several conditions are used which play into the overall cost of the system [18]. Using individual parameters, the composite criteria for optimization is defined as

$$\mathrm{c} = \mathrm{RMSE} * (\mathrm{n} + 1) * \mathrm{d} \tag{5}$$

In (5) RMSE is the root mean square error between the reconstructed data and the original signal, n is the total number of SCRs found in the signal, and d represents the deflection of the signal. The first criteria for optimization is that deflections of the driver and remainder of the signal should approach zero between deflections. The deflection of the driver and remainder of the signal is measured by looking at the discreteness of the signal represented as d in the cost function [18]. Since the deflection of the remainder and driver should approach zero between SCRs d is counted as the number of succeeding time points with values above a threshold of .2 for the driver and .005 for the remainder. The length of these successive sections are then found and squared; the individual lengths are then summed and divided by the total signal length [18]. When there are long time sections that exceed threshold values, calculated *d* values are high increasing the overall cost. The second condition considered for optimization is the number of phasic responses, n. A higher number of phasic responses allows for a closer fit with the data, but also opposes the parsimony of the model. Finally, the last condition for optimization is the root mean square error (RMSE). RMSE between the reconstructed data and original signal is used to indicate overall goodness of fit of the reconstructed data [18]. To perform optimization, c is minimized based on (5) through gradient descent and returns an optimal reconstructed phasic estimate and optimal τ_1 and τ_2 for the signal [18].

System Limitations

As previously discussed, Ledalab estimated the tonic level by using nonnegative deconvolution. This unique form of deconvolution involves assuming that the activity of the sweat glands produces single SCRs which are compact in time [18]. Using this assumption when estimating the tonic and phasic components of the signal means that potential compound SCRs in the signal will not be captured. Additionally, the nonnegative assumption also increases the peak amplitude of individual SCRs during the estimation procedure. Increasing the amplitude of SCRs could cause SCRs which do not meet the minimum amplitude requirements to be labeled as SCRs. To address the concerns of using Ledalab's tonic estimation methodology a new tonic estimation procedure needs to be investigated. An additional limitation of Ledalab is that optimization procedures are required to optimize the analysis. As previously mentioned, to identify SCRs in a signal, Ledalab uses the Bateman equation, with preset values for $\tau_1$and τ_2, to identify the onsets of individual SCRs. To improve the goodness of fit, Ledalab then uses gradient descent to optimize the $\tau_1$and τ_2 parameters across the signal [18]. This methodology is slow due to the optimization process and not robust towards artifacts, making it difficult to scale for longitudinal ambulatory signals. To move towards a more scalable and robust solution for identifying SCRs in ambulatory signals, sparse recovery methods has been proposed and shown in previous work to be a promising solution although has not been used to create or release a full EDA analysis software package.

2.2.4 Sparse Recovery

Sparse recovery has been proposed as an ideal way to model EDA data due to the flexibility sparse recovery algorithms have. Additionally, EDA signals can be represented as sparse signals due to the inherent sparsity of responses typically seen in EDA data. When a participant is at rest and no

specific stimuli is given, with the intension of producing a SCR response, SCRs are typically seen at a rate of 1-3 responses/min [10]. Recent studies have also shown that a sparse recovery algorithm could be used to fit to the tonic and phasic components of a signal [26]. Before introducing this study, a brief background of sparse recovery is presented.

General Information

Over the last ten to fifteen years there has been increased focus on representing a sparse signal in ways other than traditional Fourier transform representations [27], [28]. This has led to the creation of many different waveform dictionaries that can be used to represent various signals. A dictionary is represented by a matrix of individual columns, typically called atoms, which represent a specific waveform [29]. These atoms can be linearly added together to form an approximation of the original signal [30]. There are many standard dictionaries, such as Wavelet dictionaries, Gabor dictionaries, and Chirplet dictionaries, to name a few, that have been developed for general signal analysis [27], [31]. While these redefined dictionaries are typically easier to use and lead to fast algorithms they may led to poor reconstructions [32]. Therefore, many dictionaries are created through sample data and adaptive training methods to create atoms that will be representative of either a process or a specific type of signal [29]. Using learned dictionaries as opposed to redefined dictionaries have been shown to perform well with processing natural signals, particularly towards denoising these signals [31].

One method to select the atoms to make up the estimate of a signal is through sparse recovery; a technique that can be used to estimate a signal by linearly adding columns from a dictionary of predefined waveforms [27], [33]. The general equation being solved by sparse recovery techniques is shown below.

$$D\gamma = x \tag{6}$$

D is the dictionary, γ is the coefficient matrix that gives the weight of the selected atoms, and *x* is the original signal [27], [33]. The dictionary, as previously stated, is a matrix built up of atoms each representing a specific waveform. While there are many standard dictionaries with predefined waveforms, many new applications will build a dictionary containing application specific atoms which better represent the original signal [34]. In many applications, the chosen dictionary will be over-complete, so choosing a set of vectors from a large, and sometimes redundant, set is NP hard [35], [31] and can require combinatorial optimization to solve for an exact solution [30]. Many different greedy heuristic algorithms have been developed to reduce the computational complexity of finding sparse signal approximation [35]. The most popular methods fall into two categories; a greedy approach that approximates the signal through a sequence of incremental updates, termed matching pursuit algorithms (MP) and the basis pursuit method which approximates the signal using linear programing methods [30]. Greedy approaches constrain the problem using an L_0 norm constraint, which solves for the approximation that has the fewest nonzero elements possible [30]. The basis pursuit method relaxes the L_0 norm condition assumed in the greedy approach to an L_1 approximation which forces the algorithm to be nonlinear [36].

Basis Pursuit

One way to approximate a sparse signal is to use a basis pursuit (BP) that relaxes the L_0 norm condition employed in MP algorithms for an L_1 norm assumption, and then solve the equation through linear programming [27]. Relaxation of the L_1 norm forces the algorithm to be nonlinear, making optimization more difficult, and for many years impossible to solve for long or complex signals. The Basis Pursuit method became possible in the early 1990's when the interior point method was developed, allowing optimization to be completed in a reasonable amount of

time [27]. The basis pursuit algorithm aims to pick the dictionary coefficients that have the minimum L_1 norm using the following optimization [27].

$$\min \left\| \underline{\gamma} \right\|_1 \quad \text{Subject To } D\underline{\gamma} = x \tag{7}$$

In (7) γ again represents the estimated coefficient vector, $\|.\|_1$ denotes the L_1 norm, D is the dictionary, and x is the original signal [27], [32]. Basis Pursuit algorithms allows for a more accurate match of the original signal, but requires higher computational complexity to compute [36]. The tradeoff between computational complexity and accuracy of the approximation is important to consider when choosing an algorithm to complete sparse recovery. In most cases, MP algorithms converge faster than BP methods [36], [32]. For our application we chose to use a MP algorithm, specifically an orthogonal matching pursuit algorithm, given the method's potential speed coupled with the relative accuracy found in OMP algorithms. While it has been shown that BP algorithms typically produce more accurate matches [36], this is not as much of a concern when matched with the speed that can be gained using MP. Since we are using the recovery algorithm to index the locations of SCRs, we are not concerned with the approximation being a perfect match to the original signal as long as the match picks up peaks in the signal that correspond to the locations of SCRs.

Matching Pursuit

There are many different variations of greedy approaches which allow equation (6) to be solved as an approximation with reduced computational complexity and faster run times [27]. The most popular greedy approaches fall under the category of Matching Pursuit (MP) algorithms, all of which follow the same basic algorithm with only slight variations [32], [37]. To create an

estimate of the original signal, MP algorithms aim to solve the optimization problem shown in equation (8)

$$\underline{\gamma} = \underset{\underline{\gamma}}{\text{Argmin}} \left\|\underline{\gamma}\right\|_0 \quad \text{Subject To} \left\|x - D\underline{\gamma}\right\|_2^2 \leq \epsilon \tag{8}$$

In (8), γ is the estimated coefficient vector, $||.||_0$ represents the L_0 norm, x is the original signal, and D is the dictionary. Additionally, $D\gamma$ is the estimated representation of the original signal [34], [28], [32]. As a greedy approach MP algorithms solve the optimization equation through a series of iterative steps which can be generalized to 2 main steps. The first step is the atom selection step. Here, an atom from the dictionary is selected based on the atoms correlation to the current residual error. The second step is the residual update step. Here, the error term is updated to reflect the newly added atom [27], [38]. These steps are repeated until the stopping criteria is satisfied. The stopping criteria is typically an error based criteria but other criteria, such as sparsity based, have also been introduced [38]. For the traditional matching pursuit algorithm, atom selection is performed over all vectors in the dictionary at each step, meaning it is possible for atoms to be selected multiple times during the analysis, termed atom re-selection [35]. This decreases the rate of convergence of the algorithm and causes the output signal to suffer from poor sparsity results [30], [32]. One methodology that improves upon the traditional MP algorithm is the orthogonal matching pursuit algorithm (OMP). This procedure introduces an orthogonalization step between the atom selection step and the residual update step to avoid atom re-selection. After an atom is selected, the OMP algorithm projects that atom into an orthogonal space; this reduces that run time of the algorithm and enforces better sparsity of the estimate since the same atom cannot be selected more than once [34]. A traditional OMP algorithm's pseudo code is shown in Figure 3

1 Input: Dictionary (**D)**, phasic signal (x), target sparsity (K)

2 Initialize: Set $I := ()$, $r := x$, $\gamma := 0$

3 while sparsity < K

4 $\quad \hat{k} := \text{argmax} \left| d_k^T r \right|$

5 $\quad I := (I, \hat{k})$

6 $\quad \gamma_I \coloneqq (\mathbf{D_I})^+ x$

7 $\quad r = x - \mathbf{D_I} \gamma_I$

8 end

9 Output: Sparse representation γ such that $x \approx \mathbf{D}\gamma$

Figure 3: Pseudocode for a traditional OMP algorithm

In the above pseudo code I represents the indices of the selected atoms which is used to update γ, r is the residual at each iteration and $\boldsymbol{D\gamma}$ is the sparse representation of x [34]. These same variables will be used throughout the manuscript when discussing sparse recovery and OMP algorithms.

OMP and EDA analysis

Orthogonal matching pursuit (OMP) was proposed as an ideal MP algorithm for EDA analysis [26] and has been shown to perform well for fitting to EDA signals collected in lab settings. Chaspari et al. showed that using a knowledge driven dictionary allowed them to achieve good fits to lab-collected EDA signal with high accuracy [26]. They also showed the ability to identify SCRs in the signal using the OMP methodology with a dictionary made up of columns to represent both the tonic and phasic components and some post processing of the selected atoms. This study showed that OMP with a knowledge driven dictionary could be efficiently used to model EDA signals and detect SCRs in the signal; however, the study did not consider (1) the computational complexity or run time required for the design, (2) the use of a data driven or learned

dictionary, although the possibility of using this type of dictionary was discussed, (3) a comparison with an existing, gold standard, software including fully reporting on the performance measures achieved by the OMP methodology or (4) artifact detection or any discussion about how artifacts could affect the design. Additionally, the study didn't consider the possibility of removing the tonic level which could have an impact on the overall performance. The tonic level is much more variable than the phasic response [1] meaning it could make the dictionary more difficult to generalize to new data as well as significantly increasing the size of the required dictionary; therefore, increasing the computational complexity.

2.3 ARTIFACT DETECTION

For most studies involving EDA data, artifacts are removed either by applying smoothing or through filtering [6], [12]. These studies will typically use these techniques in conjunction with manual review of the data which makes the process time consuming and prone to human error [6], [7], [16]. For the typical smoothing a variety of filters and windowing length have been used, including hamming windows, exponential, and median [6], [16] with window lengths varying from 1 second up to 1 min. Filtering is typically done with a low-pass filter with cutoff frequencies ranging between .35 Hz up to 1 Hz [7], [12] , [24]. To our knowledge no compressive study has been to study the different filtering and smoothing techniques.

While the above approaches have been shown to work for many artifacts, high magnitude artifact are not effectively removed using either of these methods [12]. In fact, high magnitude artifacts can end up looking more like artifacts after the smoothing or filtering is performed making manual analysis difficult [24]. As experiments shift to more ambulatory data collection, the

potential for high magnitude artifacts increases due to factors including participant movement and participants touching the sensor. Traditional methods for artifact detection require either manual inspection of the data after some processing, which is time consuming and prone to varying interpretations, or through collection of EDA from multiple sites (such as the ankle and wrist), which may not be possible in out of lab collections [12], [4], [16]. A recent study attempted to address these issues and find more robust ways to identify artifacts using machine learning techniques such as support vector machines (SVM) and naïve Bayesian classification [12], [24]. It was found that SVM with a radial basis function was successfully able to distinguish between clean and nosy sections of data with test accuracies of about 96% [12], [24]. While this method addresses the issue of missing high magnitude artifacts, it performs classification by looking at sections of data instead of being able to classify individual SCRs or artifacts. Potential issues with sectioning the data are 3 fold: (1) If there are multiple responses in a section (including SCRs and artifacts) there is no way to classify the responses individually, (2) sectioning the data could miss responses if an SCR or artifact falls between two sections (i.e. the onset of an SCR could be at the end of one section and recovery in the beginning of the next), depending on how sectioning is handled, and (3) simply identifying sections of noisy data may still require manual clean up before further analysis can be completed.

3.0 METHODOLOGY

In an attempt to address the needs required to fully analyze Electrodermal Activity Data (EDA) collected in both ambulatory and lab settings our novel method proposes to use an updated orthogonal matching pursuit (OMP) method with a dictionary built to represent individual SCRs. To accomplish the research goals outlined in section 1.3.1 we split the tasks into three distinct experiments. Our first experiment was used as a proof of concept to test our proposed methodology and evaluate its speed and complexity when applied to ambulatory data. This work was presented at a conference last spring [39]. Our second experiment built upon and improved our initial methodology to better be able to capture SCRs of varying shape and length. Our second experiment also introduced a new tonic estimation methodology to be used in tangent with our analysis. Our final experiment investigated the fit of artifacts and evaluated the ability to distinguish between an artifact and SCR. The following sections break down the basics of the methodology which will be used throughout our experiments. Sections 4.1.3 and 4.2.3 also provide experimental specifics for our overall methodology.

3.1 TONIC ESTIMATION

To estimate the tonic several different methodologies were used throughout the various experiments. For our initial experiments we used the tonic estimation method developed by Ledalab as presented in section xx. As previously mentioned Ledalab estimates the tonic level by first deconvolving the signal with the Bateman function using predefined τ_1 and τ_2 parameters. After the initial estimate is completed optimization is used to redefine the τ_1 and τ_2 parameters and then the tonic level is re-estimated get a better estimation of the tonic level [18]. To test the necessity of the optimization piece of Ledalab's tonic estimation we used Ledalab's methodology with 0 and 1 round of optimization and compared both methods.

During our initial runs we determined that Ledalab's estimation was not robust towards all signals, see section 4.1.5 for full discussion. To produce a better tonic and phasic estimate across different signals we introduce a new tonic estimation methodology in our second experiment and compare our methodology to Ledalab to determine the benefits and drawback of both methods.

3.1.1 Our tonic estimation methodology

To estimate the tonic and phasic responses of each signal, the first step was to find local minima throughout the data. A strict search pattern was employed for minima identification, meaning that a minimum was only identified if the directional derivative was greater than zero on both the right and left edges. Once all of the local minima were located, they were filtered using a threshold to set a minimum time distance required between minima, M_T. Using this threshold each minimum was checked in relation to the previous minimum and, if the difference was not greater than M_T, the second minimum was removed. After the minimums were filtered, interpolation through the

minimums was done using a linear fit to create the tonic estimate. Finally, the tonic estimate was subtracted from the original EDA data to obtain an estimate of the phasic component.

3.2 DICTIONARY CREATION

The initial step for any MP algorithm is to select or create an appropriate dictionary. Instead of selecting a predefined dictionary for our analysis, a dictionary was created based on a priori knowledge of the shape of an individual SCR [40]. Each individual column in the dictionary represents a single SCR whose shape is defined by the Bateman function, $\tau_1 = .75$ and $\tau_2 = 20$, for a 30-second-time window. These parameters lead modeling each SCR using $f(\theta) = e^{\frac{-t}{20}} + e^{\frac{-t}{.75}}$. The SCR onset time is then shifted in each column by one sample so that an SCR can be correlated to an SCR located at any point in time in the original signal. This means that the SCR onset for the first column was set to equal time 0 and was then shift by one sample for each additional column until the length of the signal was reached. This allowed for an SCR to be identified at any time point within the signal. The initial τ_1 and τ_2 parameters were chosen based on existing literature as reasonable τ_1 and τ_2 values for SCR responses [1]. Additionally, we chose to initially use an SCR length of 30 seconds due to the fact that the average time for a SCR half recovery is between 2 and 10 seconds, and the average total length of a typical SCR is between 10 and 30 seconds [1]. These assumptions lead to our model of a 30 second SCR characterized by a steep onset and slow recovery back to zero. Our final dictionary was a square dictionary of n by n where n is the length of the EDA signal being analyzed. Figure 4 shows an example of the first 5 columns of the dictionary and the shape of the SCR created using $F(\theta)$. Each column in the dictionary represents a single SCR while the rows define the time, in seconds, in which the SCR is elicited.

$$\begin{bmatrix} F(\theta) & 0 & 0 & 0 & 0 \\ 0 & F(\theta) & 0 & 0 & 0 \\ 0 & 0 & F(\theta) & 0 & 0 \\ 0 & 0 & 0 & F(\theta) & 0 \\ 0 & 0 & 0 & 0 & F(\theta) \end{bmatrix}$$

$$F(\theta) = e^{\frac{-t}{20}} + e^{\frac{-t}{.75}}$$

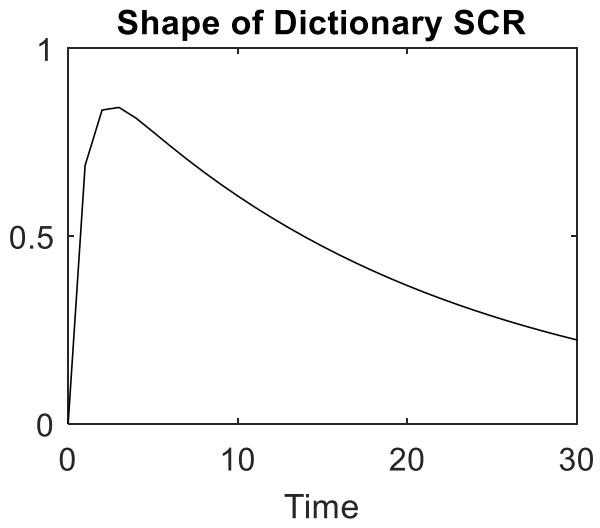


Figure 4 (left): Template used to create a knowledge driven dictionary with **(right)** SCR produced using $\boldsymbol{F(\theta)}$

After our initial experiment it was determined that a data-driven dictionary would provide better ability to detect SCRs. Labeled SCRs where therefore used in our second experiment to add additional columns, based off of new SCR shapes, to the dictionary using the same procedure described above. Each additional shape added to the dictionary increased the dictionary length by n. For example, if 2 additional shapes were added the dictionary dimensions increased from n by n to by n by 3n. See section 4.2.3.2 for details about the dictionary expansion methodology used.

3.3 BATCH ORTHOGONAL MATCHING PURSUIT

To achieve less computational complexity and a faster convergence time we selected an update OMP algorithm previously developed and shown to be efficient for the sparse-coding of large sets of signals [34]. The OMP based method implemented for our analysis to approximate the phasic estimate can be run using either a sparsity-constrained design or an error-constrained design. The

sparsity contained algorithm is controlled by (9) below, while the error based algorithm is based on (10) [34]

$$\underline{\gamma} = \underset{\underline{\gamma}}{\text{Argmin}} \left\| x - D\underline{\gamma} \right\|_2^2 \text{ Subject To } \left\| \underline{\gamma} \right\|_0 \leq K \tag{9}$$

$$\underline{\gamma} = \underset{\underline{\gamma}}{\text{Argmin}} \left\| \underline{\gamma} \right\|_0 \text{ Subject To } \left\| x - D\underline{\gamma} \right\|_2^2 \leq \epsilon \tag{10}$$

In each equation x is the original signal, D is the dictionary that we are coding over, and γ is the estimated coefficient vector. In (9) K is the sparsity constraint value, and ε is the error constraint value in (10) [34]. For our specific analysis x is the estimated phasic component, D is a dictionary created using the procedure in section 3.2, and γ will be used to determine the onset of located SCRs. Like any OMP algorithm, BOMP uses a greedy approach with the 2 main steps introduced earlier. The first step uses the dictionary to select an atom with the highest correlation to the current residual. The residual is initialized as the entire phasic response so initially the atom with the highest correlation to the entire phasic response is selected. The second step then takes the selected atom and subtracts it from the current residual to get an updated residual. For the first iteration the selected atoms response is subtracted from the phasic response to get the residual that will be used in the second iteration. Finally, since the methodology used is a variation on the traditional OMP algorithm, the selected atom is projected into an orthogonal space and removed from the dictionary to prevent the atom from being selected a second time. The orthogonalization step is a computationally expensive step and causes OMP algorithms to have a slow convergence, especially as the length of the signal and the dictionary increase. To further improve upon the

traditional OMP algorithm, the BOMP algorithm reduces the computational complexity by introducing a Cholesky factorization [34]. As mentioned OMP algorithms orthogonalize each selected atom which introduces a matrix inversion at each iteration. The orthogonalization step, with the matrix inversion, is shown in equation (11)

$$\begin{aligned} \underline{\gamma} &= (\mathrm{D}_I)^{+} x \\ &= (\mathrm{D}_I^T \mathrm{D}_I)^{-1} \mathrm{D}_I^T x \end{aligned} \tag{11}$$

From equation (11) it can be seen that at each iteration the $(\mathrm{D}_I^T \mathrm{D}_I)$ matrix remains non-singular due to orthogonalization; the $(\mathrm{D}_I^T \mathrm{D}_I)$ matrix is also a symmetric positive-definite (SPD) matrix which is updated each iteration by simply adding a single row and column to the end of the matrix [34]. To improve performance and decrease computational complexity, Cholesky factorization was used to only require the computation of the last row of the new matrix, replacing the need to invert the larger $(\mathrm{D}_I^T \mathrm{D}_I)$ matrix at each step [34]. This leads to less computational complexity and faster BOMP algorithm (summarized in Figure 5).

Input: Dictionary $\boldsymbol{D}$, signal x, target sparsity K
Initialize: Set $I := ()$, $\boldsymbol{L} := [1]$, $r := x$, $\gamma := 0$, $\alpha := \boldsymbol{D^T} x$, $n := 1$
while sparsity < K
    $\hat{k} := argmax\ |d_k^T r|$
    if $n > 1$
        $w :=$ Sovle for $w \{\boldsymbol{L}w = \boldsymbol{D_I^T} d_{\hat{k}}\}$
        $\boldsymbol{L} := \begin{bmatrix} \boldsymbol{L} & 0 \\ w^T & \sqrt{1 - w^T w} \end{bmatrix}$
    end
    $I := (I, \hat{k})$
    $\gamma_I :=$ solve for c $\{\boldsymbol{L}\boldsymbol{L^T} c = \alpha_I\}$
    $r = x - \boldsymbol{D_I} \gamma_I$
    $n = n + 1$
end
Output: Sparse representation γ such that $x \approx \boldsymbol{D}\gamma$

Figure 5: Pseudocode for the BOMP algorithm used in our analysis

In the above pseudo code I, r and $\boldsymbol{D\gamma}$ are the same variables presented in Figure 3. $\boldsymbol{L}$ is introduced to represent the Cholesky factorization, α stores the product of $\boldsymbol{D}^T r$ to reduce computational cost by completing pre-computation of large or commonly used products, and n is a counter used with the Cholesky factorization.

While Figure 5 shows the pseudocode specific to the sparsity constrained algorithm the algorithm presented in Figure 5 can be used to solve either the error constrained or sparsity constrained optimization problems introduced in (9) and (10) by simply modifying the stopping criteria used in line 3 [34]. For our analysis, we use the sparsity constrained stopping criteria due initial to the speed of the sparsity algorithm. In initial runs of the BOMP method, we found that the sparsity constrained method was significantly faster and produced fits to the original data that were comparable or better than the error constrained method. Additionally, using sparsity as the stopping criteria further enforces the sparsity of the returned γ which we found could further limit the number of false positives identified. Using sparsity as the stopping criteria makes K into a system parameter that needed to be optimized for each signal. As the sparsity controls the number of atoms that can be selected during analysis the overall performance of the system will be highly dependent on the sparsity value used. Low sparsity will likely lead to under fitting the signal and may cause many SCRs to be missed. On the opposite side high sparsity levels will likely lead to overfitting which could lead to fitting to noise as well as true SCRs. To determine the optimal sparsity, K, for each signal, the BOMP analysis was run over a range of K values, 10 to 600 and

then the best K value was chosen based on the calculated performance measures and plotted ROC curves.

3.4 PERFORMATION CALCULATIONS

To facilitate comparisons done throughout this manuscript performance measures will be calculated based on the ability to accurately identify SCRs. For each study done the performance measures will include the calculation of accuracy, sensitivity, specificity, and precision based on categorizing each identified response as either as true positive (TP), false positive (FP), true negative (TN), or false negative (FN). Before the results section for each study there will be a specific explanation of how the categorization for each value was defined and then to give specifics on the performance measures calculated within the definition of each study. This section will define generally how performance is measured and the formulas for each of the previously mentioned measures.

For each study the goal is to positivity identify SCRs in the signal while avoiding identifying pieces of the signal that do not correspond to SCRs, either because of noise or artifacts. To measure our ability to balance true identification with false identifications we will calculate statistical performance measures for a binary classification. Positivity identifying a true SCR will be labeled as a true positive (TP), identifying either an artifact or noise as an SCR will be counted as a false positive (FP), failing to identify a true SCR will be labeled as a false negative (FN), and correctly leaving sections of the signal that do not contain any SCRs blank will be counted as true

negative (TN). Using these definitions equations (12 – 15) below show the calculations that will be used for performance

Sensitivity or true positive rate:

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

Specificity or true negative rate:

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

Precision or positive predictive value

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

Accuracy

$$Accuracy = \frac{(TP + TN)}{TP + FP + FN + TN} \tag{15}$$

3.5 STATISTICAL EVALUATION

Along with the performance measures described above different methodologies throughout this manuscript will be evaluated using statistical testing. Statistical testing was introduced as a second method of comparison to give a better sense of if a set of data was truly different and if they were how much did the populations differ. The ranksum test is a nonparametric rank test which tests the equality of the medium for two populations. The null hypothesis assumes that the medium of the two populations is equal while the alternative hypothesis assumes that one population's medium is greater than the other [41]. The Wilcoxon test was chosen over other statistical tests since it does

not assume that the data follows any specific distribution. Instead, the magnitude of the differences between the populations is used to determine the statistical significance.

4.0 EXPERIMENTAL DESIGN

4.1 PROOF OF CONCEPT WITH AMBULATORY DATA[1]

Our first experiment employed the use of data collected using an ambulatory paradigm. This data however, did not include expert labels for SCRs or artifacts. Instead this study used labels produced by Ledalab, a gold standard software, as the basis of comparison. The following sections will first describe the goals of the study and then break down the data collection procedures, the data analysis methodologies, present the results with discussion, and finally discusses the limitations found and the future direction.

4.1.1 Study Goals

Common limitations in current analysis software become exaggerated and more prominent when long EDA signals, typical of ambulatory collected EDA data, are analyzed. As previously discussed, the two major limitation towards analyzing ambulatory data are:

[1] Based on M. Kelsey, A. Dallal, S. Eldeeb, M. Akcakaya, I. Kleckner, C. Gerard, K. S. Quigley and M. S. Goodwin, "Dictionary Learning and Sparse Recovery for Electrodermal Activity Analysis," in *SPIE Commercial Scientific Sensing and Imaging*, Baltimore, 2016.

1) the inability to distinguish between an artifact and a true SCR within the signal
2) the computational complexity associated with current methods that lead to long computation time required for signal analysis.

In the design of the first experiment we aim to address the second limitation by introducing a new methodology using sparse recovery algorithms that will decrease the computational complexity of EDA analysis over Ledalab, a current gold standard software package. Sparse recovery methodologies were chosen as a promising approach given the sparse nature of SCRs in EDA signals. Sparse recovery methods have also been shown to perform well when applied to the analysis of lab collected signals. Using a sparse recovery method, we aim to show a significant improvement in the analysis time required for analysis of large EDA datasets that.

4.1.2 Experimental Data

Ambulatory EDA data used in the first study came from five minimally-verbal and intellectually impaired individuals on the autism spectrum, 9-20 years old, receiving educational instruction at the Groden Center. The first goal of this data collection was to see if we could record a substantial amount of physiological and physical activity data including EDA, 3-axis accelerometery, and skin surface temperature levels, ideally for 60 consecutive school days with the Q sensor. We were also interested in whether we could collect meaningful and rich annotations and connect them with the Q sensor data during an automated data analysis paradigm. This kind of plentiful natural-environment data has never been collected before to our knowledge.

Beginning each school day, researchers synchronized sets of Q sensors, stopwatches, computers, and video cameras to Internet time. Q sensors were then consistently put on the right ankle of each student using an athletic band. All five students tolerated this placement and wearing

of the sensors quite well. One student wore the sensor on the left ankle since an injury required her to soak her right ankle daily. Students wore the Q sensor for the entirety of the school day (9:00am-3:00pm) and it was only taken off if it had the potential of getting wet.

For each day of recording, teachers were given a clipboard, data collection sheets, and a stopwatch and video camera time-synchronized with his or her student's Q sensor. Teachers used these tools to record the setting (e.g., classroom, workshop, cafeteria, gym), activity (e.g., educational, vocational, eating, physical activity), and clinically relevant behavior (e.g., inappropriate laughter, perseveration, rocking, tantrums, aggression) start and stop times. Behaviors and activities were operationally defined for each study participant. When clinically relevant behaviors were observed, teachers were asked to fill out a detailed behavior report indicating which behavior, events that preceded it, consequences of the behavior, actions taken, and perceived function. Videos of clinically relevant behavioral episodes were also taken periodically when it was safe to do so. At the end of each day, Q sensors were taken off of students and their physiological data, teacher's annotations, and videos were uploaded, filed, and stored on designated laptops.

We collected more than 1300 hours of autonomic data from five children on the autism spectrum. For these data we also obtained over 6000 annotations from teachers describing ordinary school-day activities as well as unusual challenging behaviors. To our knowledge, this is the largest time scale autonomic data collection ever conducted in a natural setting with individuals on the autism spectrum. Table 1 below summarizes the data collection broken down for each participant.

Table 1: Data collection broken down by participant

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Number of Days	64	62	61	70	55
Avg. time/day (hours)	4.55	4.26	4.49	4.23	3.64
Total recording time (hours)	290.96	264.05	273.99	296.06	200.46

4.1.3 Analysis Methods

From the close to 1325.5 hours of data collected over the 60-day study 30 datasets were randomly chosen totally close to 100 hours of EDA data. The chosen files included data from all 5 participants as well as data from multiple days for each participant. Additionally, all analysis was performed on all of the chosen 30 data sets. For our first study our proposed method for identifying the location of SCRs within an ambulatory signal uses the previously introduced batch orthogonal matching pursuit (BOMP) method [34].

During initial data collection, EDA signals were collected with a sampling frequency of 8 Hz, which was down sampled by a factor of 8 to a frequency of 1 Hz for analysis. Down sampling was completed after filtering was performed. Each signal was filtered using a low pass filter with cutoff frequency of .35 Hz to remove higher frequencies, which would be considered artifacts, from the signal. A cutoff frequency of .35 Hz was chosen as it roughly corresponds to a response quicker than 3 seconds. Since a typical SCR has a rise time between 1-3 seconds we chose to filter anything with a rise time faster than 3 seconds out as noise. Later studies will be used to evaluate different cutoff frequencies which match the 1-3 second SCR rise time knowledge.

After preprocessing was complete the signal was analyzed using Ledalab to identify the location of SCRs within the signal. These labels were considered to be the ground truth and were used as our basis for comparison. This initial run of Ledalab was also used to determine the analysis

time required to complete Ledalab's analysis. After using Ledalab to determine SCR locations, the down sampled signals were again analyzed and SCR locations found using our methodology.

For our methodology the tonic level was estimated and removed from the signal to leave only the phasic component, note that SCR identification was performed on the phasic component only. Tonic estimation was done in two ways and the results were analyzed separately to determine the most efficient estimation method using both analysis run time and performance measures to compare the methods. The first way the tonic level was estimated was to use the same tonic estimation methodology employed by Ledalab with one round of optimization. As optimization procedures can be computationally complex and time consuming the second method was to again use Ledalab's tonic estimation method, but this time without the additional optimization done by Ledalab. Removing the optimization done will likely reduce the run time significantly but could also have a negative impact on the ability to identify SCRs. After the phasic signal was estimated BOMP was used to identify SCR onsets within the phasic estimate.

BOMP analysis returns a coefficient vector, γ, which represents the weights of the selected atoms. For our purpose the γ signal reflects which atoms from the dictionary best represent the original signal and give us the location of the onsets of individual SCRs throughout the signals. The located SCRs were then compared to the number of SCRs and the locations of SCRs found by Ledalab to determine the performance of each method. Further details about performance calculations can be found in section 4.1.3.1.

To find the best BOMP estimate, BOMP was run using the sparsity constrained stopping criteria, K. The K value affected the goodness of fit for each data set, therefore, each data set was analyzed using sparsity levels varied between 10 and 600. The optimal sparsity value was then chosen based on the calculated performance measures. In addition to the performance measures,

run time for Ledalab and our method were compared to determine their respective computational complexity. Figure 6 shows a flow chart of the detailed analysis used for each dataset.

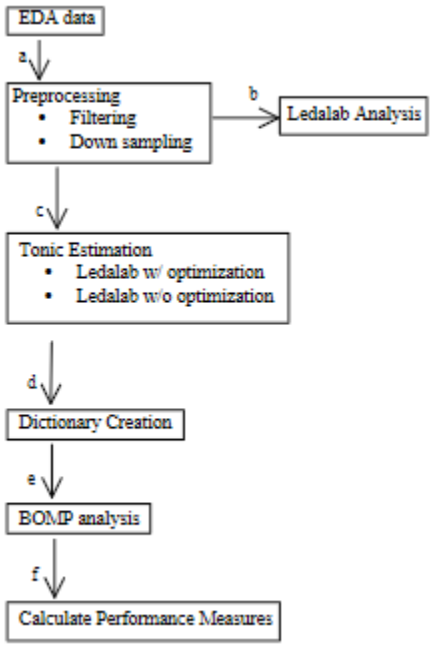


Figure 6: Flow chart to depict the general analysis methodology used for experiment 1

4.1.3.1 Comparisons and Performance

To determine the performance measures of our method, sensitivity, specificity, precision, and accuracy were calculated. Performance was calculated by identifying the number and onsets of each individual SCR found by our method and then comparing our identified SCRs to the locations of the SCRs found in Ledalab.

Using the onset location and assumption that a typical SCR is 30 seconds long, each SCR in the signal was classified as either a true positive or false positive, while the absence of an SCR was either classified as a true negative or false negative. A true positive was counted if the SCR

found in Ledalab was also found at the same location in our estimated signal with a shift tolerance of 15 samples. A false positive was counted if an SCR was identified by Ledalab but not found at the same location in our estimated signal using the same shift tolerance of 15 samples. To count false positives, the locations of SCRs found in our estimated signal was compared to the location of that SCR in Ledalab, again using a shift tolerance of 15 samples to confirm that no SCR was present in the Ledalab signal. Finally, true negatives were counted if the location of no SCRs in our estimated signal also didn't contain any SCRs in Ledalab for 80 percent of the estimated no SCR length.

4.1.4 Results and Discussion

4.1.4.1 Analysis with optimization

First, our BOMP method was run with the phasic estimation produced using Ledalab's methodology with one round of optimization. The BOMP fit to a phasic estimation is shown below in Figure 7. To show the fit in a clearer format, Figure 7b (left) shows the signal zoomed in between 6200 and 7200 seconds. Both figures below show the estimated phasic response in black with the fit generated by BOMP shown on top in red. While the BOMP estimate is not a perfect match, it generally shows peaks at the same places as those in the original signal suggesting that SCR in the signal have been captured.

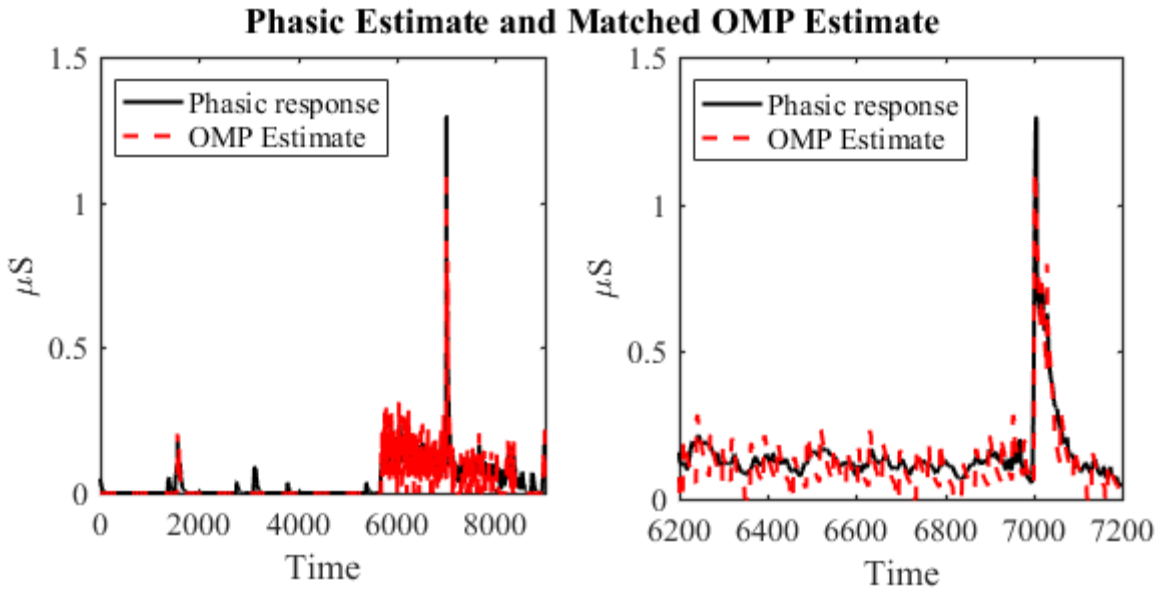


Figure 7: Phasic estimate created using Ledalab with the fit created from BOMP overlaid

Using a more quantitative approach to evaluate BOMP, the SCRs identified by our analysis were compared to those found by Ledalab and then the performance was calculated. An average accuracy of about 81%, over the 30 datasets used for analysis, was found. Table 2 below shows the calculated averages for sensitivity, specificity, accuracy and precision.

Table 2: Average performance measures of the 30 data sets analyzed.

	Average Values over 30 DS
Sensitivity	90.19
Specificity	55.29
Precision	80.52
Accuracy	80.98

Sparsity level, k, used for our analysis played a large role in the BOMP analysis of each data set and subsequently the calculated accuracy, precision, specificity, and sensitivity. For the average calculation, the best sparsity value, chosen based on the calculated performance, was used

for each individual data set. The best sparsity chosen for each dataset showed a large variation spanning the entire range tested, a sparsity level of 10 up to a sparsity level of 600. Table 3 lists the sensitivity, specificity, precision, and accuracy at the best sparsity level for each data set to show this variation as well as the individual performance achieved on each participant. Additionally, the table shows what the best sparsity level was for each data set.

Table 3: Performance measures and ideal sparsity found for each participant based on phasic estimates created using 1 round of optimization.

Data Set	**DS1**	**DS2**	**DS3**	**DS4**	**DS5**	**DS6**	**DS7**	**DS8**	**DS9**	**DS10**
Sparsity	25	25	25	100	25	75	150	600	50	300
Sensitivity	.9512	.8000	.6000	.8219	.7222	.9872	.8387	.9896	.9412	.9596
Specificity	.6667	.6364	.5357	.4674	.5714	.5625	.6111	.7143	.6129	.4400
Precision	.9512	.7143	.4800	.5505	.5200	.9167	.8478	.9502	.8000	.9169
Accuracy	.9149	.7234	.5625	.6242	.6304	.9149	.7752	.9473	.8171	.8898
Data Set	**DS11**	**DS12**	**DS13**	**DS14**	**DS15**	**DS16**	**DS17**	**DS18**	**DS19**	**DS20**
Sparsity	150	25	10	50	100	75	75	150	75	25
Sensitivity	.8788	1.000	1.000	.7297	.8031	.9053	.7925	.9359	.8816	.8696
Specificity	.5185	.5000	.4000	.4800	.6200	.6250	.4247	.6341	.6000	.5600
Precision	.9305	.9762	.4545	.5094	.8430	.9053	.5000	.8295	.77001	.6452
Accuracy	.8356	.9767	.6000	.5862	.7514	.8487	.5794	.8319	.7698	.7083
Data Set	**DS21**	**DS22**	**DS23**	**DS24**	**DS25**	**DS26**	**DS27**	**DS28**	**DS29**	**DS30**
Sparsity	50	10	100	75	200	400	200	600	600	25
Sensitivity	.8261	1.000	.9795	1.000	.9126	.9976	.9873	.9704	.9764	1.000
Specificity	.5500	.5714	.8421	0.000	.5472	.8475	.6471	.6024	.2467	NaN
Precision	.8636	.7500	.9795	.9791	.9158	.9789	.9750	.9598	.7440	1.000
Accuracy	.7640	.8125	.9636	.9792	.8555	.9791	.9646	.9362	.7514	1.000

For the best sparsity level, BOMP was able to successfully identify most of the SCRs, typically having a sensitivity higher than 80%. However, the general trend is that our analysis produced a low specificity which is an indicator of number of false positives and true negatives. The lower specificity is likely due to the way true negative values were calculated as opposed to the number of false positives. The average number of false positives in a signal was about 21

compared to identifying an average of 150 TP. However, compared to the sparsity assumption made, the average number of TN was low, average of 23 TN. Since a true negative was counted for a span of time when both Ledalab and the BOMP analysis did not show any elicited SCRs, the time span can vary widely. This means that the counted negative section can either be a very short period of the signal or span several hundred seconds. This variation means that a negative response is not consistent with the 30 second length assumed for a SCR response such that the result in a low number of overall negatives in each signal. Given that EDA is considered a sparse signal, the number of negatives that counted in our analysis do not accurately reflect the true sparsity of the original signal.

As previously mentioned, changing the sparsity level for a signal greatly affected overall performance. Going above the ideal sparsity for each data set led to overfitting the data, and while true positives increased, trending towards a sensitivity of one, the false positives also increased dramatically, causing specificity to drop. On the other hand, using a sparsity less than the optimal sparsity level typically led to under fitting, which caused the specificity to be close to one but yielded many false negatives that drove down sensitivity. For any given data set, it was therefore impossible to achieve perfect sensitivity and specificity. The best sparsity level for a given dataset was therefore determined by balancing between sensitivity and specificity. Figure 8 shows the average sensitivity, specificity, accuracy and precision versus the sparsity (ROC curves) over the 30 datasets. Due to the fact that optimal sparsity differs for each individual dataset, general values do not reflect the best values found, rather they serve to show general trends seen when varying sparsity.

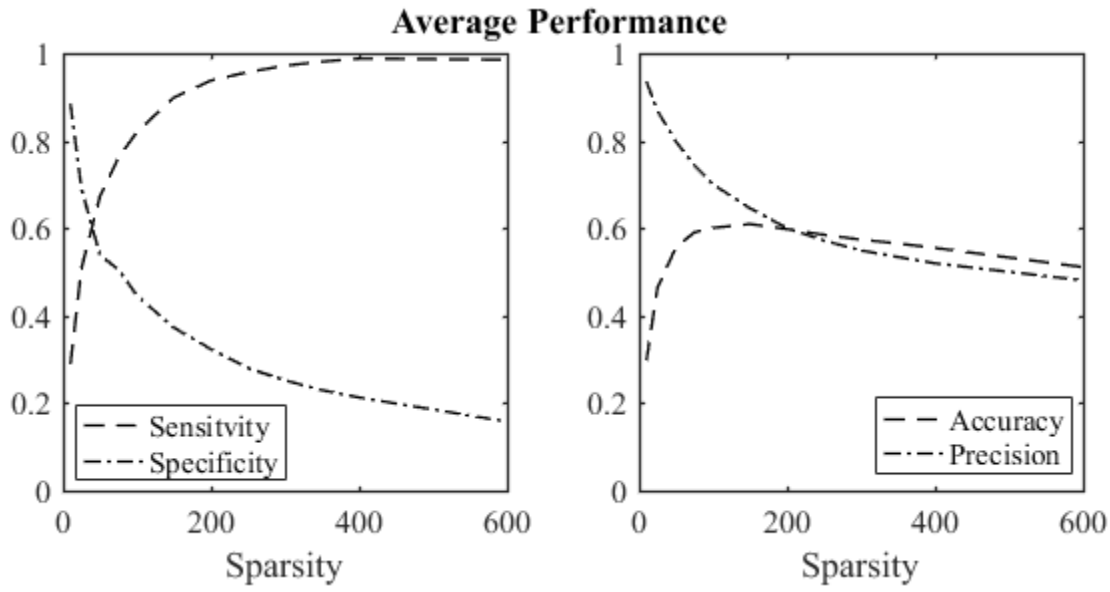


Figure 8: Average performance measures versus sparsity

Overlaying the sensitivity and specificity plots and the accuracy and precision plots gives an ideal sparsity of between 50 and 200. However, this also leads to lower sensitivity and specificity averages then those reported above. The ideal is instead to find the best sparsity level for each individual dataset to achieve the highest sensitivity and specificity possible.

Another important measure of the viability of our methodology was to keep the run time low. Using optimization to create the phasic estimate was a time consuming process, leading to generally longer run times, especially as the length of the signal increased. Additionally, increasing the sparsity threshold increased the time required to produce an estimated signal. Figure 9 below shows the time difference, in seconds, between Ledalab and the BOMP method versus the length of the data set using sparsity constraints of 25, 300, and 600.

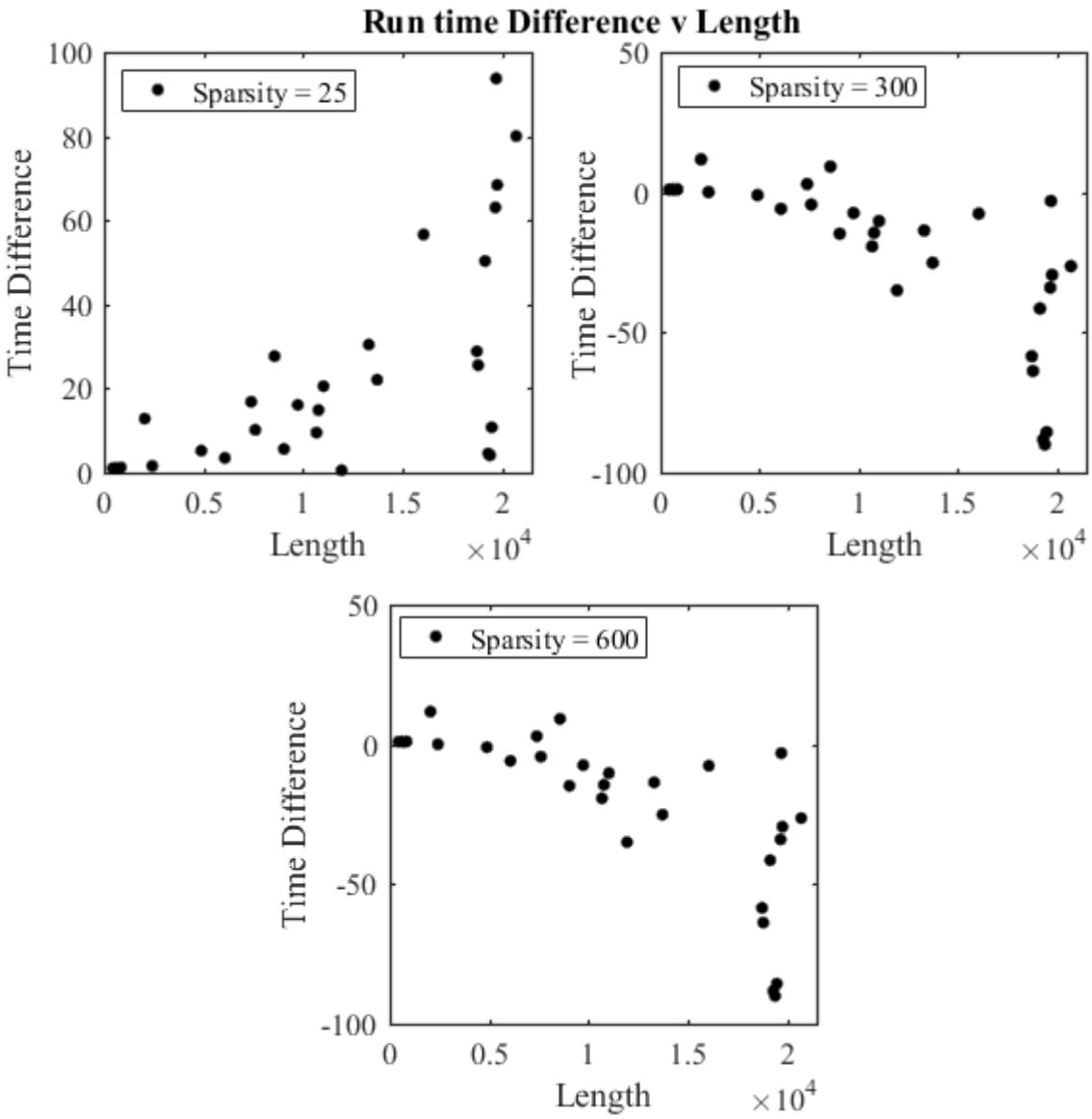


Figure 9 (top left): Run time for sparsity of 25, **(top right)** run times for sparsity of 300, and **(bottom)** run times for sparsity of 600

The time difference for all sparsity thresholds shows an exponential trend as signal length increases, but for higher sparsity values this trend favors Ledalab. However, and as previously mentioned, the average sensitivity and specificity for ideal sparsity is around 50 which would lead to faster times for the OMP method over Ledalab. Figure 10 shows this time plot.

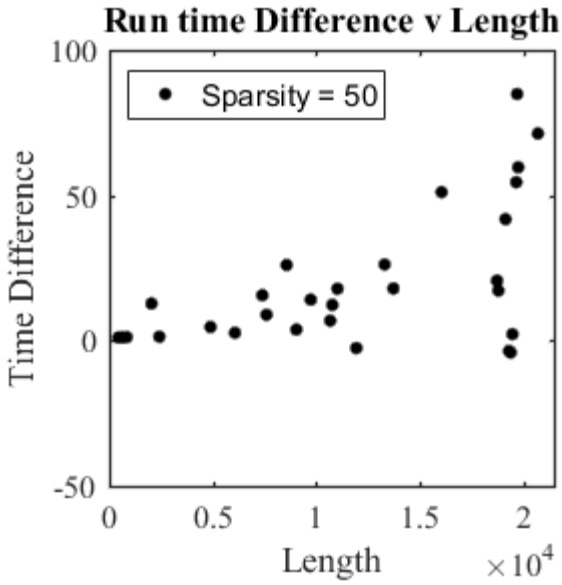


Figure 10: Run time for sparsity of 50, the average ideal sparsity.

4.1.4.2 Analysis without optimization

Using the same procedure as above, we ran the analysis is run a second time, using Ledalab's phasic estimation without optimization. The goal of using this modification was to improve the overall run time for higher sparsity thresholds and dataset length without sacrificing the performance of our analysis. Figure 11a shows the phasic estimation created using no optimization and the OMP estimation. Figure 11b shows the same signal zoomed in between 1200 and 2300 seconds to better illustrate the match between the original data and the BOMP estimate, and as before, yields similar trends in the signal.

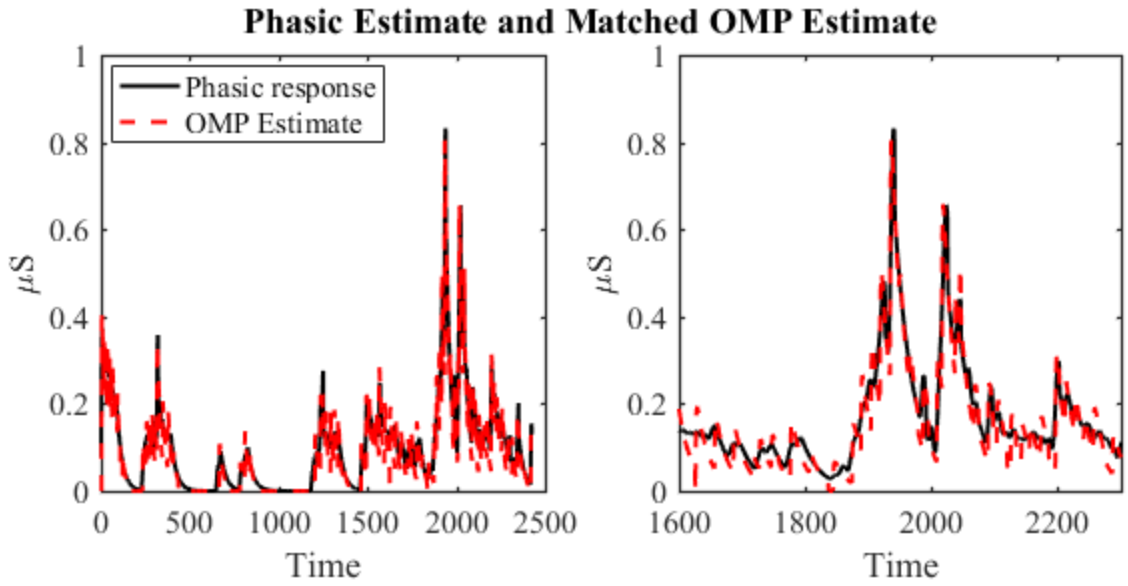


Figure 11: Phasic estimate created using Ledalab without optimization with the fit created from BOMP overlaid

However, the phasic estimate without optimization was not as robust and did not perform well for across the participants when the optimization step was excluded. For two of the 30 datasets used for analysis, the phasic estimate returned did not accurately reflect the true phasic response of the signal. If the phasic estimate was poor, BOMP similarly, returned poor performance in SCR identification. Accordingly, datasets 10 and 29 were excluded from further analysis. Figure 12 (left) shows the phasic estimate produced using no optimization and Figure 12 (right) shows what the phasic estimation looked like when optimization was used to illustrate the problem with the phasic estimate.

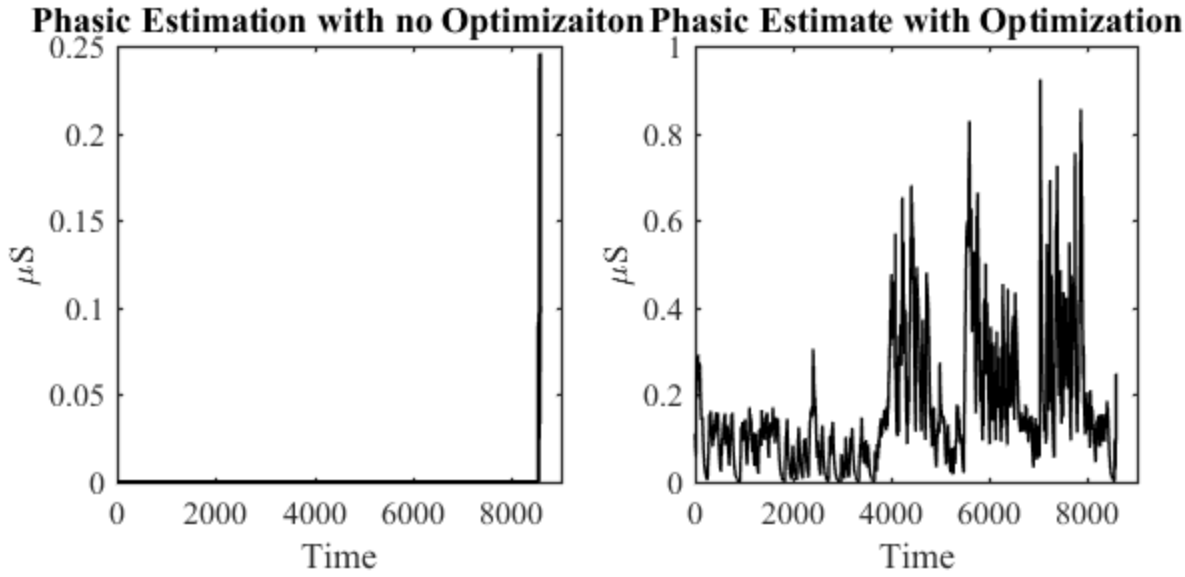


Figure 12: Limitation of creating phasic estimate without optimization

Using BOMP to identify SCRs using the phasic estimate created using Ledalab's estimation method with no optimization led to an average accuracy of 80% over the 28 datasets. This average accuracy is only slightly lower than the accuracy found when optimization was used. Table 4 below shows average values for the sensitivity, specificity, precision, and accuracy.

Table 4: Average performance measures for the 28 participants

	Average Values over 28 DS
Sensitivity	89.57
Specificity	52.82
Precision	80.12
Accuracy	80.13

Again, the sparsity value used played a large role in the performance found for each data set. Again no specific trend was found when considering the best sparsity used for each individual dataset; instead sparsity showed large variation ranging from a sparsity value of 10 up to a sparsity of 600 needed to achieve a balanced performance. Table 5 shows the calculated values for each individual data set as well as the best sparsity chosen for each. Once again, specificity is not as high as we would like it to be due to how true negative values were calculated.

Table 5: Performance measures and ideal sparsity found for each participant based on phasic response estimated with no optimization.

Data Set	**DS1**	**DS2**	**DS3**	**DS4**	**DS5**	**DS6**	**DS7**	**DS8**	**DS9**	**DS11**
Sparsity	25	25	25	100	25	75	150	600	500	150
Sensitivity	.9268	.8400	.7500	.7808	.7778	.9615	.8172	.9689	.9412	.8889
Specificity	.6667	.6111	.6296	.4302	.5714	.4444	.6486	.5076	.6333	.4286
Precision	.9500	.7500	.6000	.5377	.5385	.8824	.8539	.8960	.8136	.9167
Accuracy	.8936	.7442	.6809	.5912	.6522	.8646	.7692	.8831	.8272	.8319
Data Set	**DS12**	**DS13**	**DS14**	**DS15**	**DS16**	**DS17**	**DS18**	**DS19**	**DS20**	**DS21**
Sparsity	25	10	50	100	75	75	150	75	25	50
Sensitivity	1.000	1.000	.7297	.8031	.9158	.7736	.9359	.8289	.8696	.7826
Specificity	.333	.4545	.5536	.5957	.5909	.3824	.6081	.5417	.5417	.5238
Precision	.9535	.4545	.5192	.8430	.9063	.4940	.8343	.7412	.6452	.8438
Accuracy	.9545	.6250	.6237	.7471	.8547	.5537	.8304	.7177	.7021	.7222
Data Set	**DS22**	**DS23**	**DS24**	**DS25**	**DS26**	**DS27**	**DS28**	**DS30**		
Sparsity	10	100	75	200	400	200	600	25		
Sensitivity	1.000	.9452	1.000	.9056	.9833	.9831	.9692	1.000		
Specificity	.5714	.7222	0.000	.5472	.5405	.5882	.5955	Nan		
Precision	.7500	.9650	.9792	.9152	.9238	.9708	.9562	1.000		
Accuracy	.8125	.9207	.9792	.8496	.9168	.9567	.9322	1.000		

Figure 13 shows average sensitivity and specificity versus sparsity and Figure 13 shows average accuracy and precision vs sparsity over the 28 data sets. The general values do not reflect

the best values found, but instead serve to show the general trend seen when varying the sparsity given the fact that optimal sparsity differs for each individual data set.

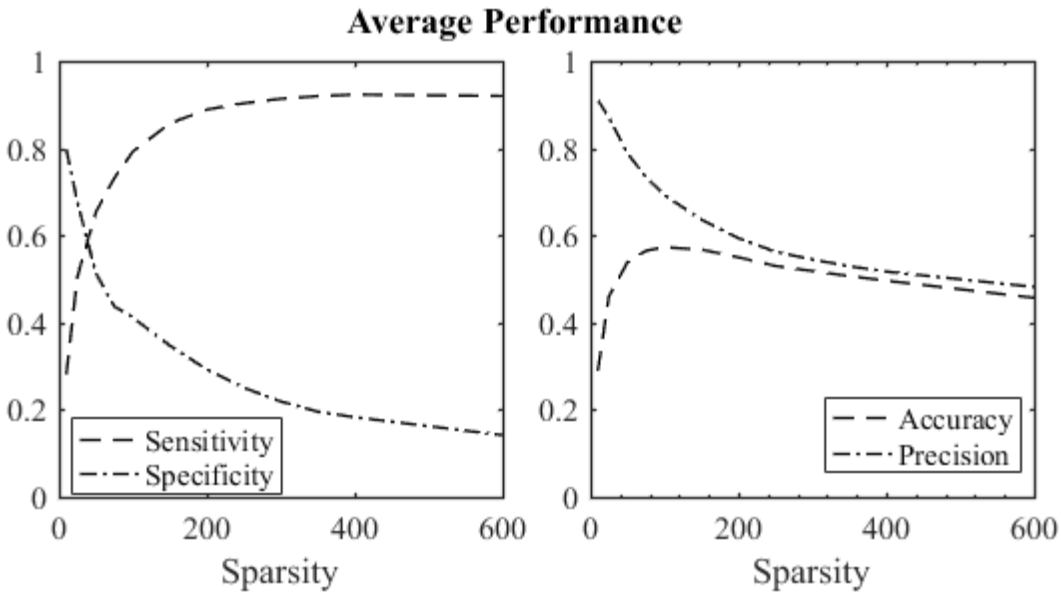


Figure 13: Average performance measures versus sparsity

Overlaying the sensitivity and specificity plots and the accuracy and precision plots yielded an ideal sparsity between 50 and 100. However, this also led to lower sensitivity and specificity averages then reported above. The ideal is instead to find the best sparsity for each individual data set to achieve the best performance possible.

Not using optimization did not have a significant impact on the overall performance achieved in our data sets. While it was slightly lower for most of our tested data sets, it only suffered a small amount and the gains achieved in run time make this decrease acceptable. Using a one-sided Wilcoxon ranksum test, the statistical significance between estimating the phasic response with and without optimization was determined. To testing the difference between the

tonic estimation methods each of the calculated performance measure were evaluated. It was determined that there was no statistical difference between the two methods except in regards to the specificity. Table 6 shows the calculated p-values found.

Table 6: Wilcoxon ranksum test evaluating the statistical difference between the phasic estimates created with and without optimization

Performance Measure	Test decision	p-value
Accuracy	0	.7146
Sensitivity	0	.6261
Specificity	1	.0304
Precision	0	.7675

As previously mentioned, using the phasic estimation created without optimization achieved better run times. For all of the different sparsity values tested, we found that the time was exponentially better, with increasing length, when compared to the run time achieved by Ledalab. Figure 14a-c shows the time difference, in seconds, between Ledalab and the BOMP method versus the length of the data set using sparsity constraints of 25, 300, and 600.

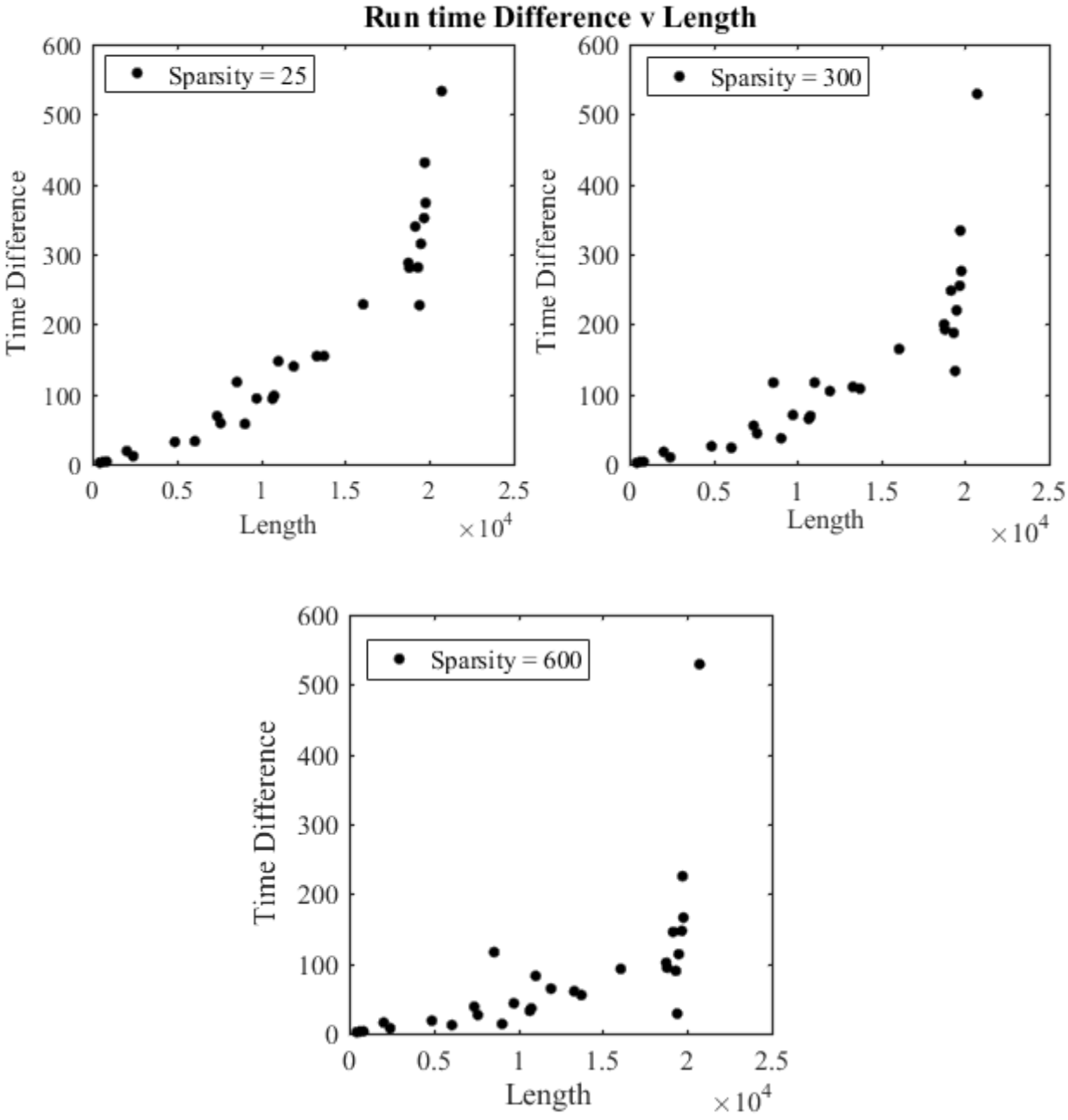


Figure 14 (top left): Run time for sparsity of 25, **(top right)** run times for sparsity of 300, and **(bottom)** run times for sparsity of 600

4.1.5 Limitations and Future direction

This initial study gives us an incipient method that we could expand on to further improve the efficient analysis of large EDA data sets typical of ambulatory recordings. However, it also highlighted several drawbacks that remain to be addressed. One of the issues that was determined through this study was that a more robust tonic estimation methodology is needed. The tonic estimation method employed by Ledalab works well for some signals but as was seen in Figure 11 when no optimization is used some signals produce a phasic estimate that does not contain any information about the original signal. Additionally, to get a good tonic estimate using Ledalab's tonic estimation will sometime require multiple optimizations which requires trial and error. Using trial and error to determine the optimal number of optimizations for a data set is time consuming and requires manual analysis of each individual data set. Finally, as there is no standardized methodology for estimating the phasic and tonic components of EDA signals, as many groups perform analysis without removing the tonic level [2], further analysis is required to fully evaluate the performance of Ledalab's tonic estimation. To be better able to evaluate Ledalab's tonic estimation in our next study we will use labeled data to determine how well Ledalab is able to identify SCRs throughout a signal once the tonic level is removed and how well the tonic estimation matches the original signal. Additionally, in our next study we will introduce a new tonic estimation methodology that is less computationally complex and will be robust across data sets and participants. Our method and Ledalab's method will be compared to evaluate the performance and validity of both methodologies.

The second issue to be addressed in our next stud is to transform the dictionary we are using to be a data-driven dictionary instead of knowledge-driven dictionary. In the current study

we used a single τ_1 and τ_2 value to fit to the SCRs in each data set. The chosen τ_1 and τ_2 values were chosen as reasonable values that could be used to represent an SCR [1], [21] however, using a single parameter pair does not take into account variations in SCR shapes seen between participants and within a participant. To be better able to identify SCRs we will use expert labeled SCRs with the Bateman equation to determine additional τ_1 and τ_2 parameter pairs that represent common SCR shapes within the given data. These common shapes can then be used to extend our dictionary as new dictionary atoms. Determining common SCR shapes across multiple data sets as the basis to expand our dictionary would also allow our analysis to be more robust across participants and would generalize our analysis to make it easy to extend to new data sets.

The third remaining issues that needs to be addressed is the ability to accurately distinguish between SCRs and artifacts in EDA signals. While filtering can be used to remove low magnitude artifacts, it may not be as affective at removing higher magnitude artifacts [12]. To be able to include atoms specific to artifacts into future dictionary builds the first step is to determine if artifacts can be modeled using the Bateman equation. Once we have determined models for artifacts, we need to determine if the returned parameters needed to model an artifact are separable from the parameters used to model SCRs. Finally, the artifact parameters can be able to be added as new atoms to the dictionary and performance evaluated for the ability to detect SCRs and artifacts

4.2 METHODOLOGY EXPANSION WITH LABELED DATA[2]

Our second experiment employed the use of data collected in a laboratory setting. After collection the data was labeled by experts in the field. The expert labels will be used as the basis of comparison to determine the detection ability of both Ledalab and our methodology. This study will also evaluate the performance of different tonic estimation methodologies to determine the optimal one that can be used across data sets and participants. The following sections will describe the goals of the study and then introduce the data collection procedures and the data analysis performed.

4.2.1 Study Goals

Expanding on the methodologies introduced in our first experiment [39], our goals for this study are 3 fold:

1) We aim to show that using an expanded data driven dictionary improves the performance of SCR identification over the use of a knowledge driven dictionary.
2) In an attempt to generalize the dictionary while limiting the dictionary size, therefore keeping runtimes low, we also investigate the removal of the tonic level using different techniques. We then compare our estimation to the estimation performed by Ledalab.
3) We provide a holistic view of both methodologies by reporting on the performance measures of our method and Ledalab's, which, to our knowledge, has yet to be done.

[2] Based on manuscript currently being prepared for submission.

4.2.2 Experimental Data

Data were collected at Northeastern University (Boston, MA) from 2013-2014 and included 73 participants (35 females) recruited from Northeastern University and the Boston area for being regularly physically active. Seven participants did not provide EDA data for the following reasons: three participants were lost due to a computer error, two participants were lost because they stopped the study early due to lack of interest, and two participants were lost because the researcher ended the study early due to lack of participant compliance. Out of the remaining 66 participants, the EDA data from eleven participants were not annotated for the following reasons: data from six participants did not also have a video recording due to a computer error, and data from five participants were due to a human-based computer error. The final dataset consisted of 55 participants (24 females) with age range 18 - 38 years ($M \pm SD = 24.3 \pm 5.5$ years).

Participants were greeted and consented in accordance with Northeastern University's institutional review board. Participants completed a health questionnaire asking about the intake of caffeine, alcohol, and medications, whether they were suffering from any illnesses, and the amount of time they slept the prior night. Participants were asked to abstain from caffeine, alcohol, and recreational drugs for the 12 hours leading up to the study. Participants who were too ill to perform the study tasks (e.g., sitting still and not coughing during physiological recording) were asked to come back to the lab when are healthy. Participants were measured for height and weight. Participants were fitted with pre-gelled ConMed (Westborough, MA) Cleartrace Ag/AgCl sensors to obtain a modified lead II ECG, a respiration belt, impedance cardiography sensors, and electrodermal activity sensors on the palm of the right hand. Physiological channels were sampled at 1000 Hz using BioLab v. 3.0.8-3.0.13 (Mindware Technologies LTD; Gahanna, OH). After being connected to physiological recording equipment, participants sat quietly for 2-10 minutes

during which they also completed a demographics questionnaire and the PANAS-X questionnaire [42]. Next, participants completed a five-minute seated baseline during which they were asked to sit still. They then completed a heartbeat detection task (data reported previously in [43]).

Next, participants completed a task viewing and providing ratings in response to each of 103 full-color photos were from the International Affective Picture System (IAPS), which are used to induce various affective experiences [44]. Per the IAPS instructions [44], participants were informed to remain still during the task, to immerse themselves in the experience of each photo, and to rate how they feel in response to each photo. The particular photos were selected based on normative ratings of pleasantness/unpleasantness (valence) and arousal experienced when viewing them (all IAPS photo numbers are shown in Table 7). Photos were separated into an initial "anchor" block of three photos, and then ten additional blocks of ten photos each: two blocks of unpleasant-high arousal, two blocks of pleasant-high arousal, two blocks of unpleasant-low arousal, two blocks of pleasant-low arousal, and two blocks of pleasant or neutral valence-low arousal. Participants viewed all instructions and photos sequentially on a high definition television screen two meters away while seated. The blocks were presented in a counter-balanced order across participants, always starting with the anchor block, to familiarize participants with the task [44], and then starting with either the unpleasant high arousal block or the pleasant high arousal block, and alternating between unpleasant, neutral, and pleasant blocks. The order of the photos within each block was randomized within participants. This task was implemented using BioLab v. 3.0.8-3.0.13 and an in-house MATLAB program (Mathworks, Natick, MA) that utilized PsychoPhysics Toolbox extensions [45], [46], [47].

For each of the 103 trials, participants first viewed a screen indicating they are free to move if desired (e.g., stretch, adjust posture), then, upon pressing the mouse button, they viewed a "get

ready" screen for 3-8 sec (duration jittered) indicating that the picture was about to be shown, then they viewed the picture for six seconds, then they rated their response to the picture in terms of (1) valence, using a continuous scale with nine anchor images for valence from the self-assessment manikin scale (SAM; [48]), (2) arousal, using a continuous scale with nine anchor images for arousal from the SAM, and (3) confidence in their responses, using a continuous scale anchored from "least confident" to "most confident" with "intermediate" in the middle. After the completion of each block of images, participants also made the same valence, arousal, and confidence ratings in response to the entire block. The participant rating data are not reported here. Finally, participants completed additional tasks not related to this study. Participants were compensated $5 per half hour.

A group of four expert EDA raters (all trained by an author, I.K.) labeled the EDA data in response to each photo by simultaneously viewing the EDA data, respiration belt data, and a video of the participant (the EDA sensors were visible in the video) from photo onset to 4 seconds after photo offset (10 seconds total). Each EDA data segment was labeled as one of the following possibilities: (1) only one SCR, which is a biologically induced EDA response of at least 0.01 μS amplitude, (2) only two SCRs, (3) only three SCRs, (4) only four SCRs, (5) an artifact, which is a change in EDA not due to an SCR (e.g., the participant or other objects touching the EDA sensors or tension on the EDA leads), (6) participant movement (e.g., scratching their nose) without an EDA response, (7) at least one SCR likely caused by participant movement, (8) an EDA artifact likely caused by participant movement, (9) no SCR, participant movement, or artifacts, or (10) no data available. Using these annotations, the EDA data were categorized into three groups: (1) segments that elicited a clear SCR (cases 1-4 above), which were analyzed in experiment 2 (section 4.2.4), (2) segments that involved a clear EDA artifact (cases 5, 6, and 8), which were analyzed in

experiment 3 (section 4.3.3), and (3) all other segments, which were not analyzed further (cases 7, 9, and 10).

Table 7: Pictures presented from the International Affective Picture System (IAPS)

Block	IAPS #	Block	IAPS #	Block	IAPS #
Anchor	3000	Anchor	8499	Anchor	7010
Positive High Set 1	8400	Neutral Low Set 1	7020	Negative High Set 1	9183
Positive High Set 1	8178	Neutral Low Set 1	7175	Negative High Set 1	3110
Positive High Set 1	8034	Neutral Low Set 1	7009	Negative High Set 1	6350
Positive High Set 1	8341	Neutral Low Set 1	7030	Negative High Set 1	6520
Positive High Set 1	8030	Neutral Low Set 1	7150	Negative High Set 1	3080
Positive High Set 1	8163	Neutral Low Set 1	7016	Negative High Set 1	9413
Positive High Set 1	8501	Neutral Low Set 1	7034	Negative High Set 1	9902
Positive High Set 1	8179	Neutral Low Set 1	7012	Negative High Set 1	3500
Positive High Set 1	8370	Neutral Low Set 1	7050	Negative High Set 1	3550.1
Positive High Set 1	8190	Neutral Low Set 1	7185	Negative High Set 1	9921
Positive Low Set 1	2352	Negative Low Set 1	2141		
Positive Low Set 1	7508	Negative Low Set 1	3300		
Positive Low Set 1	4624	Negative Low Set 1	9140		
Positive Low Set 1	8497	Negative Low Set 1	9332		
Positive Low Set 1	1463	Negative Low Set 1	9342		
Positive Low Set 1	4628	Negative Low Set 1	2900		
Positive Low Set 1	1720	Negative Low Set 1	9832		
Positive Low Set 1	2274	Negative Low Set 1	2301		
Positive Low Set 1	2310	Negative Low Set 1	9295		
Positive Low Set 1	5215	Negative Low Set 1	9220		
Negative High Set 2	3100	Neutral Low Set 2	7025	Positive High Set 2	7650
Negative High Set 2	9414	Neutral Low Set 2	7011	Positive High Set 2	8300
Negative High Set 2	3180	Neutral Low Set 2	2440	Positive High Set 2	8210
Negative High Set 2	9635.1	Neutral Low Set 2	7018	Positive High Set 2	5470
Negative High Set 2	3170	Neutral Low Set 2	7055	Positive High Set 2	8251
Negative High Set 2	3191	Neutral Low Set 2	2397	Positive High Set 2	8502
Negative High Set 2	3301	Neutral Low Set 2	2512	Positive High Set 2	8470
Negative High Set 2	2703	Neutral Low Set 2	2396	Positive High Set 2	8496
Negative High Set 2	9903	Neutral Low Set 2	7031	Positive High Set 2	5833
Negative High Set 2	3266	Neutral Low Set 2	7041	Positive High Set 2	8193
Negative Low Set 2	9426	Positive Very Low Set 2	1659		
Negative Low Set 2	3216	Positive Very Low Set 2	2510		
Negative Low Set 2	2457	Positive Very Low Set 2	5711		
Negative Low Set 2	3181	Positive Very Low Set 2	1604		
Negative Low Set 2	9331	Positive Very Low Set 2	5200		
Negative Low Set 2	9470	Positive Very Low Set 2	7530		
Negative Low Set 2	9265	Positive Very Low Set 2	1740		
Negative Low Set 2	9610	Positive Very Low Set 2	2314		
Negative Low Set 2	7359	Positive Very Low Set 2	1601		
Negative Low Set 2	3160	Positive Very Low Set 2	2530		

4.2.3 Analysis Methods

This section provided a brief overview of the proposed methodology with further detail in sections the following sections (see Figure 15). For most of the analysis done a test set of 10 randomly selected participants where used to complete analysis. The base methodology used in our second study follows the same general procedure introduced in 1.0 and then used in experiment 1 (section 4.1.3). Signals were initially filtered and downsampled before the tonic and phasic components were estimated. In this study along with using the tonic estimation done by Ledalab, a novel tonic estimation methodology (introduced in 3.1.1) is also used. The various tonic estimation methodologies are compared to determine the performance of each. Similar to experiment 1, SCRs were initially identified from the phasic components using the batch OMP (BOMP) in conjunction with a knowledge-driven dictionary. To update the dictionary to a data-driven one, performance measures were calculated, based on the knowledge-driven dictionary, and misses in each signal were analyzed to find common trends. The common trends were then used to expand the knowledge-driven dictionary to a data-driven one. Using the expanded dictionary, the SCRs in the phasic components were again identified with the BOMP methodology and new performance measures calculated. As the last step, and to verify that our methodology did not over fit the data, validation was done using all 55 participants with our proposed method and data-driven dictionary. The following sections will provide details on the updates made to our initial methodology.

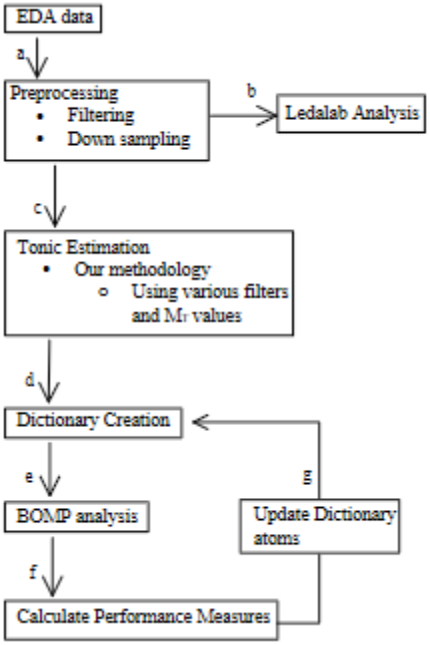


Figure 15: Flow chart to depict the general analysis methodology used for experiment 2

4.2.3.1 Preprocessing and Tonic Estimation

To determine ideal filtering and downsampling, signals were analyzed using 3 different filters and a range of sampling frequencies. All filters used were low pass filters with cutoff frequencies of .35 Hz, .5 Hz and 1 Hz respectively. These cutoff frequencies were chosen based on the average rise time for an SCR. The typical rise time is between 1 – 3 seconds [1], which corresponds to 1 - .35 Hz signals. The sampling frequencies, F_s, were ranged between 1 – 3 Hz to match each filter and avoid aliasing; and were obtained by downsampling the raw data from the original F_s of 1000 Hz. The lowest sampling frequency possible was used for each filter, as higher sampling frequencies did not show any benefits towards SCR detection and significantly increased the run time required to complete BOMP analysis.

A comparison of different tonic estimation methodologies was then done to determine the ability of each to accurately capture the tonic and phasic components of a signal and to evaluate the time required to complete the estimation. A comparison was done between our methodology,

introduced in section 3.1.1, and Ledalab's methodology. To estimate the tonic and phasic level using Ledalab's methodology no optimization was done. It was shown in our first experiment that using no optimization did not have a large impact on the general estimation but was significantly faster so this method was the focus for our comparison.

To fully test our novel tonic estimation methodology (see section 4.2.4.1) and to determine the optimal M_T, a range of thresholds was tested and the corresponding performance evaluated. Using the knowledge that the average length of an SCR is between 3 – 10 seconds [10] the M_T range investigated was 1 – 10 seconds using a step size of 1 second.

Note that in order to identify the SCRs in a signal, our methodology fits the atoms in the generated dictionary to the phasic component only. Therefore, the further references to the estimated signals refer to the fit of the BOMP to the phasic component instead of the full EDA signal.

4.2.3.2 Dictionary Expansion

In this study, two BOMP analyses were used sequentially to first generate and expand the dictionary and then to calculated final performance measures. To avoid overfitting of the BOMP analysis, the dictionary was expanded based on a test set of signals, and final performance measures were calculated on a validation set. Initial BOMP analysis of each test signal was completed using a base dictionary, as introduced in 3.2. Each individual column (i.e. atom) in the base dictionary represents a single SCR whose shape is defined by the Bateman equation, $\tau_1 = .75$ and $\tau_2 = 20$, for a 30 second time window. The format of the base dictionary is show in Figure 4

After each test signal was initially analyzed with the BOMP methodology and base dictionary, missed SCRs were identified, and new τ_1 and τ_2 values were computed to fit to the

misses using the Bateman equation. This process returned the optimal τ_1 and τ_2 values for each missed SCR. Using the optimal parameters for all of the misses, histograms were plotted and used to identify the common τ_1 and τ_2 values across the misses. The identified common τ_1 and τ_2 values were then used to create new SCR shapes that were added as additional columns to the dictionary. Each new SCR shape was added to the dictionary using the same methodology that was used to create the base dictionary to produce a dictionary using the format shown in Figure 16. Each new τ_1 and τ_2 parameter pair added to the dictionary increased the dictionary length by the length of the signal. If the original dictionary had a dimensionality of n by n and 3 additional atoms were added the dictionary would increase in size to be an n by 3n dictionary.

$$\begin{matrix} F(\theta) & 0 & 0 & 0 & 0 & G(\theta) & 0 & 0 & 0 & 0 \\ 0 & F(\theta) & 0 & 0 & 0 & 0 & G(\theta) & 0 & 0 & 0 \\ 0 & 0 & F(\theta) & 0 & 0 & 0 & 0 & G(\theta) & 0 & 0 \\ 0 & 0 & 0 & F(\theta) & 0 & 0 & 0 & 0 & G(\theta) & 0 \\ 0 & 0 & 0 & 0 & F(\theta) & 0 & 0 & 0 & 0 & G(\theta) \end{matrix}$$

$$F(\theta) = e^{\frac{-t}{20}} + e^{\frac{-t}{.75}} \; and \; G(\theta) = e^{\frac{-t}{\tau_1}} + e^{\frac{-t}{\tau_2}}$$

Figure 16: Template used to create knowledge driven dictionary.

4.2.3.3 BOMP Thresholding and Performance

Recall from (9), the output of the BOMP analysis is a matrix (γ) which represents which atoms from the dictionary were chosen (i.e. the weights for each atom). The dictionary is comprised of individual SCRs so each selected atom corresponds to a single SCR identified in the phasic component. This means that each non-zero value in γ represents the onset of an SCR. However, to get an accurate count of SCR onsets, post processing on γ was done to remove values that didn't align with our knowledge of SCR shape or previously made assumptions. For example, since SCRs are presented as concave inflections, any negative γ values were removed as they correspond to

convex inflections in the estimate. Additionally, if atoms within 1 second of each other were selected, these atoms added together into a single SCR as opposed to representing multiple onsets; filtering was therefore done to remove these redundant onsets.

To facilitate the comparison between our methodology and Ledalab, the average performance measures, including accuracy, sensitivity, specificity, and true SCR identification percentage (TIP), were reported over a set of labeled EDA data (see section 4.2.2 for details on the EDA data labeling). Each measure was then statistically assessed using a one-sided Wilcoxon ranksum test [41] to determine the statistical significance of the differences seen between BOMP and Ledalab's returned performance.

4.2.4 Results and Discussion

4.2.4.1 Tonic and Phasic Estimation

To determine a robust way to estimate the tonic and phasic level, we first compared the performance of our tonic estimation procedure with Ledalab's tonic estimation. We found that our methodology was more robust across different signals and led to a better ability to detect SCRs. To further improve upon our procedure, we then tested different filters and M_T values with our methodology to find the optimal combination based on calculated performance measures.

We started by comparing our methodology to Ledalab's tonic estimation using a .35 Hz filter and M_T equal to 10 seconds. Figure 17a shows the original data and Figure 17b shows the resulting Ledalab phasic estimate for one of the participants. The phasic estimate produced a signal that is non-negative but may overestimate the amplitude of some parts of the signal to maintain the non-negativity. In this phasic estimation, Ledalab shows most of the signal centered around .5 μS instead of the signal falling to zero between SCRs. This is caused by the non-

negative deconvolution and serves to increase the overall amplitude of the signal and therefore the amplitude of each individual SCR. Recall that in Figure 4 the dictionary was built on the assumption that a SCRs onset starts at 0 μS. The increase in amplitude that Ledalab gives appears to skew the phasic estimate and possibly allow phasic fluctuations that should have fallen below the threshold amplitude to be counted as true SCRs. Additionally, the Ledalab estimate seems to miss several compound SCRs around 400 seconds, and instead model them as a single peak. In comparison, Figure 17b shows the phasic estimate produced by using our phasic estimation methodology. This method keeps the phasic estimate centered around zero, however, it also produces negative results. The possible compound SCR around 400 is also captured more fully.

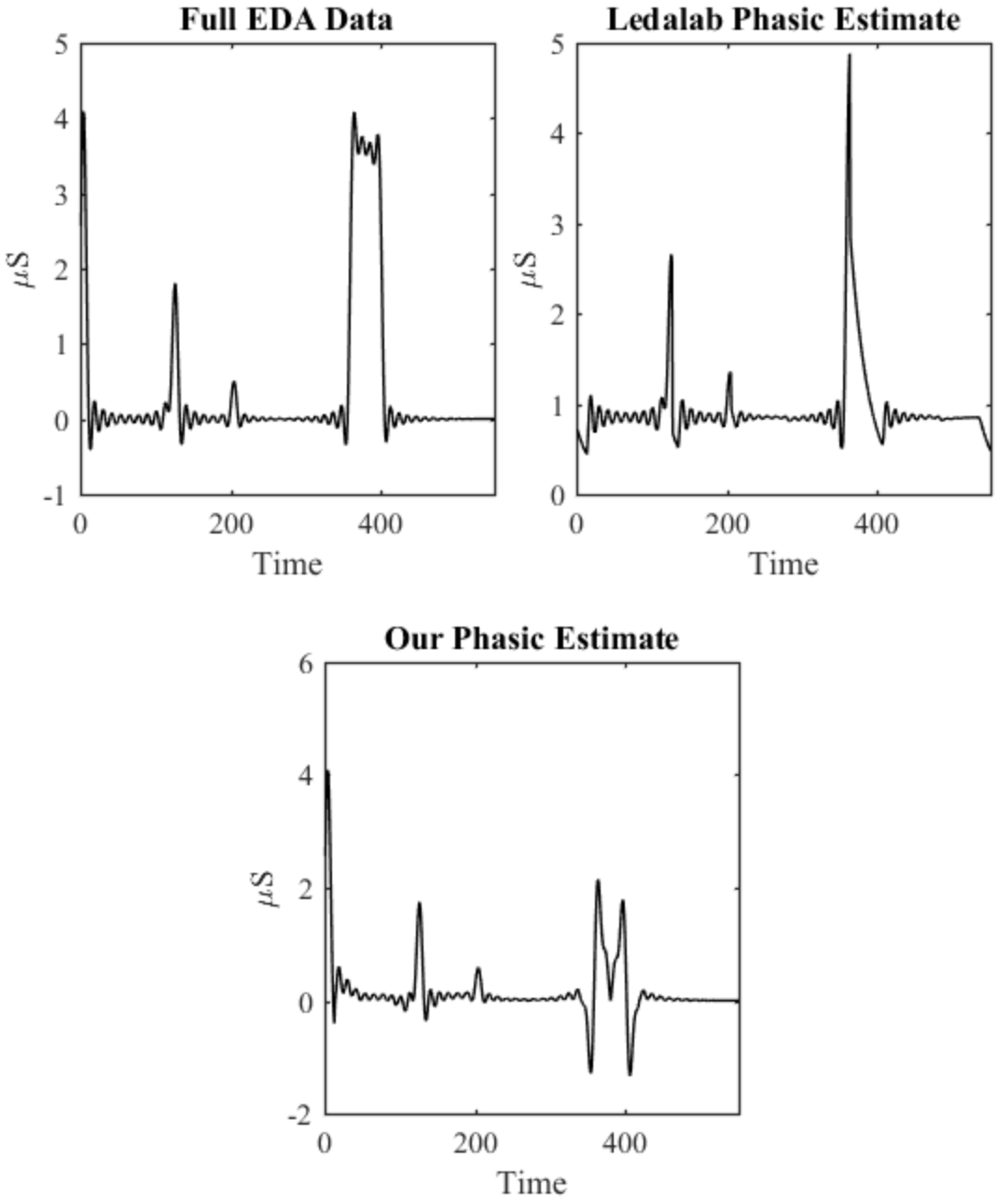


Figure 17 (top left): Original EDA data after filtering with a low pass filter. **(top right):** Phasic estimate produced using Ledalab's methodology compared to **(bottom)** the phasic estimate produced using our methodology.

Next Ledalab's tonic estimation is evaluated in more depth. Figure 18 shows the original signal plus Ledalab's tonic estimation, and the tonic plus phasic reconstruction for Ledalab. As can be seen from the reconstructed signal, much of the reconstructed signal is below the original

signal, and therefore does not produce a good match; this suggests that information from the original signal has been lost during the estimation process. While Ledalab's estimation works well for some signals, it is not robust to all signals as it is dependent on the number of optimizations used during the tonic estimation process. To get a good match for some signals, 2 or more optimizations are required which is time consuming and requires trial and error to determine optimal optimization number for each individual signal. Finally, the performance was low when using Ledalab's phasic estimation to identify SCRs with the BOMP methodology. Only 27.83% of the possible SCRs in the signals were detected.

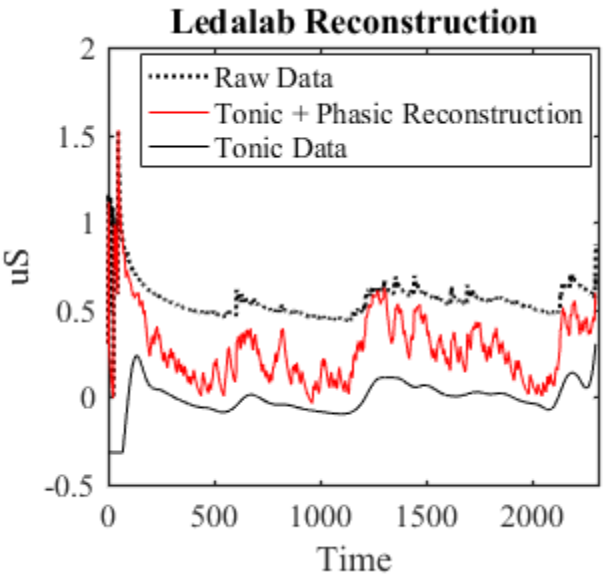


Figure 18: Ledalab's phasic and tonic estimations

In contrast, our estimation, shown in Figure 19, is based on the minima throughout the signal so the fit to the signal follows the original signal more closely and minimal information is lost. The response was also significantly better using our estimate, detecting 60.18% of the possible SCRs.

The major issue with the original methodology used for our tonic and phasic estimation was that, due to the chosen M_T, not every minimum within the signal was used. Due to this, it was possible to get negative responses in the estimated phasic component. Negative responses are hard to fit when attempting to identify SCRs within the signal, and are not possible to achieve naturally. Additionally, having these negative responses in the phasic estimation doesn't align with the underlying assumptions about how the signal is produced. Figure 19 shows an examples of a section of a signal where the raw data is below the tonic estimation, meaning when this tonic is subtracted from the raw data, it will produce a negation response in the phasic component.

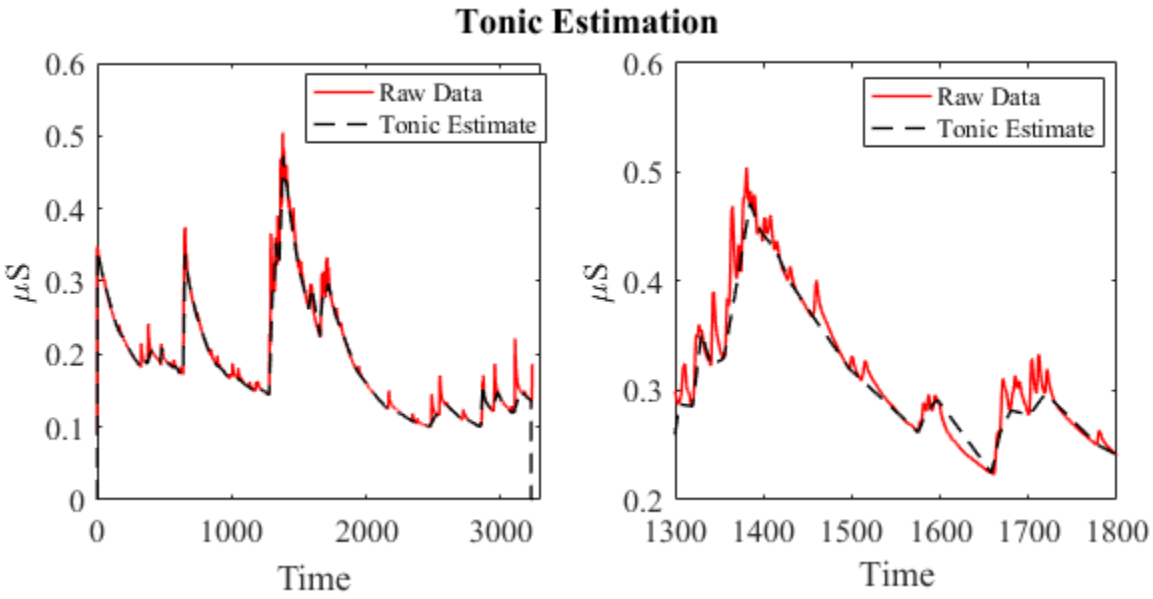


Figure 19: Proposed tonic estimation methodology

In an attempt to find a more robust way to estimate the tonic level and avoid negative responses in our methodology, several filtering and thresholding values were examined and compared. To determine the performance of each tonic estimation method, SCRs where identified

in each estimated phasic component using BOMP, and the accuracy, sensitivity, specificity, and true SCR identification percentage (TIP) were computed. These computed performance measures were then used to determine which method gave the best detection ability. Table 8 shows the average performance for each filtering and M_T combination investigated. It was determined that for the 10 test participants, the best performance was achieved by filtering with a 1Hz low pass filter and an $M_T = 1$ second, producing an average accuracy of 68.63% and a TIP of 66.29%.

Table 8: Average performance calculated for each combination of filtering and thresholding evaluated.

Low Pass Filter	Min Threshold (sec)	Accuracy	Sensitivity	Specificity	TIP (/442)
.35 Hz					
	10	67.96	59.49	73.22	266
	3	67.63	64.07	71.77	267
	2	66.13	64.49	68.94	268
	1	65.89	63.98	69.51	265
.5 Hz					
	5	60.40	61.35	62.48	248
	3	62.50	63.69	62.31	263
	2	63.01	64.72	62.82	262
	1	60.57	63.31	59.82	256
1 Hz					
	5	61.22	61.25	63.59	251
	3	61.29	62.07	63.33	256
	2	63.01	64.72	62.82	262
	1	68.63	68.80	70.83	293

Much of the first study centered around comparing the run time for Ledalab and our methodology. It was shown in 4.1 that the BOMP methodology significantly improved the overall run time of the analysis [39]. While the focus of this study is not on the run time required for our methodology some evaluation of the run time was performed to determine if the changes to the methodology significantly affected the complexity of our analysis. It was determined that in this study the filter chosen when creating the tonic and phasic estimates affected the run time and that

higher cutoff frequencies slowed the run time. We investigated the performance of the tonic estimation by first assessing the .35 Hz filter. This filter was the same used in the previous study, and allowed us to use a sampling frequency of 1 Hz. As in the last study, this study found that using the .35 Hz filter with the BOMP methodology was faster than Ledalab's run time, regardless of the M_T value. However, moving to the 1 Hz filter significantly increased the run time of the BOMP method due to the higher sampling rate needed to avoid aliasing, F_S=3 Hz. This time increase caused our methodology to have a longer run time than Ledalab. However, it was also shown in our previous study that Ledalab's run time increased dramatically more than the BOMP method as the length of the signal increased [39]. This suggests that longer signals could still favor the BOMP methodology, even with the increase in run time caused by the 1 Hz filter. Therefore, further analysis of run time with ambulatory data is need to be able to get an accurate picture of the run time of both systems. As the best overall performance measures were achieved using the 1Hz filter with M_T = 1 second the phasic estimation for all further models used this parameter set.

4.2.4.2 Dictionary Expansion

Using our tonic estimation method with the base dictionary, we were only able to achieve an accuracy of 52.10%. To improve the accuracy, we expanded the dictionary from a knowledge driven dictionary to a data driven one using general trends found from the fits of missed SCRs. To be able to identify SCRs within the signal, the Bateman equation was used to fit approximately 150 missed SCRs and return the ideal $\tau_1$and τ_2 pair for each miss. After completing an initial analysis of the returned parameters, about 30% of the fit data was removed due to bad fits. A bad fit was determined either if convergence wasn't reached in 100000000 iterations or using visual inspection of the returned fit. Figure 20 shows an example of a good fit as well as an example of a bad fit that would have been removed during visual inspection. This fit misses the peak of the SCR

and instead just fits to the recovery giving an overall linear trend. Along with visual inspection, value pairs were removed if the τ_1 or τ_2 value was greater than 1000. It was determined that fits with a τ_1 or τ_2 value over 1000 either led to SCRs with sharp peaks, meaning they had fast rise and recovery times, or produced responses similar to Figure 20. Shapes with fast rise and recovery times are more akin to noise than to true SCRs and were therefore ignored.

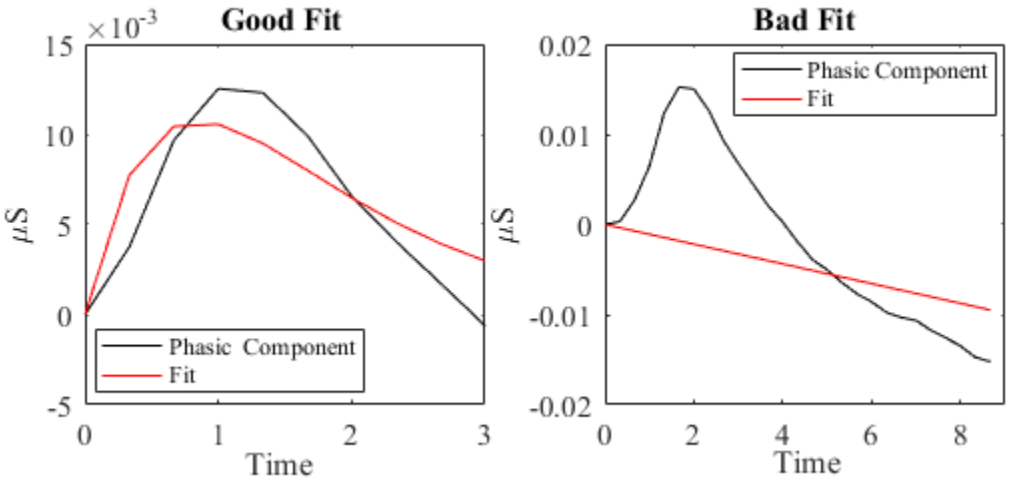


Figure 20: Estimated fits generated from missed SCRs. **(left)** Good fit to be included in further analysis versus **(right)** bad fits removed after inspection

After the bad fits were determined and removed, there were approximately 100 $\tau_1$and τ_2 pairs left which were plotted using a bivariate histogram, shown in Figure 21a, to determine the common values. The histogram showed most of the values falling into a linear trend between the $\tau_1$and τ_2 values, with the τ_2 values being slightly larger. The rest of the parameters generally fell into bins where the τ_2 value was significantly larger than the τ_1 values. Plotting the $\tau_1$and τ_2 values in a scatter plot, Figure 21b, showed 3 or 4 major clusters existing within these two trends.

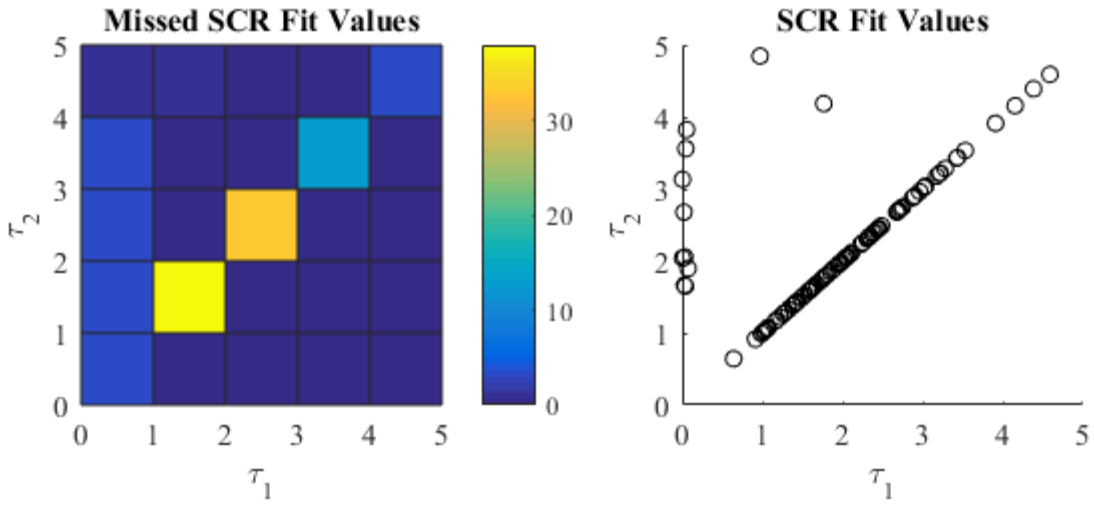


Figure 21: (left) Histogram of $\boldsymbol{\tau_1}$ and $\boldsymbol{\tau_2}$ values returned from estimated fits. **(right)** scatter plot of the same values

Using these clusters and the most common bins produced from the histogram, 6 additional SCR shapes were chosen to expand the dictionary; the new dictionary therefore included atoms for each new shape as well as the original shape used. Figure 22 shows the shape of the additional atoms added to the dictionary. While determining the additional shapes, it was found that the general shape of each new atom was relatively the same, and the major difference between the new shapes were the times required for the SCR to reach half and full recovery. The 6 new shapes, therefore, gave a range of SCR length between 5 and 20 seconds, which matches with previously reported ranges of SCR lengths. This new dictionary was then used to re-identify SCRs in each test signal with significantly improved performance over the initial accuracy reported above (see section 4.2.4.3for details).

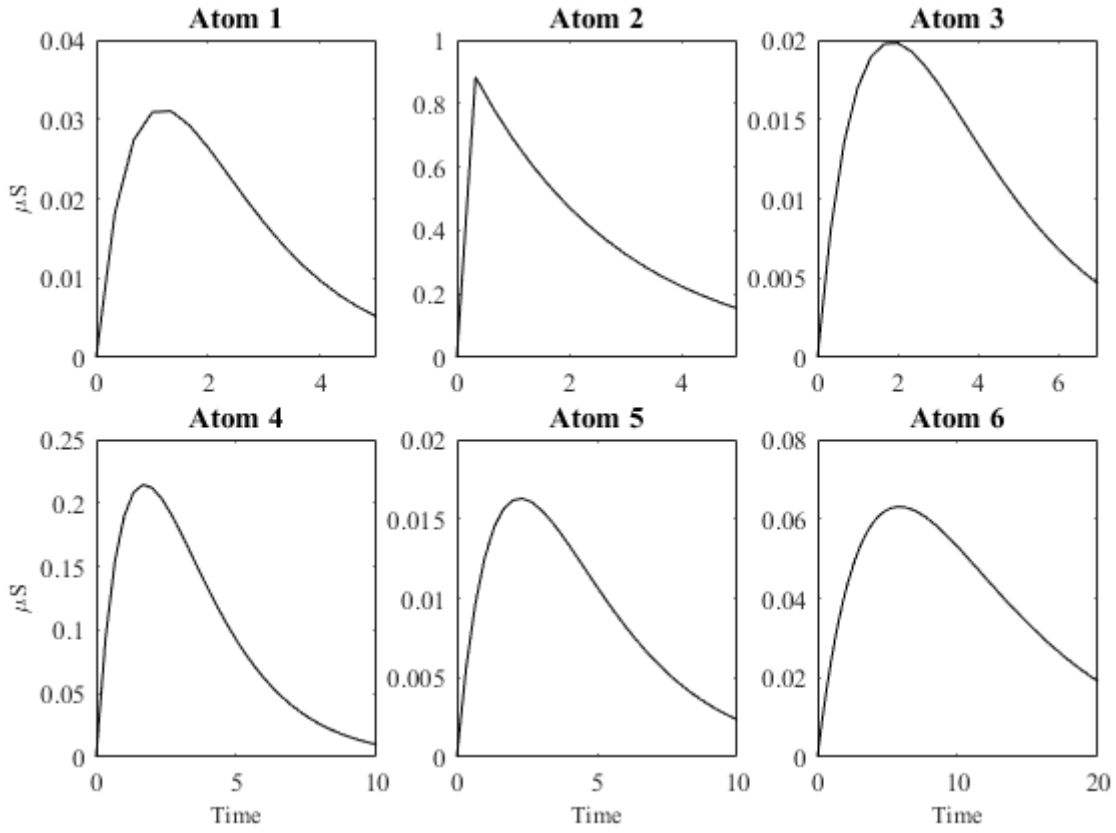
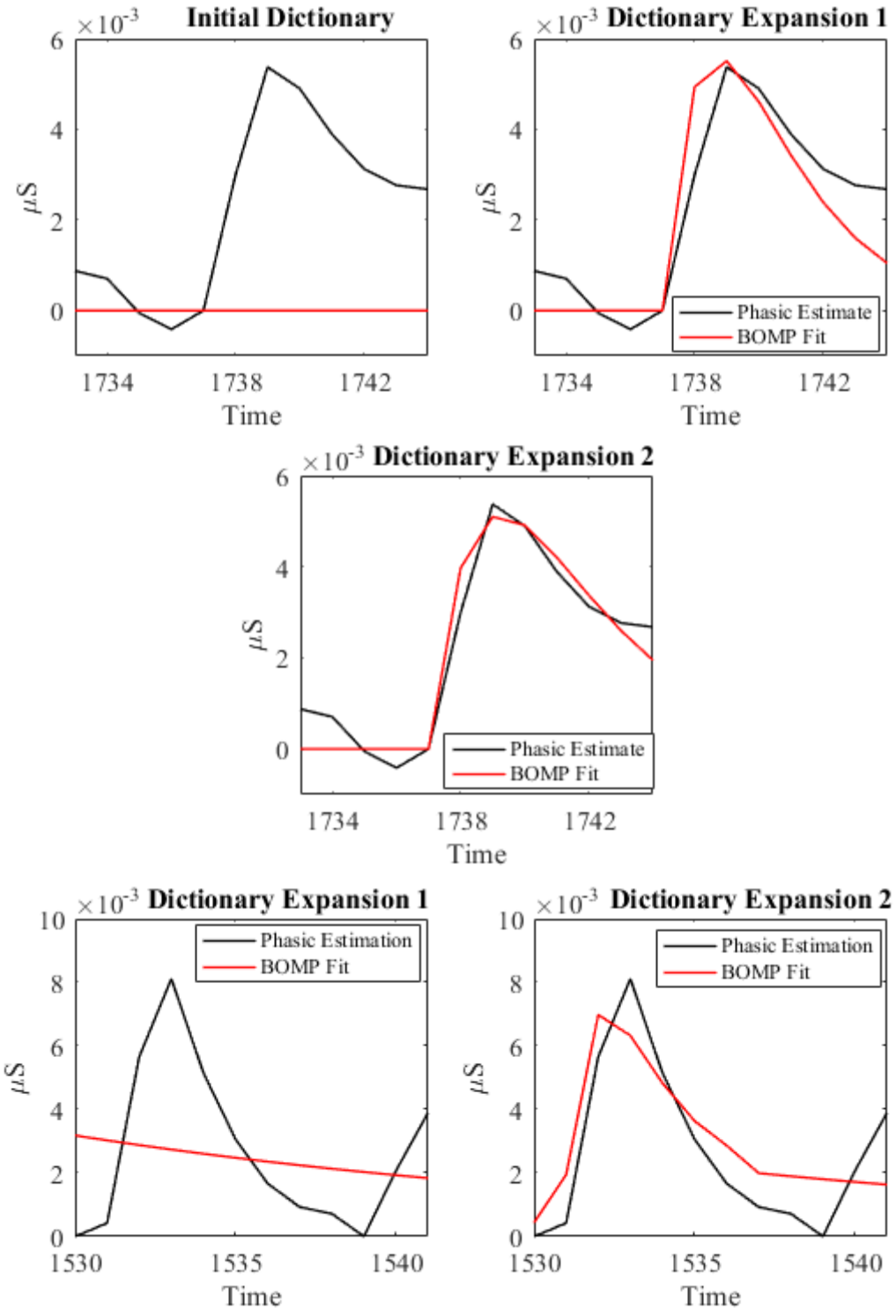


Figure 22: New SCR shapes used to expand our initial dictionary

To determine the 6 shapes shown above and ensure that redundant atoms were not added, two separate expansions were done. The first expansion lead to the addition of 2 of the additional atoms. To investigate the utility of the added atoms, SCRs that were missed when the initial dictionary was used were evaluated after the addition of the new atoms. It was found that adding the 2 additional atoms increased the overall accuracy of the system from 52% up to 63% and that many of the missed SCR were captured. However, there were still a significant number of SCRs that were missed. When the histogram method was used to determine new atoms it returned atoms with different parameters then were previously added. A second expansion of the dictionary was therefore done which added the remaining atoms shown previously. This expansion again increased the overall accuracy of our system up to 67%. Evaluating missed SCRs based on the

second dictionary expansion showed that the missed shapes fell mostly into the same bins as previously added atoms. Adding additional shapes would therefore not have given us significant improvements in accuracy and would have likely slowed our system. Figure 23 below shows the progression of SCR detection through the 3 different dictionaries used. Figure 23a shows an SCR which was missed by the initial dictionary but was then captured after the first expansion and was again captured after the second expansion. Figure 23b shows an SCR which was missed after the first expansion of the dictionary but was then successfully identified after the second expansion.



(a)

(b)

Figure 23a: SCR successfully captured after the first dictionary expansion and **b:** SCR successfully captured only after the second dictionary expansion

4.2.4.3 BOMP vs. Ledalab

In this section we compare the ability of two algorithms, Ledalab and our novel approach, to identify SCRs in an EDA dataset using accuracy, sensitivity, specificity, and number of SCRs identified as performance measures. It was found that our novel approach produced results that were more robust than those of Ledalab. Additionally, the ability of our method to be extended to new data was tested. We found that using the BOMP algorithm with a data driven dictionary also showed a good ability to fit to new EDA signals.

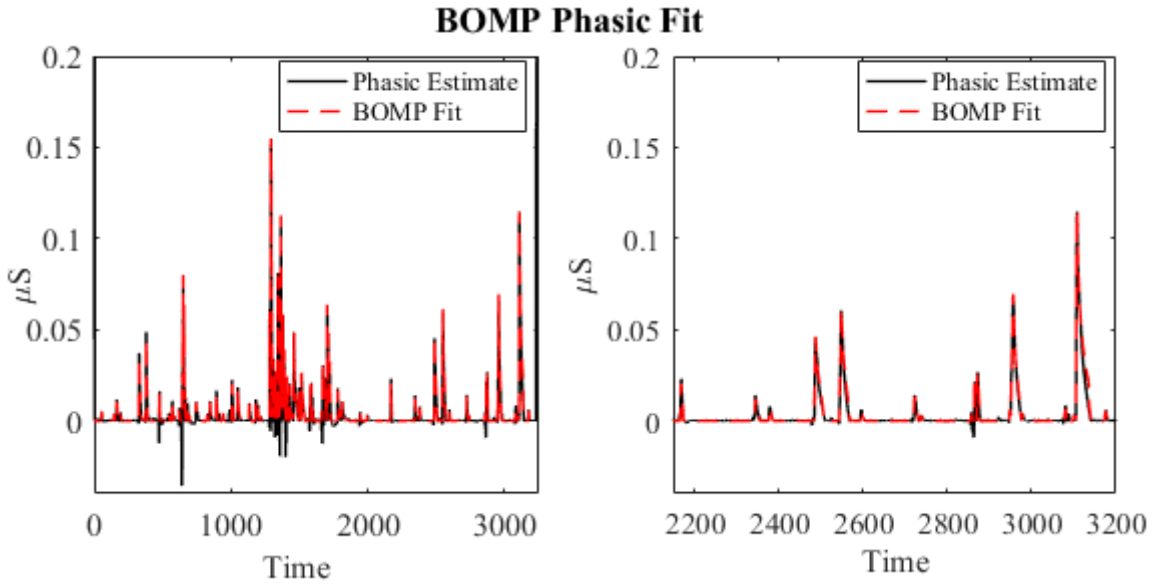


Figure 24: Phasic estimated created using our methodology with the BOMP fit overlaid

Figure 24a shows the BOMP fit based off of our phasic estimation which, through visual inspection, suggests that BOMP captures the overall peaks and trends of the original signal. Capturing the general trends indicates that SCRs throughout the signal have been captured. To give a clearer view of the original and fit, Figure 24b shows Figure 24a zoomed in between 2150 and 3200 seconds. Quantitatively, the BOMP approach, with post processing of the estimated

gamma values, was able to identify SCRs with an average accuracy of 68.63% compared to the Ledalab average of 69.01%. Looking at the sensitivity, 68.80% for BOMP versus 41.46% for Ledalab, shows an advantage towards the BOMP methodology. Additionally, the BOMP methodology detected 66.29% of all possible SCRs within the 10 test participants, a significant improvement over the 45.02% of SCRs detected by Ledalab. Though our methodology increased sensitivity of SCR detection compared to Ledalab, specificity was lower. However, as the increase in sensitivity was significantly greater (~30%) than the decrease in specificity (~17%) our methodology gives an overall improvement in the ability to accurately detect SCRs. Table 9 shows the performance measures calculated for each of the 10 test participants and the averages for both Ledalab and our novel approach.

Table 9: Performance by participant for our methodology (left) and Ledalab's analysis (right)

	Sparsity	Accuracy	Sensitivity	Specificity	# SCRs		Accuracy	Sensitivity	Specificity	# SCRs
1	400	71.11	77.27	69.12	17/22	1	83.82	50.00	86.36	11/22
2	200	89.81	85.71	92.31	12/14	2	87.25	7.14	100.00	1/14
3	400	55.74	51.39	62.00	37/72	3	71.19	62.50	84.78	45/72
4	350	66.67	77.03	50.00	57/74	4	47.66	32.43	81.82	24/74
5	400	63.33	41.67	71.21	10/24	5	72.41	0.00	100.00	0/24
6	350	60.26	52.46	78.95	32/59	6	53.09	44.07	77.27	26/59
7	300	74.34	84.91	65.00	45/53	7	77.78	94.34	64.06	50/53
8	350	66.05	88.24	61.63	15/17	8	88.00	64.71	92.77	11/17
9	350	68.69	60.87	86.67	42/69	9	45.45	26.09	90.00	18/69
10	350	70.37	68.42	71.43	26/38	10	63.46	33.33	81.54	13/38
Avg	-	68.80	70.83	69.00	293/442	Avg	69.01	41.46	85.86	199/442

To determine the statistical significance of these performance measures across the test participants a one-sided Wilcoxon test was used. Testing the difference between Ledalab's accuracies and our accuracies returned no statistical difference, p-value =.5733, between the techniques. However, results indicated the sensitivity of our approach is statistically significantly greater than Ledalab's (p = .005), whereas Ledalab's specificity was significantly greater than

ours, p = .007. This supports our prior conclusions, indicating that while the accuracies of the two methodologies are statistically the same, our methodology has greater sensitivity while Ledalab

We noted in section 2.1.3 that the sparsity system parameter, K, was varied between 10-400 to find the optimal value. As can be seen from Table 9a, the ideal K was between 200 and 400 for the participants. The variability seen is due to the role that sparsity plays in the goodness of fit and subsequently the calculated performance measures. With the lowest sparsity, specificity will be the maximum achievable for the participant, and shows a decrease as sparsity increases. Inversely, sensitivity is low with low sparsity, and increases as sparsity increases. Figure 25 shows this general trend for a single participant. This suggested that by adjusting the sparsity, sensitivity and specificity could be adjusted. Therefore, the best sparsity value was picked to maximize and balance both the sensitivity and specificity.

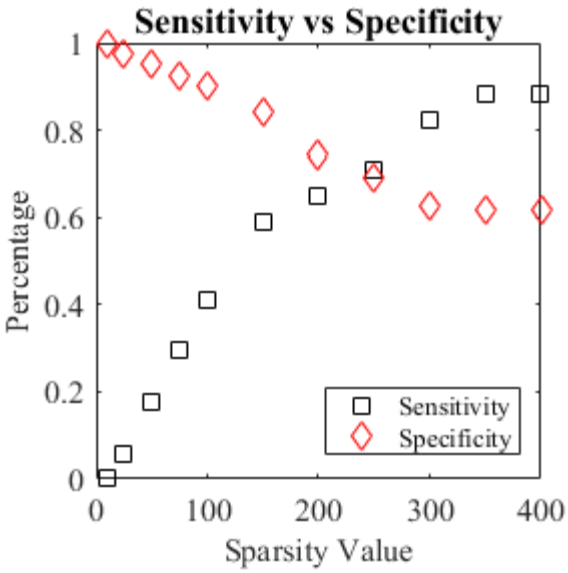


Figure 25: Sensitivity vs average performance

To show the ability of our method to be generalized, our full method was applied to the validation signals that were set aside at the beginning of the analysis. Table 10 shows the average performance values across the 42 validation participants. The returned performance measures from the validation set are similar, if not slightly improved, to what was found using the testing set. Additionally, the ranksum test was again used to determine the statistical similarity between the test set and the validation results. It was found that all of the performance measures reported were statistically equivalent, each accepting the null hypothesis. The similarity in the validation sets performance shows the ability of the BOMP methodology to be extended to new data without requiring modification of the dictionary. This ability to successfully identify SCRs in the validation set suggests that overfitting was avoided and that our methodology can be easily expanded to ambulatory data without requiring significant modifications.

Table 10: Average performance found using the validation participants.

Method	Accuracy	Sensitivity	Specificity	Total SCRs detected
Ledalab analysis	66.96%	19.54%	94.52%	350/1727
BOMP analysis	70.49%	73.82%	62.55%	1189/1727

4.3 ARTIFACT DETECTION WITH SPARSE RECOVERY

Our final study focused on artifacts and looked into the seperability of artifacts and SCRs if both were modeled using the Bateman equation. Artifact detection is important in EDA analysis due to the nature of EDA signals and how they are collected. Since EDA is a measure of the conductance of the skin, caused by the activity of sweat glands, EDA can be highly affected by artifacts if the environment and movement are not carefully controlled for during collection. Reducing the need of careful control during collection by developing a robust method to identify artifact would

improve the analysis of EDA signals and allow for more ambulatory collections to be fully analyzed.

4.3.1 Study Goals and Experimental Data

In our final experiment we aim to investigate the classification ability between artifacts and SCRs, using the Bateman equation parameters as features, to determine the possibility of including artifact columns in the dictionary.

To fit and analyze artifacts the data used in experiment 2 was again used this time focusing on participants that had labeled artifacts see section 4.2.2 for further details.

4.3.2 Analysis Methods

To investigate the possibility of being able to distinguish between artifacts and SCRs, several labeled artifacts were fit using the Bateman equation, which returned the τ_1 and τ_2 value pairs for each artifact. The returned τ_1 and τ_2 parameters for artifacts and previously determined τ_1 and τ_2 value for SCRs were then used as features in support vector machine (SVM) and discriminate analysis classification. For SVM classification, linear and radial basis function classification schemes were investigated. To determine the optimal parameters for sigma and the box constrain a range between .001 and 1000 were tested for both. Additionally, linear and quadratic schemes were used in the discriminate analysis classification. Due to a low number of labeled artifacts, both classification methodologies used leave-one-out analysis to split the data into testing, training, and validation sets.

4.3.3 Results and Discussion

To determine the seperability between artifacts and SCRs, the parameters returned form the Bateman equation, SVM and discriminant analysis were used to perform classification. Quadratic discriminant analysis (QDA) gave better classification, showing accuracies of 73.66%. After labeling was completed on the 65 participates, approximately 30 sections of data across 11 participants that contained labeled artifacts. This lead to approximately 50 τ_1 and τ_2 parameters returned after fits were found for each artifact. The low number of identified artifacts was expected due to signal collection conditions. As the lab based experiment minimized artifact, it was difficult to include a more holistic artifact detection scheme. Still, 50 fits allowed an investigation into the shape of artifacts and the possibility of distinguishing between artifact and SCR fits in future work. Plotting the fit parameters for both artifacts and SCRs, Figure 26, suggests that some seperability may exist. This was supported by the classification achieved using QDA. QDA gave a classification accuracy of 73.66% with 81.25% sensitivity for artifact classification and 71.34% sensitivity towards SCR classification.

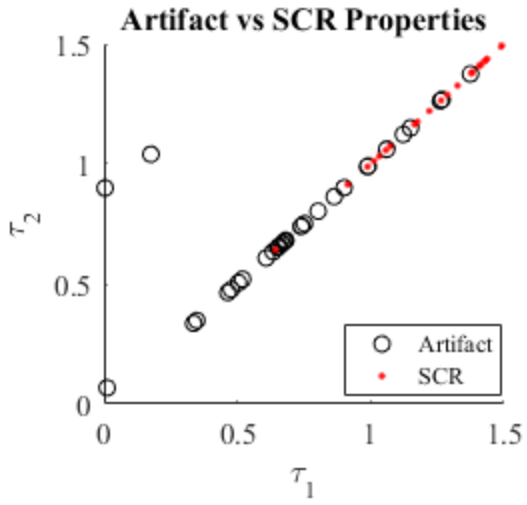


Figure 26: Parameters required to fit to SCRs and Artifacts

5.0 CONCULSIONS AND FUTURE WORK

5.1 CONCLUSIONS

Skin conductance response has been a widely used measure of physiological states in part because of its sensitivity to changes in arousal as well as the ease in which EDA data can be measured [1]. The recent emergence of ambulatory data collation with EDA data has allowed many studies to be expanded, and data to be collected for longer time periods and in a variety of settings [3], [8], [16]. However, this new data collection style has introduced a number of challenges in processing and analyzing EDA signals. One of the major challenges is simply the time that is required to analyze large datasets typical of ambulatory recordings. The second major issue is the need for a robust way to automate the processing of EDA signals, and remove the need for manual work. Automation of EDA analysis should include the ability to robustly and discreetly identify SCRs and artifacts in the signal. This thesis presented a novel application of curve fitting, sparse recovery, and dictionary learning to addresses the major challenges associated with analyzing ambulatory EDA data.

In addressing the first issue our first experiment showed that our methodology could be run with an average run time of 30 seconds for a signal data set with average length of 3 hours. Ledalab, an existing EDA analysis program, in comparison produced a run time average of 175.5 seconds

for the same length data. Additionally, it was shown that our methodology had an exponential decrease in run time compared to Ledalab.

In our second experiment we were able to address several issues found during our initial experiments while also fully evaluating our systems ability to successfully identify SCRs in a signal. Using a data-driven dictionary our system achieved an average accuracy of 68.80%, sensitivity of 70.83% and specificity of 69.00% with an ability to positively detect 66.29% of all the SCRs in the test signals. This was a significant improvement in sensitivity and SCR detection over currently available EDA analysis software. We were also able to generalize our dictionary by developing a robust tonic estimation methodology that could be used to estimate the phasic component of the signal, removing the need for tonic atoms in the dictionary. Additionally, we showed the ability for our methodology to be extended to new data with an accuracy of 70.49% and detection rate of 68.85% without the dictionary or methodology needing to be modified. This flexibility suggests that our methodology can continue being extended to new data. including ambulatory data, without requiring significant modification of the methodology or dictionary. Finally, to begin addressing the issue of artifact identification our final experiment showed that using only the τ_1 and τ_2 values returned from the Bateman equation as features artifacts could be distinguished from SCRs. Using extract τ_1 and τ_2 parameters to describe both artifacts and SCRs, classification accuracies of 73.66% were achieved using a simple QDA analysis. Achieving high classification accuracies of artifacts and SCRs suggests that artifacts could be included as new atoms in the dictionary, allowing our analysis to detect artifacts, without compromising the ability to detect SCRs.

5.2 FUTURE WORK

Our expanded data driven dictionary significantly improves upon our previously introduced methodology, but leaves several outstanding issues. Future work is needed to further improve the robustness and efficiency of EDA analysis. The two major outstanding drawbacks to our methodology are better balancing false positives and true positives and to improve upon artifact detection.

About 17% of SCRs missed after our best run were caused by either (1) the filtering and phasic estimation or (2) incorrectly counting misses due to post processing. The first type of misses could be better handled through further analysis of the optimal parameters for our tonic estimation methodology. Prior to our analysis we hypothesized that a lower minimum time distance, M_T, would allow for more compound SCRs to be correctly identified; but could increase our false positive detection due to noise. Higher M_T values, in contrast, would likely reduce fitting to noise but could limit our ability to detect compound SCRs. Unfortunately, due to a low number of labeled compound SCRs, this could not be satisfactorily studied within the current manuscript. Using data collected with a rapid stimuli paradigm, however, could generate more compound SCRS, allowing us to fully investigate our hypothesis, and better identify optimal parameters. A study which employs a rapid stimuli paradigm would purposefully attempt to elicit new responses before the previous response could fully recover [10], [2] increasing the probability of compound SCRs. Additionally, data with more compound SCRs, would allow us to address the second issue, miscounting identified SCRs as misses. In our analysis we used a threshold of 1 second between chosen dictionary atoms, based on prior knowledge, in favor of reducing double counting SCRs. This thresholding, however, limited our ability to fully capture compound SCRs that were elicited less than a second apart. Further analysis of the typical parameters of compound SCRs, as

mentioned above, could allow us to balance capturing all of the SCRs while avoiding additional false positives.

The second issue remaining is to include artifact atoms into our dictionary as it has been shown that artifacts can be separated from SCRs with high accuracy. To incorporate these atoms, data with artifacts not removed by filtering is needed. In addition, longer more complex signals such as those collected using ambulatory sensors will aid in the testing and development of artifact atoms.

We hope to keep the computational complexity of our analysis low. As we continue to expand our dictionary to include more atoms to describe both SCRs and artifacts the dictionary could increase to an unmanageable size, both in terms of memory requirements as well as increase computational complexity. To develop a universal dictionary that could be used across data sets we therefore want to investigate the use of dictionary learning techniques that could be used to reduce the dictionary size [28], [32]. We envision using our fully expanded dictionary a library that can then be pruned, through dictionary learning algorithms. This will create a smaller dictionary that is specific to the given data set. Further analysis of current dictionary learning techniques and the development of new algorithms will need to be investigated to determine an efficient way to prune a large library while keeping run times low and allowing for high SCR detection ability.

Further improvements on the proposed methodology as discussed above could be used to increase the efficiency of the analysis. Additionally, investigation into the application of other methodologies could show further benefits in decreasing computational complexity or in detection ability over our current approach. One appropriate wavelet could be found additional SCR shapes could be tested without needing to continue expanding our dictionary but simply by transforming

the mother wavelet. This could further decrease the computational complexity of our analysis and would potentially address the issue of our dictionary becoming too memory intensive to be run efficiently. Another area to investigate is if deep learning techniques could be used efficiently. Over the past decade increased investigations into deep learning has increased the efficiency and utility of these methodologies. While deep learning is typically computationally complex, if training could be done off line and remain general enough to be applied to novel EDA deep learning techniques could be used without increasing analysis time [49]. One deep learning technique that has shown promise when applied to other biological signals, including P300 detection in EEG, is convolutional Neural Networks [49], [50].

Finally, our work could be extended to address outstanding questions specific to EDA and phycology. One potential extension would be to investigate the shape of SCRs elicited by specific stimulus or responses that result based on specific emotional responses. Being able to distinguish between specific emotions or be able to link a SCR shape to a specific response would offer greater insight into the analysis of EDA data. While it is unlikely this type of seperability could be found this type of investigation, to our knowledge, has not been previously done and could offer a deeper knowledge of the link between SCRs and the autonomic nervous system. More generally, investigations into the differences between specific and non-specific response shapes could be evaluated to determine again if the type of responses could be distinguished giving the same insight that being able to classify individual responses could [19]. Additionally, being able to distinguish between specific and non-specific responses would further enhance the ability to identify SCRs of interest in a signal since typically specific SCRs contain more relevant information and would be the more desirable SCRs to identify [19].

BIBLIOGRAPHY

[1] J. T. Cacioppo, L. G. Tassinary and G. Berntson, Handbook of Psychophysiology, Cambridge: Cambridge Up, 2007.

[2] C. L. Lim, C. Rennie, R. J. Barry, H. Bahramali, I. Lazzaro, B. Manor and E. Gordon, "Decomposing Skin conductnace into tonic and phasic components," *Journal of Psychophysiology*, vol. 25, pp. 97-109, 1997.

[3] C. Kappeler-Setz, F. Gravenhorst, J. Schumm, B. Arnrich and G. Troster, "Towards long term monitoring of electrodermal activity in daily life," *Personal and ubiquitous computing,* vol. 17, no. 2, pp. 261-271, 2013.

[4] H. Storm, A. Fremming, S. Odegaard, O. G. Martinsen and L. Morkid, "The development of a softwawre program for analyzing spontaneous and externally elicited skin conductance changes in infants and adults," *Clinical Neurophysiology,* vol. 111, no. 10, pp. 1889-1898, 2000.

[5] J. T. Cacioppo and L. G. Tassinary, "Inferiring psychological Significance from Psysiological Signals," *The American Psychologist,* vol. 45, no. 1, pp. 16-28, 1990.

[6] J. Hernandez, R. R. Morris and R. W. Picard, "Call Center Stress Recognition with Person-specific Models," Springer, 2011, pp. 125-134.

[7] A. Sano and R. W. Picard, "Stress Recognition using Wearable Sensors and Mobile Phones," in *Humaine Association Conference on Affective Computing and Intelligent Interaction (IEEE)*, 2013.

[8] S. Doberenz, W. T. Roth, E. Wolburgh, N. I. Maslowski and S. Kim, "Methodological considerations in ambulatory skin conductance monitoring," *International Journal of Psychophysiology,* vol. 80, no. 2, pp. 87-95, 2011.

[9] W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. Roth, M. E. Dawson and D. L. Filion, "Publication Recommendations for Electrodermal

Measurements," *Psychophysical Psychophysology,* vol. 49, no. 8, pp. 1017-1034, 2012.

[10] D. M. Alexander, C. Trengrove, P. Johnston, T. Cooper, J. August and E. Gordon, "Separating individual skin conductance responses in a short interstimulus-interval paradigm," *Journal of Neuroscience Methods,* vol. 146, no. 1, pp. 116-123, 2005.

[11] R. Edelberg, "Mechanisms of Electrodermal adaptions for locomation, manipulation, or defense," *Progress in Physiological Psychology,* vol. 8, pp. 155-209, 1973.

[12] S. Taylor, N. Jaques, W. C. Chen, S. Fedor, A. Sano and R. Picard, "Automatic Identification of Artifacts in Electrodermal Activity Data," in *Engineering in Medicine and Biology SCociety (IEEE),* 2015.

[13] D. C. Fowles, M. J. Christie, R. Edelberg, W. W. Grings, D. T. Lykken and P. H. Venables, "Publication Recommendations for Electrodermal Measurements," *Psychophysiology,* vol. 49, pp. 1017-1034, 2012.

[14] C. L. Lim, C. Rennie, R. J. Barry, H. Bahramali, I. Lazzaro, B. Manor and E. Gordon, "Decomposing skin conductance into tonic and phasic components," *International Journal of Psychophysiology,* pp. 97-109, 1997.

[15] D. R. Bach, G. Flandin, K. J. Friston and R. J. Dolan, "Modelling event-related skin conductance responses," *International Journal of Psychophysiology,* vol. 75, no. 3, pp. 349-356, 2010.

[16] R. Kocielnik, N. Sidorova, F. M. Maggi, M. Ouwerkerk and J. H. Westerink, "Smart Technologies for Long-Term Stress Monitoring at Work," in *IEEE International Symposiumon Computer-Based Medical Systems* , 2013.

[17] D. R. Bach, "A Head-to-head comparison of SCRalyze and Ledalab; Two model-based methods for skin conductance Analysis," *Biolocial Psychology,* vol. 103, pp. 63-68, 2014.

[18] M. Benedek and C. Kaernbach, "Decomposition of skin conductance data by means of nonnegative deconvolution," *Psychophsiology,* vol. 47, no. 4, pp. 647-658, 2010.

[19] D. R. Bach, K. J. Friston and R. J. Dolan, "Analytic measures for quantification of arousal from spontaneous skin conductance fluctuations," *Journal of Psychoophysiology,* vol. 76, pp. 52-55, 2010.

[20] D. R. Bach, K. J. Friston and R. J. Dolan, "An improved algorithm for model-based analysis of evoked skin conductance responses," *Biological Psychology,* vol. 94, pp. 490-497, 2013.

[21] M. Benedek and C. Kaernback, "A Continuous Measure of Phasic Electrodermal Activity," *Journal of Neuroscience Methods,* vol. 190, no. 1, pp. 80-91, 2010.

[22] A. Blanch, F. Balada and A. Aluja, "Presentation and AcqKnowledge: An application of software to study human emotions and individual differences," *Computer Methods and Programs in Biomedicine,* vol. 110, pp. 89-98, 2013.

[23] S. R. Green, P. A. Kragel, M. E. Fecteau and K. S. LaBar, "Development and Validation of an Unsupervised Scoring System (Autonomate) for Skin Conductance Response Analysis," *Journal of Psychophysiology,* vol. 91, pp. 186-193, 2014.

[24] V. Xia, N. Jaques, S. Taylor, S. Fedor and R. Picard, "Active Learning for Electrodermal Activity Classification," in *Signal Processing in Medicine and Biology Symposium*, 2016.

[25] D. R. Bach, J. Daunizeau, K. J. Friston and R. J. Dolan, "Dynamic Causal Modelling of Anticipatory Skin Conductance Responses," *Biological Psychology*, vol. 85, pp. 163-170, 2010.

[26] T. Chaspari, A. Tsiartas, L. I. Stein, S. A. Cermak and S. S. Narayanan, "Sparse Representation of Electrodermal Activity With Knowledge-Driven Dictionaries," *IEEE Transactions on Biomedical Engineering* , vol. 62, no. 3, pp. 960-971, 2015.

[27] G. Rath and A. Sahoo, "A comparative Study of some Greedy Pursuit Algorithms for Sparse Approximation," in *Signal Processing Conference (IEEE)*, 2009.

[28] M. J. Gangeh, A. Ghodsi and M. S. Kamel, "Kernelized Supervised Dictionary Learning," *IEEE Transactions on Signal Processing,* vol. 61, no. 19, 2013.

[29] S. Chen and D. Donoho, "Basis Pursuit," in *28th Asilomar Conference of Signals, Systems, and Computers,* 1994.

[30] G. Z. Karabulut and A. Yongacoglu, "Estimation of Time-varying Channels with Orthogonal Matching Pursuit Algorithm," in *IEEE/ Sarnoff Symposium on Advance in Wired and Wireless Communication,* 2005.

[31] J. Mairal, F. Bach, J. Ponce, G. Sapiro and A. Zisserman, "Supervised Dictionry Learning," in *Advances in Neural Information Processing* , 2008.

[32] M. Aharon, M. Elad and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing,* vol. 54, no. 11, 2006.

[33] K. Kreutz-Delgado, J. F. Murray and B. D. Rao, "Dictionary Learning Algorithms for Sparse Representation," *Neural Computation,* vol. 15, pp. 349-396, 2003.

[34] R. Rubinstein, M. Zibulevsky and M. Elad, "Efficient Implementation of the K-SVD Algorithm using Batch Orthogonal Matching Pursuit," *CS Technion,* vol. 40, no. 8, pp. 1-15, 2008.

[35] S. Kunis and H. Rauhut, "Random Sampling of Sparse Trigonometric Polynomials, II. Orrthogonal Matching Pursuit versus Basis Pursuit," *Foundations of COmputational Mathematics,* vol. 8, no. 6, pp. 737-763, 2007.

[36] J. Alder, B. Rao and K. Kreutiz-Delgado, "Comparision of Basis Selection Methods," in *30th Asilomar Conference on Signals, Systems, and Computers*, 1996.

[37] H. Zhang, Y. Zhang and T. S. Huang, "Simultaneous discriminative projection and dictionary learning for sparse representation based classification," *Pattern Recognition,* vol. 46, pp. 346-354, 2013.

[38] T. T. Cai and L. Wang, "Orthogonal Matching Pursuit for Sparse Siignal Recovery with Noise," *IEEE Transactions on Information Theory,* vol. 57, no. 7, 2011.

[39] M. Kelsey, A. Dallal, S. Eldeeb, M. Akcakaya, I. Kleckner, C. Gerard, K. S. Quigley and M. S. Goodwin, "Dictionary Learning and Sparse Recovery for Electrodermal Activity Analysis," in *SPIE Commercial Scientific Sensing and Imaging*, Baltimore, 2016.

[40] D. R. Bach, G. Flandin, K. J. Friston and R. J. Dolan, "Time-series analysis for rapid event-related skin conductance analysis," *Journal of Neuroscience Methods*, vol. 184, no. 2, pp. 224-234, 2009.

[41] J. D. Gibbons and S. Chakraborti, "Nonparametric Statistical Inference," in *International Encyclopedia of Statistical Science*, Springer Berlin Heidelberg, 2011, pp. 977-979.

[42] D. Eatson, L. A. Clark and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the PANAS scales," *Journal of Personality and Social Psychology,* vol. 54, no. 6, pp. 1063-1070, 1988.

[43] I. R. Kleckner, J. B. Wormwood, W. K. Simmons, L. F. Barrett and K. S. Quigley, "Methodological recommendations for a hearbeat detection-based measure of interoceptive sensitivity," *Psychophysiology,* vol. 52, no. 11, pp. 1432-1440, 2015.

[44] P. J. Lang, M. M. Bradley and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," University of Florida, Gainesville, Fl, 2008.

[45] D. H. Brainard, *The psychophysics toolbox 10,* Spatial Vision, 1997.

[46] M. Kleiner and D. P. D. Brainard, *What's new in psychtoolbox-3*, Perception, 2007.

[47] D. G. Pelli, *The VideoToolbox software for visual psychhophysics: Transforming numbers into movies,* Spatial Vision, 1997.

[48] M. M. Bradely and P. J. Lang, "Measuring emotion: the Self-Assessment Manikin and the Semantic Differential," *Journal of Behavior Therapy and Experimental Psychiatry,* vol. 25, no. 1, pp. 49-59, 1994.

[49] L. Deng and D. Yu, "Deep Learning MEthods and Applications," *Foudations and Trends in Signal Processing,* vol. 4, no. 4, pp. 297-416, 2012.

[50] H. Cecotti and A. Graser, "Convolutional Neural Networks for P300 Detection with Applications to Brain-Computer Interfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, no. 3, pp. 433-445, 2011.