Biomedical Signal Processing and Control 40 (2018) 58–70

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Research paper

Applications of sparse recovery and dictionary learning to enhance analysis of ambulatory electrodermal activity data

Malia Kelsey[a,*], Murat Akcakaya[a], Ian R. Kleckner[b], Richard Vincent Palumbo[c], Lisa Feldman Barrett[c], Karen S. Quigley[d], Matthew S. Goodwin[c]

[a] University of Pittsburgh, Pittsburgh, PA, United States, United States
[b] University of Rochester Medical Center, Rochester, NY, United States, United States
[c] Northeastern University, Boston, MA, United States, United States
[d] Edith Nourse Rogers Memorial (VA) Medical Center, Bedford, MA and Northeastern University, Boston, MA, United States, United States

ARTICLE INFO

Article history:
Received 9 December 2016
Received in revised form 17 June 2017
Accepted 24 August 2017

Keywords:
Skin conductance response
Electrodermal activity
Sparse recovery
Orthogonal matching pursuit
Artifact detection

ABSTRACT

Electrodermal Activity (EDA) – an index of sympathetic nervous system arousal – is one of the primary methods used in psychophysiology to assess the autonomic nervous system [1]. While many studies collect EDA data in short, laboratory-based experiments, recent developments in wireless biosensing have enabled longer, 'out-of-lab' ambulatory studies to become more common [2]. Such ambulatory methods are beneficial in that they facilitate more longitudinal and environmentally diverse EDA data collection. However, they also introduce challenges for efficiently and accurately identifying discrete skin conductance responses (SCRs) and measurement artifacts, which complicate analyses of ambulatory EDA data. Therefore, interest in developing automated systems that facilitate analysis of EDA signals has increased in recent years. Ledalab is one such system that automatically identifies SCRs and is currently considered a gold standard in the field of ambulatory EDA recording. However, Ledalab, like other current systems, cannot distinguish between SCRs and artifacts. The present manuscript describes a novel technique to accurately and efficiently identify SCRs and artifacts using curve fitting and sparse recovery methods We show that our novel approach, when applied to expertly labeled EDA data, detected 69% of the total labeled SCRs in an EDA signal compared to 45% detection ability of Ledalab. Additionally, we demonstrate that our system can distinguish between artifact and SCR shapes with an accuracy of 74%. This work, along with our previous work [3], suggests that matching pursuit is a viable methodology to quickly and accurately identify SCRs in ambulatory collected EDA data, and that artifact shapes can be separated from SCR shapes.

1. Introduction

Electrodermal Activity (EDA) – an index of sympathetic nervous system activity – is one of the primary methods employed in psychophysiological research [4] and is widely used to quantify autonomic and psychological arousal [5]. Formally, EDA is a measure of electrical conductance on the skin surface, which changes as sweat is released by eccrine sweat glands [6]. Fluctuations in skin conductance are linked to a specific set of brain circuitry [7], and can be used to reveal when psychologically salient events occur. Using this link, EDA has been widely employed to investigate a variety of psychological states, including stress, depression, anxiety, attention, pain, and information processing [8,9,1].

EDA signals are traditionally separated into three distinct components: skin conductance level (SCL); skin conductance response (SCR); and artifacts. SCL, or tonic response, is a slowly fluctuating response that typically ranges between 2 and 20 μS and reflects general trends in level of activation. It is common to remove the tonic level from an EDA signal during analyses given that 1) it is less clear how psychological events relate to tonic changes [1] and 2) EDA baselines are rarely consistent within or between individuals due to hydration status, recording site, eccrine sweat gland density at site of recording, and psychological state [1]. In contrast, SCRs are quick responses superimposed on the tonic response that can be more directly linked to psychological events [10]. SCRs typically have a predictable shape that can be characterized by rise time, amplitude, and half recovery time. In healthy adults, rise time is

* Corresponding author.
E-mail address: mak341@pitt.edu (M. Kelsey).

http://dx.doi.org/10.1016/j.bspc.2017.08.024
1746-8094/

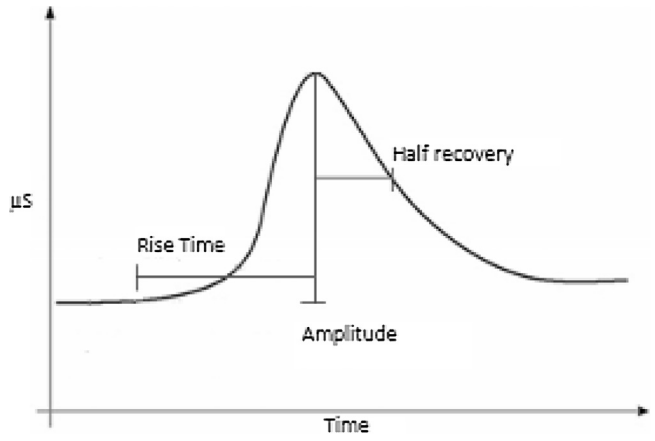

Fig. 1. Parameters used to characterize SCRs.

usually between 1 and 3 s, amplitude often varies, but a minimum is commonly set between 0.01 and 0.05 μS, and half recovery time is typically between 2 and 10 s [1]. Fig. 1 shows the typical shape and parameters that can be 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 compound SCRs, indicates that two separate stimuli or psychologically different events have occurred [11]. As compound SCRs may be caused by different stimuli, accurate identification of each SCR is important during analysis. Finally, a common feature in EDA data are artifacts resulting from contact changes (i.e., increased or decreased pressure of the senor on the skin), wearer movement, shifts in ambient environmental temperature, or electrical interference. While the curvature of an artifact can vary widely, they are often, and problematically, similar in shape and phase to SCRs. Due to this similarity between artifacts and SCRs, identifying artifacts using current practices is a challenging and manually intensive endeavor.

Until the early 2000s, most studies employing EDA were restricted to short-term assessments in laboratory settings [9]. The recent advent and wider availability of ambulatory recording devices has made it increasingly feasible to gather EDA longitudinally in daily life, opening the exciting possibility of evaluating unique variance across time-scales and settings. For example, a study investigating panic disorders found that SCL trends in participants with panic disorders were significantly elevated during longer ambulatory recordings than in shorter-term assessments in a laboratory setting [12]. While advances in wireless biosensing have allowed for more studies to be conducted in ambulatory settings, the challenges associated with artifact detection and robust SCR identification have hindered efficient and accurate analyses of these signals [9].

To further the utility of ambulatory EDA data, the current manuscript presents a novel strategy for automatically identifying SCRs and removing artifacts. We present the performance of our methods compared to expert manually labeled EDA data. EDA data used for testing was acquired from 55 healthy participants in a lab setting in response to a standardized set of evocative photos. While we will ultimately apply our novel approach to ambulatory data, using data collected in a lab setting provided two major benefits: 1) using standardized evocative photos as a stimulus is a well-studied and widely used approach to elicit SCRs and 2) expert human coders provided labels, coded from videos, for the responses enabling a ground truth with which to compare our method's performance. Using the expert labels as ground truth, we evaluated our method's accuracy in automatically identifying SCRs compared to a current gold-standard software, Ledalab. We also report the separability between SCR and artifact shape as a first step towards moving our method to ambulatory collected EDA data. Finally, we present the possible directions this work could take in the future work section.

1.1. Current analysis methods

1.1.1. SCR detection

Traditionally, EDA signals are analyzed by hand, and, in fact, the Society for Psychophysiological Research still recommends manual analysis for identifying SCR locations and removing artifacts [11]. However, manual analysis is time-consuming and prone to human error and inconsistency. As a first step towards more automated analysis methods, many groups have developed different models to represent the shape of an SCR. A popular model used in several recent studies is the Bateman equation:

$$b(t) = e^{\frac{-t}{\tau_2}} + e^{\frac{-t}{\tau_1}} \quad (1)$$

In (1), 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, controlled by τ_1 and τ_2 respectively [4]. Because the Bateman equation relies on only two parameters, minimal computation complexity is required to estimate optimal parameters and fit to an SCR, making it ideal for different SCR detection software [10], [13]. Using this model as the basis for an SCR shape, several groups have created software capable of analyzing EDA data and determining the location of SCRs; however, most of these methods were developed for short, laboratory-based studies and have not been optimized for longer ambulatory recordings [14]. Model-based approaches employ psychophysiological assumptions to develop mathematical models describing how an underlying process generates observed data [14]. Two model-based systems currently considered gold-standard for EDA analysis are SCRalyze and Ledalab [15,16]. However, while both systems have been shown to perform well when analyzing EDA signals collected in the lab, they may not perform well with ambulatory signals [15,17,16]. One of the major drawbacks of SCRalyze is that it relies on convolution with a driver function to locate SCRs in the signal before employing probabilistic inversion to estimate the parameters of the SCRs. This convolution and subsequent estimation relies on prior knowledge about the location of a stimulus or event that evoked an SCR [15], [14]. When this prior knowledge is unknown, for instance when EDA is collected outside of a controlled laboratory setting, these systems may not accurately locate SCRs. For further details the reader is referred to the original papers [15,17,18]. Similar to SCRalyze, Ledalab uses the Bateman equation as an impulse response that, when deconvolved with the signal, is used to identify the onsets of individual SCRs. To improve goodness of fit, Ledalab uses gradient descent to optimize the $\tau_1$and τ_2 parameters to better fit SCRs found across the signal [16]. Ledalab is slow due to its optimization process and not robust to artifacts, making it difficult to scale to longer and more artifact-laden ambulatory signals.

Another interesting automated SCR identification approach recently proposed is convex optimization. Convex optimization allows the problem to be solved efficiently using a sparse QP-solver [19]. While the algorithm appears conceptually promising, in-depth quantitative analyses of its performance is currently based on simulated data, while only an observatory analysis is provided for the SCR detection with real data [19]. Because a full quantitative analysis of non-simulated data is not provided, a true comparison between this method and our novel approach is not possible at this time. Additionally, this algorithm only considers noise as iid white Gaussian but does not consider artifacts caused by movement or touching the recording sensor [19]. Not being robust to these types of artifacts could degrade the performance of this algorithm if applied to ambulatory data and make it difficult to successfully scale analysis for ambulatory data.

To move towards a more scalable and robust solution for automatically identifying SCRs in ambulatory EDA signals, sparse recovery methods have been proposed and shown to be a promising solution [20]. In a comparison paper evaluating SCRalyze and Ledalab, it was shown that SCRalyze performed better in event driven analysis [14]. However, as discussed above, SCRalyze does not allow for analysis of signals without labeled events, as Ledalab does. For this reason, we compare our novel approach to Ledalab only.

1.1.2. Sparse recovery methods

Sparse recovery is a technique that can be used to estimate a signal by linearly adding columns from a dictionary of predefined waveforms using $D\gamma = x$ (2). In (2), D is the dictionary, γ is the coefficient matrix that gives the weight of the selected atoms, and x is the original signal [21]. Typical dictionaries are represented using a matrix built up of individual columns, commonly called atoms, each representing a specific waveform. While there are many standard dictionaries with predefined waveforms, most applications build a dictionary containing application-specific atoms to better represent the original signal [22]. Dictionaries are typically designed to be fat matrices, meaning a single exact solution does not exist. Instead, greedy approaches are used which allow Eq. (2) to be solved as an approximation [21]. The most popular greedy approaches fall under the category of Matching Pursuit (MP) algorithms, which follow the same basic algorithm with slight variations. To create an estimate of the original signal, MP algorithms aim to solve the optimization problem with:

$$\underline{\gamma} = \underset{\underline{\gamma}}{\text{Argmin}} \|\underline{\gamma}\|_0 \text{ Subject To} \|x - D\underline{\gamma}\|_2^2 \leq \epsilon \quad (3)$$

In (3), γ is the estimated coefficient vector, x is the original signal, and D is the dictionary. Additionally, $D\gamma$ is an estimated representation of the original signal [22]. As a greedy approach, MP algorithms solve the optimization equation through a series of iterative steps that can be generalized to two main steps: 1) atom selection and 2) residual update. First, the atom with the highest correlation to the current residual error is selected. Then the residual error is updated to reflect the newly selected atom [21]. 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 and the residual update steps. After an atom is selected, the OMP algorithm projects that atom into an orthogonal space; this reduces run time of the algorithm and enforces better sparsity of the estimate.

It has been shown that OMP is a useful MP algorithm for EDA analysis [20], wherein a knowledge-driven dictionary achieves good fit to lab-collected EDA signals with high accuracy [20]. This method also enables the ability to identify SCRs in the signal using the OMP methodology, with a dictionary made up of columns to represent both tonic and phasic components, and some post processing of the selected atoms. While this study showed that OMP with a knowledge-driven dictionary could be efficiently used to model EDA signals and detect SCRs, it did not address: 1) the computational complexity or run time required for the design; 2) use of a data-driven or learned dictionary, although the possibility of using this type of dictionary was discussed; 3) a comparison with existing gold standard software, including fully reporting on the performance measures achieved by the OMP methodology; or 4) artifact detection. Additionally, the study did not evaluate the effect of removing tonic level, which could impact overall performance. Tonic level is much more variable than phasic responses [1], meaning they can make dictionaries more difficult to generalize to new data. They also significantly increase the size of a required dictionary, increasing computational complexity and run time.

In our previous work, we showed that a similar OMP methodology could be extended to an ambulatory EDA signal with high accuracy when compared to labels generated from Ledalab [3]. This previous work was a proof-of-concept test demonstrating that OMP could be successfully extended to ambulatory EDA signals with mean accuracy of 80%, and sensitivity and specificity of 90% and 53%, respectively. Additionally, this previous work showed that OMP significantly reduced run time for analysis compared to Ledalab [3]. With mean recording lengths of a little over 3 h, the proposed method had about an 81% decrease in run time over Ledalab.

1.1.3. Artifact detection

For most studies involving EDA data, artifacts are removed either by applying exponential smoothing or through low-pass filtering [23]. While these approaches often work well for minor artifacts, high magnitude or long duration artifacts are not effectively removed using either of these methods [23]. Other traditional methods for artifact detection require either manual inspection of the data after some processing, which is time consuming and prone to subjective interpretations, or through collection of EDA from multiple sites simultaneously (such as the ankle and wrist), which may not be possible in ambulatory settings [23–25]. As experiments shift to more ambulatory data collection, the potential for high magnitude artifacts increases due to factors including, but not limited to, participant movement and participants or other objects bumping into the sensor. A recent study attempted to address these issues and find more robust ways to identify artifacts using machine learning techniques [23]. It was found that Support Vector Machines (SVM) with a radial basis function was successfully able to distinguish between clean and noisy sections of data with test accuracies of about 96%. While this method addresses the issue of missing high magnitude artifacts, it performs classification by looking at sections of data instead of classifying individual SCRs or artifacts. Potential issues with sectioning the data in this way are three-fold: 1) If there are multiple responses in a section (including SCRs and artifacts) there is no way to classify the responses individually; 2) they can miss responses if an SCR or artifact is compounded falls between two sections (i.e., the onset of an SCR is at the end of one section and recovery is in the beginning of the next), depending on how sectioning is handled; and 3) simply identifying sections of noisy data may still require manual cleaning before further analysis can be completed.

1.2. Our contribution

In previous work, we showed that our methodology could be successfully used to automatically identify SCRs in ambulatory EDA data with low computational complexity. The current work aims to fully present the performance of our methodology when applied to expert human labeled data. Expanding on the methodologies introduced previously [3], our contributions in the present manuscript are four-fold:

1) Demonstrate that an expanded data driven dictionary improves performance of SCR identification over a knowledge-driven dictionary.
2) Investigate the removal of tonic level using different techniques, and compare our estimation to Ledalab's.
3) Report on Ledalab's performance measures – which to our knowledge, has yet to be done – and compare it to our method's performance.
4) Investigate classification accuracy between artifacts and SCRs using the Bateman equation parameters as features

In the present study, we analyze laboratory EDA data from participants viewing a series of standardized and normed evocative images that induce a range of autonomic responses. Using expert human labeled SCRs and artifacts as ground truth, we show that our proposed method, including a tonic estimation that employs a low pass filter with 1 Hz cutoff frequency, allowed us to identify SCRs with an accuracy of about 69%, sensitivity of about 69%, and specificity of about 71%. Additionally, using discriminant analysis classification allowed us to classify artifacts from SCRs with an accuracy of about 74%.

2. Methods

2.1. Experimental data

2.1.1. Participants

Data were collected at Northeastern University (Boston, MA) from 2013 to 2014 and included 73 healthy participants (35 females) recruited from Northeastern University and the Boston area. Seven participants' data were not included in this study, three because data were lost due to technical malfunctions and the remaining 4 stopping the study early due to lack of interest or lack of compliance with study instructions. Out of the remaining 66 participants, EDA data from 11 participants were not annotated due to a computer technical malfunction or human-based computer error. Thus, our final dataset consisted of 55 participants (24 females) with age ranging from $18 - 38$ years ($M \pm SD = 24.3 \pm 5.5$ years).

2.1.2. Procedure

Participants were greeted and consent was obtained in accordance with Northeastern University's Institutional Review Board. Participants completed a health questionnaire asking about their 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 h 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 healthy. Participants' height and weight were measured. 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 (only data captured from this device is analyzed in the present study) 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 connecting to the physiological recording equipment, participants sat quietly for 2–10 min while completing a demographics questionnaire and the PANAS-X questionnaire [26]. Next, participants completed a five-minute baseline period wherein they were asked to sit still. They then completed a heartbeat detection task (data reported previously in [27]).

Next, participants completed a task viewing and providing ratings in response to each of 103 full-color photos from the International Affective Picture System (IAPS), which are commonly used to reliably and validly induce various autonomic responses and affective experiences [28]. Per IAPS instructions [28], 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 A1 in the Appendix A). Photos were separated into an initial "anchor" block of 3 photos, and then 10 additional blocks of 10 photos each: 2 blocks of unpleasant-high arousal, 2 blocks of pleasant-high arousal, 2 blocks of unpleasant-low arousal, 2 blocks of pleasant-low arousal, and 2 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 counterbalanced order, always starting with the anchor block, to familiarize participants with the task [28]. Next, either the unpleasant high arousal block or the pleasant high arousal block was displayed, with subsequent images 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 [29–31].

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 s indicating that the picture was about to be shown. They then they viewed the picture for six seconds and rated their response to the picture in terms of 1) valence using a continuous scale with nine anchor images from the self-assessment manikin scale (SAM [32]); 2) arousal using a continuous scale with nine anchor images 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 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.

2.1.3. Annotating EDA data

A group of four experts (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 s after photo offset (10 s total). Each file was labeled by one of the four trained experts while simultaneously viewing the video to increase the robustness and consistency of the labeling. Additionally, each labeled file was reviewed by author I.K. to verify consistency and robustness of the labels. Each EDA data segment was labeled using one of the following possibilities: 1) only one SCR, which is a biologically induced EDA response of at least 0.01 μS in magnitude from tough to peak, regardless of whether it was caused by the photo; 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 caused by participant movement; 8) an artifact caused by participant movement; 9) no SCR, participant movement, or artifacts; or 10) no data available. Using these annotations, the EDA data was categorized into the following four groups: 1) segments that elicited a clear SCR (cases 1–4 above); 2) no response (case 9); 3) segments that involved a clear EDA artifact (cases 5, 6, and 8); and 4) all other segments (cases 7 and 10). Groups 1 and 2 were used for analysis for SCR identification (Section 3.1), group 3 was used for analysis for artifact identification (Section 3.2), and group 4 was not analyzed further. In total, there were 6175 events. A histogram showing the relative frequency for each group of events is shown in Fig. 2.

2.2. Analysis methods – SCR identification

This section provides a brief overview of our proposed methodology to automatically identify expert labeled SCRs (further detail is provided in Sections 2.2.1–2.2.4), see Fig. 3. For an analysis of

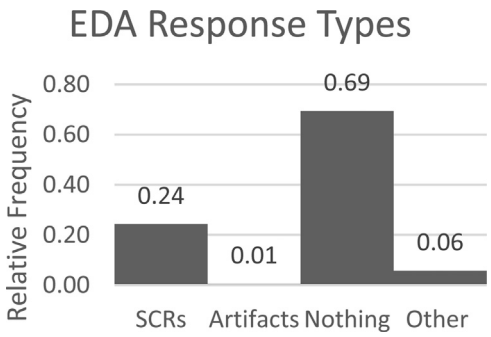

Fig. 2. Relative frequency of responses shown by group.

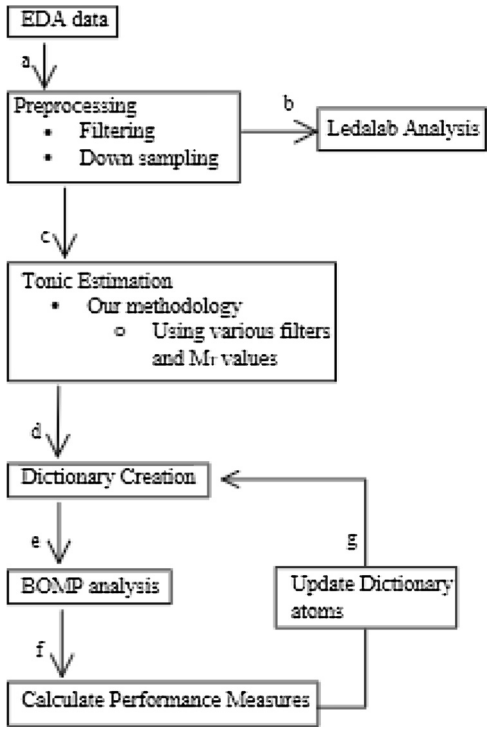

Fig. 3. Flowchart depicting analysis methodology.

EDA signals using the OMP methodology, the participants' EDA signals were initially filtered and downsampled (see below for more details). After filtering and downsampling was complete, tonic and phasic components were estimated using the tonic estimation methodology introduced in our previous paper [3]. Tonic estimates were subtracted from the EDA signal to obtain phasic components. Initially, SCRs were identified from the phasic components using the batch OMP (BOMP) in conjunction with a knowledge-driven dictionary. Corresponding performance measures were calculated and misses in each signal were analyzed to find common trends and expand the knowledge-driven dictionary to a data-driven dictionary. Using the expanded dictionary, SCRs in the phasic components were again identified with the BOMP methodology and new performance measures calculated. The following sections breakdown each aspect of our methodology in further detail.

2.2.1. Preprocessing and tonic estimation

To determine the best functioning filter and downsample unit, signals were analyzed using three different filters and a range of sampling frequencies. All filters used were low pass with cutoff frequencies of 0.35 Hz, 0.5 Hz, and 1 Hz, respectively. These cutoff

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

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

Fig. 4. Template used to create the base dictionary .

frequencies were chosen based on the average rise time of an SCR. The typical rise time is between 1 and 3 s [1], which corresponds to 1 − 0.35 Hz signals. The sampling frequencies, F_s, ranged between 1 and 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 identifying SCRs and significantly increased run time required to complete BOMP analysis.

After preprocessing was complete, tonic and phasic estimations were found for each signal using the methodology introduced in our previous work [3]. However, in our previous work, we focused more heavily on using Ledalab's methodology for estimating tonic and phasic components to facilitate comparison between the OMP methodology and Ledalab. Our tonic estimation methodology was therefore not fully investigated at that time, but is addressed more fully in the current study.

To estimate 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 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 minima were filtered, interpolation through the minima 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. Note that to identify SCRs in phasic components, BOMP fits atoms in the generated dictionary to the phasic component only. Therefore, further reference to estimated signals refer to the fit of the BOMP to the phasic component instead of the full EDA signal.

To determine the optimal M_T, a range of thresholds were tested and corresponding performance evaluated. Using the knowledge that the average length of an SCR is between 3 and 10 s [10], the M_T range investigated was 1–10 s using a step size of 1 s.

2.2.2. Dictionary creation and expansion

In this study, two BOMP analyses were used sequentially to first generate and expand the dictionary and then to calculate 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 our previous study [3], designed as follows. Each individual column (i.e., atom) in the base dictionary represents a single SCR whose shape is defined by the Bateman equation, $\tau_1 = 0.75$ and $\tau_2 = 20$, for a 30-s time window. The SCR onset for the first column was set to equal time 0 and then shifted 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 format of the base dictionary is show in Fig. 4. The initial τ_1 and τ_2 parameters used in the base dictionary were

chosen based on existing literature as reasonable τ_1 and τ_2 values for SCR responses [1]. Additionally, the initial parameters were selected so that $\tau_2 > \tau_1$ to maintain the assumption of non-negative SCR responses in the base dictionary.

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. Again, each new τ_1 and τ_2 pair was constrained to $\tau_2 > \tau_1$ to maintain only non-negative responses in the dictionary. This process returned optimal τ_1 and τ_2 values for each missed SCR. Using the optimal parameters for all 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 used to create the base dictionary.

2.2.3. Batch OMP

As previously mentioned, the specific OMP algorithm used in our analysis is referred to as batch orthogonal matching pursuit (BOMP) and is based on Eq. (4):

$$\underline{\gamma} = \underset{\underline{\gamma}}{\text{Argmin}} \|x - D\underline{\gamma}\|_2^2 \text{ Subject To} \|\underline{\gamma}\|_0 \leq K \tag{4}$$

In (4), γ is the estimated coefficient vector, x is the phasic estimate to be analyzed, D is the dictionary, and K is the sparsity constraint. Like any OMP algorithm, BOMP uses a greedy approach with the two main steps introduced earlier. To further improve upon the traditional OMP algorithm, the BOMP algorithm reduces computational complexity by introducing a Cholesky factorization [22]. OMP algorithms orthogonalize each selected atom, which introduces a matrix inversion at each iteration. The orthogonalization step, with the matrix inversion, is shown in Eq. (5):

$$\underline{\gamma} = (\mathrm{D}_I)^{+} x$$

$$= (\mathrm{D}_I^T \mathrm{D}_I)^{-1} \mathrm{D}_I^T x \tag{5}$$

From equation (5) 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 matrix [22]. To improve performance and decrease computational complexity, Cholesky factorization was used to only require 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 [22]. This leads to less computational complexity and faster BOMP algorithm (summarized in Table 1). Another difference between traditional OMP methods and BOMP is that BOMP uses sparsity as the stopping criteria for the iterations, which further enforces sparsity on the estimates. However, using sparsity as the stopping criteria makes K into a system parameter that needs to be optimized for each signal. To accomplish this, the BOMP analysis was run over a range of K values, 10–400 for each signal, and the best value was chosen based on the calculated performance measures.

2.2.4. OMP thresholding and performance measure calculations

Recall from (4), the output of the BOMP analysis is a matrix (γ) representing 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 is needed to remove values that do not align with the working knowledge of SCR shapes. For example, to maintain non-negative SCR responses, any negative γ values were removed as they correspond to convex inflections in the estimate, representing negative responses.

Table 1
Pseudocode for the BOMP algorithm used in our analysis.

1	Input: Dictionary $\boldsymbol{D}$, signal x, target sparsity K
2	Initialize: Set $I := (), \boldsymbol{L} := [1], r := x, \gamma := 0, \alpha := \boldsymbol{D}^T x, n := 1$
3	**while** sparsity $< K$
4	$\hat{k} := argmax\lvert d_k^T r\rvert$
5	**if** $n > 1$
6	$w := \text{Sovlefor} w \left\{ \boldsymbol{L}w = \boldsymbol{D}_I^T d_{\hat{k}} \right\}$
7	$\boldsymbol{L} := \begin{bmatrix} \boldsymbol{L} & 0 \\ w^T & \sqrt{1 - w^T w} \end{bmatrix}$
8	**end**
9	$I := \left(I, \hat{k}\right)$
10	$\gamma_I := \text{solveforc} \left\{ \boldsymbol{L}\boldsymbol{L}^T c = \alpha_I \right\}$
11	$r = x - \boldsymbol{D}_I \gamma_I$
12	$n = n + 1$
13	**end**
14	Output: Sparse representation γ such that $x \approx \boldsymbol{D}\gamma$

To facilitate the comparison between our methodology and Ledalab, average performance measures, including accuracy, sensitivity, specificity, F1 score, and true SCR identification percentage (number of identified SCRs divided by number of labeled SCRs), are reported over a set of labeled EDA data (see section 2.1 for details on the EDA data labeling). Using human annotated data as the ground truth, performance was evaluated using time ranges corresponding to each picture being shown. True positives were counted if the number of responses labeled for a specific picture were correctly identified in the estimate. A true negative was counted if neither the labels nor estimates contained any responses. False negatives and false positives were counted if any missed or extra responses were found.

Each performance measure was then statistically assessed, first using a two-sided Wilcoxon ranksum test, and if a statistical difference was found, then a one-sided Wilcoxon ranksum test [33,34]. Using a combination of the two-side and one-side ranksum test allowed us to determine statistical significance of the differences seen between BOMP and Ledalab's returned performance, and to determine if the performance was statistically the same, or if the populations were not the same, if our method or Ledalab's had a statistically higher median. The ranksum test is a nonparametric rank test that assesses equality of the median for two populations. For the two-sided ranksum test, the null hypothesis assumes that the medians of the two populations are the same while the alternative hypothesis assumes that the two population medians are different. For the one-sided ranksum test, the null hypothesis assumes that the median of population x is less than or equal to the median of population y, while the alternative hypothesis assumes that the median of population x is greater than the median of population y [33,34]. We chose to use the one-side test as a secondary evaluation to determine which population had a greater median value as opposed to simply determining if the populations were statistically different.

2.3. Analysis methods – artifact identification & classification

To investigate the ability to automatically distinguish between artifacts and SCRs, several artifacts, previously identified by expert human raters, 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 SVM and discriminate analysis classification. For SVM classification, linear and radial

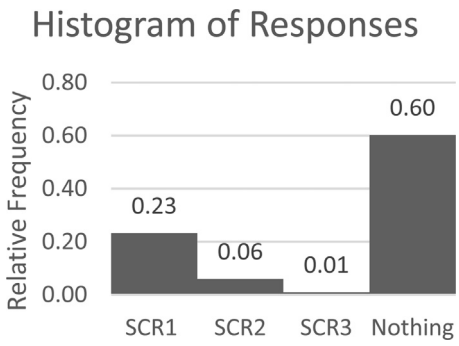

Fig.5. Histogram showing the relative frequencies of each response type analyzed in the test set (N = 10 participants, 1030 events).

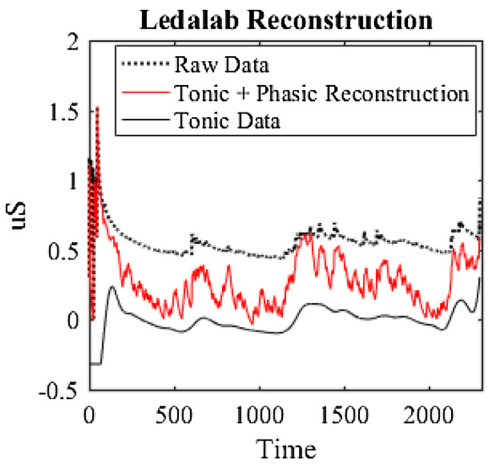

Fig. 6. Ledalab's tonic estimation shown with the raw data and signal reconstruction.

basis function classification schemes were investigated. 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.

3. Numerical experiment and results

3.1. SCR identification

To limit overfitting, the data used was split into a test set and a validation set. Most of our analysis was done using the test set, which was made up of 10 of the original 55 participants. The 10 test participants were randomly selected with no prior knowledge of the participants' EDA data being used. The remaining 45 participants formed the validation set. Sections 3.1.1 and 3.1.2 present results obtained using only the test participants. Section 3.1.3 presents a comparison of our methodology to Ledalab, first using the test participants and then using the validation participants.

The 10 selected test participants consisted of 4 females and 6 males with ages ranging between 18 and 36 years ($M \pm SD = 22.8 \pm 5.5$ years). Using 10 participants led to 1030 total events which, based on the received annotations, had 239 single SCRs (case 1), 61 events that had 2 SCRs totaling 122 SCRs (case 2), 9 events which had 3 SCRs totaling 27 SCRs (case 3), no events which had 4 SCRs (case 4), and 620 events with no response (case 9). This gave us a total of 442 total SCRs and 620 events with no response. Events labeled as artifacts, had no data, or fell into another category were not considered for this section of the study (cases 5–8, and 10). Fig. 5 shows the histogram of the frequency of the different cases analyzed.

3.1.1. Tonic and phasic estimation

To determine a robust way to estimate tonic and phasic components, we first compared performance of our tonic estimation procedure with Ledalab's. 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 0.35 Hz filter and M_T equal to 10 s. Fig. 6 shows the original signal, Ledalab's tonic estimation, and the tonic plus phasic reconstruction from Ledalab. As can be seen from the reconstructed signal (red curve), much of it is below the original signal, and therefore does not produce a good fit. This suggests that information from the original signal has been lost during LedaLab's 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 reconstruction fit for some signals, two or more optimizations are required, which is time consuming and requires trial and error to determine an 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

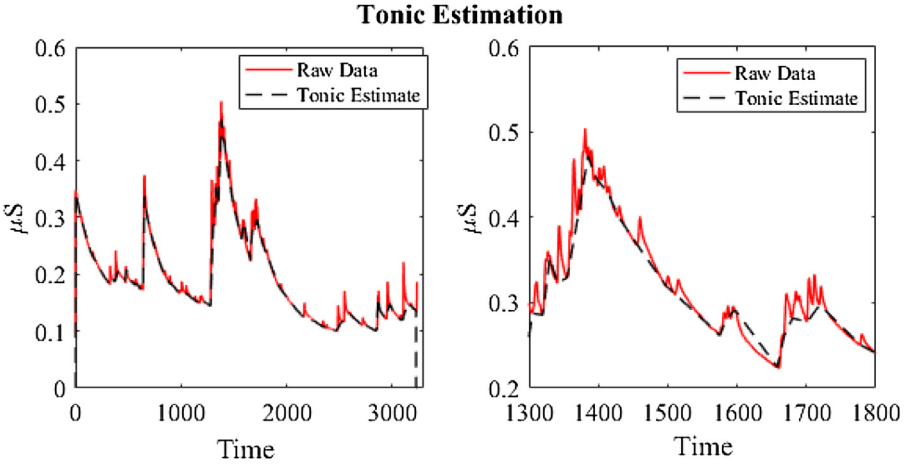

Fig. 7. Proposed tonic estimation methodology.

Table 2
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

27.83% of possible SCRs in the signals were detected. In contrast, our tonic estimation, shown in Fig. 7, was based on the minima throughout the signal, so the reconstructed signal follows the original signal more closely and minimal information is lost. The response was 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. Therefore, it was possible to get negative responses in the estimated phasic component. Negative responses violate the previously made assumption of non-negativity in EDA signals, and are not possible to achieve naturally. Fig. 7 shows an example of a signal where the raw data is below the tonic estimation, meaning when this tonic is subtracted from the raw data it produces a negative response in the phasic component.

To find a more robust way to estimate the tonic level and reduce negative responses in our methodology, several filtering and thresholding values were examined and compared. To determine performance of each tonic estimation method, SCRs were identified in each estimated phasic component using BOMP, and then average accuracy, sensitivity, specificity, and true SCR identification percentage (TIP) was computed. These computed performance measures were then used to determine which method yielded the best detection ability. Table 2 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 1 Hz low pass filter and an $M_T = 1$ s, producing an average accuracy of 68.63% and a TIP of 66.29%.

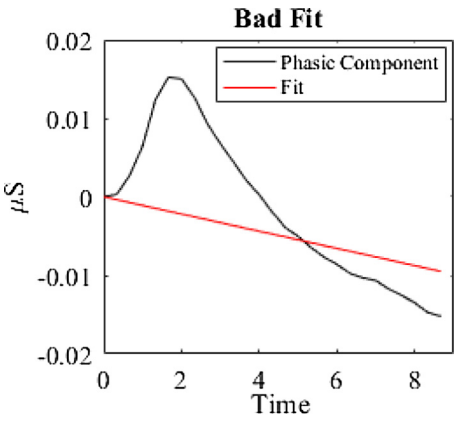

Fig. 8. Bad estimated fit generated from missed SCRs.

In our previous work, the run time for Ledalab and our methodology was compared, and it was shown that the BOMP methodology significantly improved run time [3]. In this study, the filter used to create tonic and phasic estimates slowed run time. We investigated the performance of the tonic estimation by first assessing the 0.35 Hz filter. This filter was the same as that used in the previous study, and allowed us to use a sampling frequency of 1 Hz. As in the last study, the present study found that using the 0.35 Hz filter, run with files whose average length was 46.41 min ± 7.85 min, with the BOMP methodology was faster than Ledalab's run time, regardless of the M_T value. However, moving to the 1 Hz filter with the same files increased 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 work that Ledalab's run time increased dramatically more than the BOMP method as the length of the signal increased [3]. 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 needed to get an accurate picture of run time in both systems. As the best overall performance measures were achieved using the 1 Hz filter with $M_T = 1$ s, phasic estimation for all further models used this parameter set.

3.1.2. Dictionary expansion

Using our tonic estimation method with the base dictionary, we were only able to achieve an accuracy of 52.10%. To improve this 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

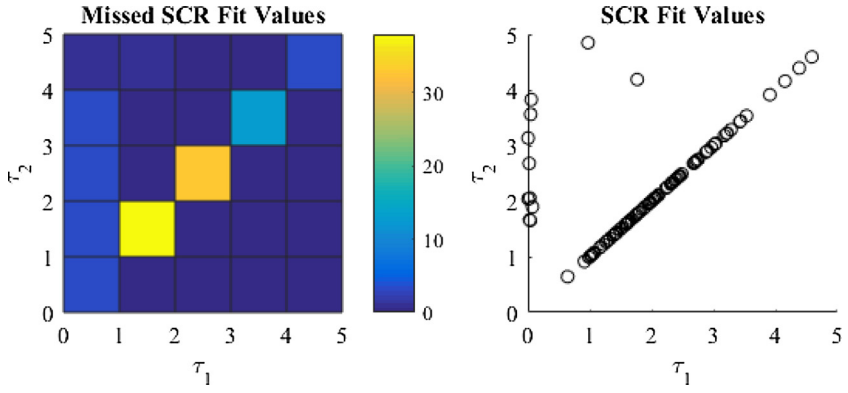

Fig. 9. (left) Histogram of τ_1 and τ_2 values returned from the estimated fits. (right) scatter plot of the same values.

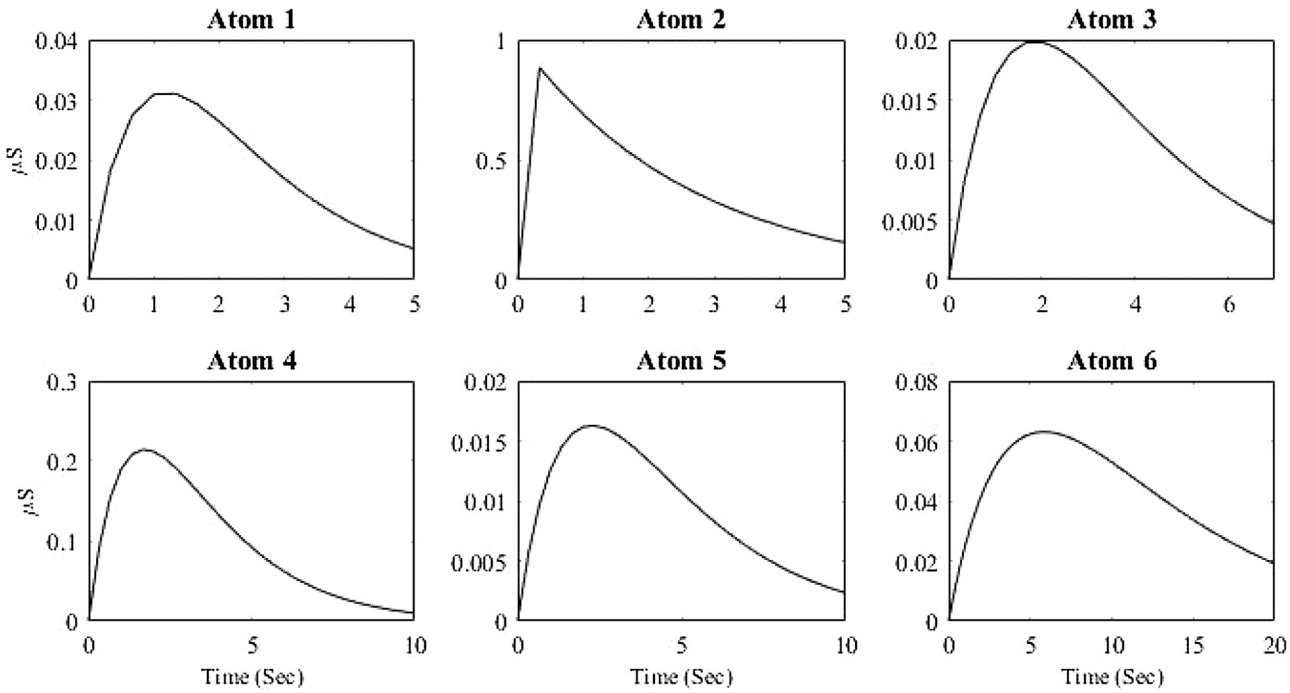

Fig. 10. New SCR shapes used to expand the initial dictionary.

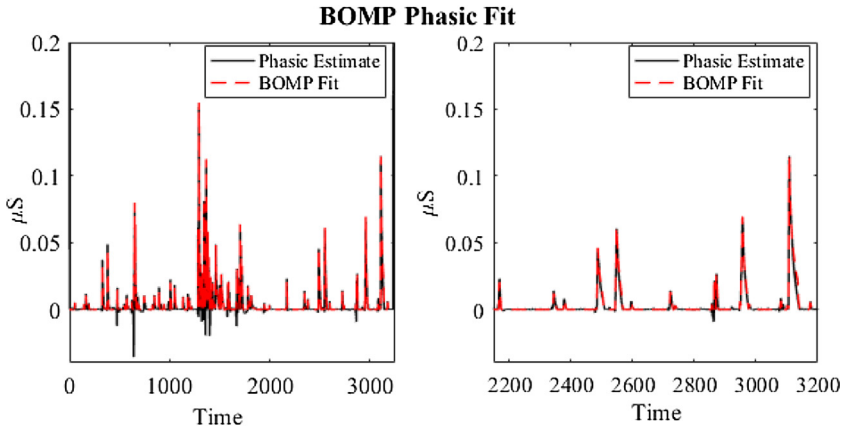

Fig. 11. Phasic estimate created using our methodology with BOMP fit overlaid.

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 was not reached in 100,000.000 iterations, or through visual inspection. Fig. 8 shows an example of a bad fit 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 Fig. 8. Shapes with fast rise and recovery times are more akin to noise than to true SCRs (Taylor, Jaques, Chen, Fedor, Sano, & Picard, 2015), and were therefore ignored. After the bad fits were determined and removed, there were approximately 100 $\tau_1$and τ_2 pairs left. They are plotted using a bivariate histogram, shown in Fig. 9 (left), to determine common values. The histogram shows 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 values were significantly larger than the τ_1 values. These trends match the previously made non-negativity assumptions and fit with the constrain of $\tau_2 > \tau_1$ required for the dictionary. Plotting the $\tau_1$and τ_2 values in a scatter plot, Fig. 9 (right), suggests three or four major clusters existing within these two trends.

Using these clusters and the most common bins produced from the histogram, six 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. Fig. 10 shows the shape of the additional atoms added to the dictionary. While determining these additional shapes, we found that the general shape of each new atom was relatively the same, and the major difference between the new shapes was 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 s, which matches previously reported ranges of SCR lengths [6]. This new dictionary was thus used to re-identify SCRs in each test signal with significantly improved performance over the initial accuracy reported above (see section 3.1.3 for details).

Table 3
Performance by participant for our methodology (left) and Ledalab's analysis (right). Note: table refers to the participants in the test set.

File	Sparsity	Accuracy	Sensitivity	Specificity	F1 Score	File	Accuracy	Sensitivity	Specificity	F1 Score
1	400	71.11	77.27	69.12	56.67	1	77.27	50.00	86.36	52.38
2	200	91.43	85.71	92.31	72.73	2	87.25	7.14	100.00	13.33
3	400	55.74	51.39	62.00	57.81	3	71.19	62.50	84.78	72.58
4	400	66.67	77.03	50.00	74.03	4	47.66	32.43	81.82	46.15
5	400	63.33	41.67	71.21	37.74	5	72.41	0.00	100.00	0.00
6	350	60.26	54.24	78.95	67.37	6	53.09	44.07	77.27	57.78
7	300	74.34	84.91	65.00	75.63	7	77.78	94.34	64.06	79.37
8	350	66.02	88.24	61.63	46.15	8	88.00	64.71	92.77	64.71
9	350	68.69	60.87	86.67	73.04	9	45.45	26.09	90.00	40.00
10	350	70.37	68.42	71.43	61.90	10	63.46	33.33	81.54	40.63
Avg	**-**	**68.79**	**68.97**	**70.83**	**62.31**	**Avg**	**68.36**	**41.46**	**85.86**	**46.69**

3.1.3. BOMP SCR identification vs. ledalab SCR identification

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, F1 score, 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.

Fig. 11 shows the BOMP fit based on our phasic estimation that, through visual inspection, suggests that BOMP captures overall peaks and trends of the original signal. To give a clearer view of the original signal and fit, Fig. 11 shows the estimate zoomed in between 2150 and 3200 samples. BOMP, with post processing of the estimated gamma values, identified SCRs with an average accuracy of 68.79% ± 9.64% (M ± SD) compared to a Ledalab accuracy of 68.36% ± 15.44%. While the average accuracy for both systems was statistically the same ($p = 0.7337$), Ledalab showed significantly higher variability across participants, suggesting that the BOMP method may better generalize across different participants. Looking at sensitivity, 68.97% ± 16.25% for BOMP versus 41.46% ± 28.17% for Ledalab, shows a statistically significant increase in BOMP over Ledalab ($p = 0.0257$). Additionally, of the 442 labeled SCRs across the 10 test participants, BOMP successfully located 66.29% of the SCRs while Ledalab only located 45.02%. While the BOMP method does a superior job detecting SCRs, specificity favors Ledalab over the BOMP method. Specificity for Ledalab, 85.86% ± 10.80%, is statistically better than the BOMP method, 70.83% ± 12.53% ($p = 0.0211$). Finally, to consider both false negatives and false positives, F1 scores were calculated. F1 scores showed a benefit towards the BOMP analysis (62.31% ± 12.83%) over Ledalab's method (46.69% ± 24.93%). While the accuracy is of our novel approach is the same as Ledalab, given the improvements seen in sensitivity and F1 scores, we feel that our novel approach performs better overall and will likely yield better performance when applied to the detection of SCRs in ambulatory data. Additionally, the high number of false positives seen in the BOMP, which causes the low specificity, could be reduced through further optimization of the dictionary, and inclusion of artifact elements in the dictionary (please see section 3.3 for full discussion of method improvements). Table 3 shows the performance measures calculated for each of the 10 test participants and the averages for both Ledalab and our novel approach.

We noted in section 2.2.3 that the sparsity system parameter, K, was varied between 10 and 400 to find the optimal value. As can be seen from Table 3 (left), the ideal K was between 200 and 400 for each participant. The variability seen appears to be due to the role that sparsity plays in goodness of fit and subsequently the calculated performance measures. With the lowest sparsity, specificity will be the maximum achievable for the participant, and show a decrease as sparsity increases. Inversely, sensitivity is low with low sparsity, and increases as sparsity increases. Fig. 12 shows this general trend for a single participant. This suggests that by adjusting sparsity, sensitivity and specificity could be adjusted. For the above results, we picked the best sparsity value to maximize and balance both sensitivity and specificity.

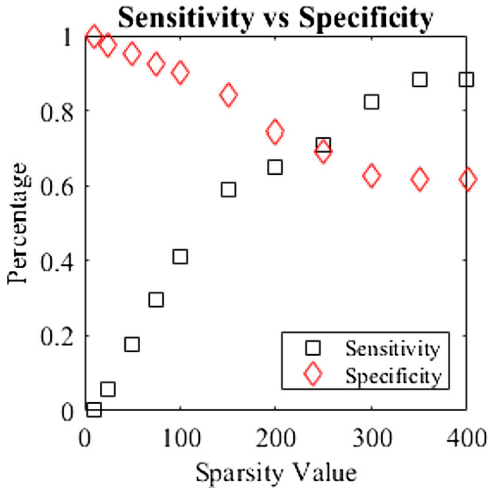

Fig. 12. Sparsity vs average performance.

Table 4
Average performance found using the validation participants.

Method	Accuracy	Sensitivity	Specificity	F1 Score
Ledalab analysis	75.58%	62.21%	73.38%	56.62%
BOMP analysis	71.27%	74.93%	60.71%	59.07%

To show the ability of our method to be generalized, our full method was applied to the validation signals set-aside at the beginning of the analysis. Table 4 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 that found in the test set. Additionally, the ranksum test was again used to determine statistical similarity between the test set and the validation results. It was found that all 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.

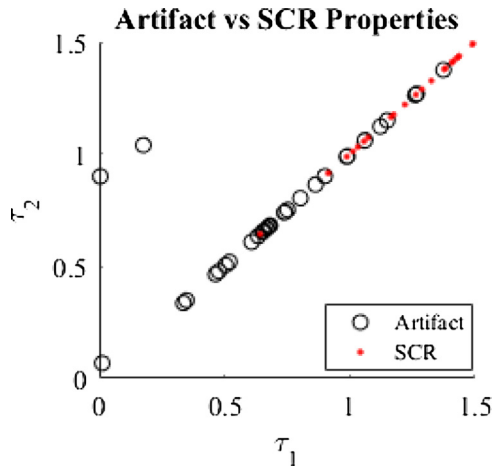

Fig. 13. Parameters required to fit to SCRs and artifacts.

3.2. Artifact detection

As seen in Fig. 2, artifacts occurred in the data set very infrequently. Therefore, there were only a few labeled artifacts throughout the data (33 events that had artifact labels, cases 5,6 and 8). These labeled artifacts were spread across 11 of the participants. The 11 participants consisted of 5 females and 6 males with ages ranging between 18 and 38 years ($M \pm SD = 24.4 \pm 6.8$ years).

3.2.1. Artifact and SCR separability

To determine separability between artifacts and SCRs, the parameters returned from the Bateman equation were used as features to complete classification with SVM and discriminant analysis. Quadratic discriminant analysis (QDA) produced the best classification, showing accuracies of 73.66%.

Using the 11 participants introduced above led to approximately 50 τ_1 and τ_2 parameters returned after fits were found for each labeled 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, Fig. 13 shows that most τ_1 and τ_2 values required to fit to artifact fall below 1, while most τ_1 and τ_2 values that fit to SCRs fall above 1. This difference shows that $\tau_1$and τ_2 parameters for artifacts and SCRs fall into different regions of the feature space with minimal overlap, suggesting that separability may exist between the two shapes. This was supported by the classification achieved using QDA. QDA yielded a classification accuracy of 73.66%, 81.25% sensitivity for artifact classification, and 71.34% sensitivity towards SCR classification. Being able to successfully classify artifacts versus true SCRs will allow for the addition of artifact columns in the dictionary in future work. This dictionary expansion may enable our novel method additional advantages over Ledalab and many of the other current EDA analysis systems. Ledalab does not include the ability to handle artifacts and they suggest artifacts be removed before the software is used [13,16], which is impractical in long-term naturalistic studies. Adding artifact elements into the dictionary would allow our novel method to automatically identify artifacts as well as SCRs so that signals do not need to be artifact-free before analysis is completed [35].

3.3. Future work

Our expanded data driven dictionary significantly improves upon our previously introduced methodology, but leaves several outstanding issues to be addressed. Future work is needed to further improve the robustness of EDA analysis and to fully test the performance of our system when applied to ambulatory EDA data. The two major outstanding areas for improvement using our methodology (further expanded upon below) are to better balance false positives and true positives and to improve upon automated artifact detection. Additionally, this study used lab-collected data for analysis since it provided us with SCRs elicited using a standard method and expert labels as ground truth. To fully address the performance of our system towards ambulatory data further, tests are needed that use expert-labeled ambulatory data.

Approximately 17% of SCRs missed after our best run were caused either by 1) 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 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 it could have increased 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 dataset used in the current manuscript. However, using data collected with a rapid stimuli paradigm 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 SCR responses before the previous response could fully recover [10], thereby increasing the presence 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 one second between chosen dictionary atoms, based on prior knowledge, in favor of reducing double counting SCRs. However, this thresholding limited our ability to fully capture compound SCRs that were elicited less than a second apart. Further analysis of typical compound SCR parameters, as mentioned above, could allow us to balance capturing all SCRs while avoiding additional false positives.

The second issue remaining is to include artifact atoms into our dictionary, as it has been shown in this work 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 testing and developing artifact atoms [35].

Finally, we would plan to further test our system's ability to be generalized to different populations, contexts, and recording devices. Running our analyses on our validation set suggested our methodology would be generalizable, but since this data was collected using the same methodology and recording device as our test set, further analyses are required. We therefore plan to test our methodology on a variety of different EDA signals collected in different situations from different populations using different methodologies [35].

4. Conclusion

The recent ability to collect EDA ambulatorily has allowed many studies to be expanded, with data collected for longer time periods and in a variety of settings [5,9], [25]. However, this new data col-

lection methodology introduces several challenges for processing and analyzing EDA signals. Principal among these issues is the need for a robust way to accurately and automatically process EDA signals, removing the need for manual work, which is error-prone and laborious in very long recordings. While the end goal of our work is toward scalable and accurate analysis of ambulatory data, the present study relied on lab-collected EDA data for two reasons: 1) the stimuli used are well-studied and effective for eliciting SCRs and 2) expert human raters labeled responses that served as the ground truth. Using this data, we were able to develop a data-driven dictionary that, when used with an OMP algorithm, detected labeled SCRs with an average accuracy of 68.80%, sensitivity of 70.83%, specificity of 69.00%, and F1 score of 62.31%, with an ability to positively detect 66.29% of all pre-labeled SCRs in the test signals. This is a significant improvement in sensitivity and F1 score 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 phasic components of the signal, removing the need for tonic atoms in the dictionary. Additionally, we showed the ability of 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, which again is a significant improvement over current methods. 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, we show that, using only τ_1 and τ_2 values returned from the Bateman equation to describe both artifacts and SCRs, classification accuracies of 73.66% can be achieved using a simple QDA analysis. Using QDA to show separability of artifacts and SCRs suggests that artifacts could be included in the dictionary without compromising the ability to detect SCRs, therefore allowing our method to identify both SCRs and artifacts. In conclusion, our methodology appears to provide significant improvement in SCR identification over currently available methods and, given the separability found between SCRs and artifact, will be extendable to automated artifact identification in the future.

Acknowledgments

This research was 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. Murat Akcakaya was supported by Air Force Office of Scientific Research (AFOSR) under award number FA9550-16-1-0386.

Appendix A.

Table A1
Pictures presented from the International Affective Picture System (IAPS).

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

Note. Picture order was randomized within each block. Block order was counter balanced across participants

References

[1] J.T. Cacioppo, L.G. Tassinary, G. Berntson, Handbook of Psychophysiology, Cambridge Up, Cambridge, 2007.
[2] M.S. Goodwin, W. Velicer, S. Intille, Telemetric monitoring in the behavior sciences, Behav. Res. Methods 40 (2008) 328–341.
[3] M. Kelsey, A. Dallal, S. Eldeeb, M. Akcakaya, I. Kleckner, C. Gerard, K.S. Quigley, M.S. Goodwin, Dictionary Learning and Sparse Recovery for Electrodermal Activity Analysis, SPIE Commercial Scientific Sensing and Imaging, Baltimore, 2016.
[4] C.L. Lim, C. Rennie, R.J. Barry, H. Bahramali, I. Lazzaro, B. Manor, E. Gordon, Decomposing skin conductance into tonic and phasic components, Int. J. Psychophysiol. (1997) 97–109.
[5] C. Kappeler-Setz, F. Gravenhorst, J. Schumm, B. Arnrich, G. Troster, Towards long term monitoring of electrodermal activity in daily life, Pers. Ubiquitous Comput. 17 (2) (2013) 261–271.
[6] W. Bourcsein, Electrodermal Activity, second edition, Springer, New York, 2012.
[7] M.M. Bradley, P.J. Lang, Motivation and Emotion, in Handbook of Psychophysiology, 2nd edition, Cambridge University Press, New York, 2006, pp. 581–607.
[8] A. Sano, R.W. Picard, Stress recognition using wearable sensors and mobile phones, Humaine Association Conference on Affective Computing and Intelligent Interaction (IEEE) (2013).
[9] S. Doberenz, W.T. Roth, E. Wolburgh, N.I. Maslowski, S. Kim, Methodological considerations in ambulatory skin conductance monitoring, Int. J. Psychophysiol. 80 (2) (2011) 87–95.
[10] D.M. Alexander, C. Trengrove, P. Johnston, T. Cooper, J. August, E. Gordon, Separating individual skin conductance responses in a short interstimulus-interval paradigm, J. Neurosci. Methods 146 (1) (2005) 116–123.
[11] D.C. Fowles, M.J. Christie, R. Edelberg, W.W. Grings, D.T. Lykken, P.H. Venables, Publication recommendations for electrodermal measurements, Psychophysiology 49 (2012) 1017–1034.

[12] R. Hoehn-Saric, D.R. McLeod, F. Funderburk, P. Kowalski, Somatic symptoms and physiologic responses in generalized anxiety disorder and panic disorder, Arch. Gen. Psychiatry 61 (2004).
[13] M. Benedek, C. Kaernback, A continuous measure of phasic electrodermal activity, J. Neurosci. Methods 190 (1) (2010) 80–91.
[14] D.R. Bach, A Head-to-head comparison of SCRalyze and Ledalab Two model-based methods for skin conductance Analysis, Biol. Psychol. 103 (2014) 63–68.
[15] D.R. Bach, G. Flandin, K.J. Friston, R.J. Dolan, Time-series analysis for rapid event-related skin conductance analysis, J. Neurosci. Methods 184 (2) (2009) 224–234.
[16] M. Benedek, C. Kaernbach, Decomposition of skin conductance data by means of nonnegative deconvolution, Psychophsiology 47 (4) (2010) 647–658.
[17] D.R. Bach, G. Flandin, K.J. Friston, R.J. Dolan, Modelling event-related skin conductance responses, Int. J. Psychophysiol. 75 (no. 3) (2010) 349–356.
[18] D.R. Back, K.J. Friston, Model-based analysis of skin conductance responses: towards causal models in psychophysiology, Psychophysiology 50 (2013) 15–22.
[19] A. Greco, G. Valenza, A. Lanata, E.P. Scilingo, L. Citi, cvxEDA: a convex optimization approach to electrodermal activity processing, IEEE Trans. Biomed. Eng. 63 (4) (2016) 797–804.
[20] T. Chaspari, A. Tsiartas, L.I. Stein, S.A. Cermak, S.S. Narayanan, Sparse representation of electrodermal activity with knowledge-Driven dictionaries, IEEE Trans. Biomed. Eng. 62 (3) (2015) 960–971.
[21] G. Rath, A. Sahoo, A comparative study of some greedy pursuit algorithms for sparse approximation, Signal Processing Conference (IEEE) (2009).
[22] R. Rubinstein, M. Zibulevsky, M. Elad, Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit, CS Tech. 40 (8) (2008) 1–15.
[23] S. Taylor, N. Jaques, W.C. Chen, S. Fedor, A. Sano, R. Picard, Automatic identification of artifacts in electrodermal activity data, Engineering in Medicine and Biology Society (IEEE) (2015).
[24] H. Storm, A. Fremming, S. Odegaard, O.G. Martinsen, L. Morkid, The development of a softwawre program for analyzing spontaneous and externally elicited skin conductance changes in infants and adults, Clin. Neurophysiol. 111 (10) (2000) 1889–1898.
[25] R. Kocielnik, N. Sidorova, F.M. Maggi, M. Ouwerkerk, J.H. Westerink, Smart technologies for long-Term stress monitoring at work, IEEE International Symposiumon Computer-Based Medical Systems (2013).
[26] D. Eatson, L.A. Clark, A. Tellegen, Development and validation of brief measures of positive and negative affect: the PANAS scales, J. Pers. Soc. Psychol. 54 (6) (1988) 1063–1070.
[27] I.R. Kleckner, J.B. Wormwood, W.K. Simmons, L.F. Barrett, K.S. Quigley, Methodological recommendations for a hearbeat detection-based measure of interoceptive sensitivity, Psychophysiology 52 (11) (2015) 1432–1440.
[28] P.J. Lang, M.M. Bradley, B.N. Cuthbert, International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual, University of Florida, Gainesville, Fl, 2008.
[29] D.H. Brainard, The Psychophysics Toolbox 10, Spatial Vision, 1997.
[30] M. Kleiner, D.P.D. Brainard, What's New in Psychtoolbox-3 Perception, 2007.
[31] D.G. Pelli, The VideoToolbox Software for Visual Psychhophysics: Transforming Numbers into Movies, Spatial Vision, 1997.
[32] M.M. Bradely, P.J. Lang, Measuring emotion: the self-Assessment manikin and the semantic differential, J. Behav. Ther. Exp. Psychiatry 25 (1) (1994) 49–59.
[33] J.D. Gibbons, S. Chakraborti, Nonparametric Statistical Inference, in International Encyclopedia of Statistical Science, Springer, Berlin Heidelberg, 2011, pp. 977–979.
[34] M. Hollander, D.A. Wolfe, in: N.J. Hoboken (Ed.), Nonparametric Statistical Methods, John Wiley & Sons Inc, 1999.
[35] M. Kelsey, R.V. Palumbo, A. Urbaneja, M. Akcakaya, J. Huang, I.R. Kleckner, L.F. Barrett, K.S. Quigley, E. Sejdic, M.S. Goodwin, Artifact Detection in Electrodermal Activity Using Sparse Recovery, SPIE, Anehiem, 2017.