AUTOMATED SELECTION OF CHANGEPOINTS USING EMPIRICAL P-VALUES AND TRIMMING (ASCEPT)

A PREPRINT

Matthew Quinn
Department of Biostatistics
Harvard T.H. Chan School of Public Health
mjq522@g.harvard.edu

Kimberly Glass
Channing Division of Network Medicine
Brigham and Women's Hospital
Harvard Medical School
kimberly.glass@channing.harvard.edu

October 31, 2020

ABSTRACT

Mobile health research, using devices produced by manufacturers such as Fitbit and Garmin, has been a growing field in recent years. One of the challenges the field faces are sudden changes in proprietary algorithms that can alter how various data are recorded over time. Multiple approaches exist for the offline detection of these changepoints in time series, but they typically require a pre-specification of the number of changepoints or a non-intuitive parameter such as the penalty in an optimization problem. In this paper, we overcome this by proposing a novel approach for the selection of an optimal set of changepoints among those found through changepoint detection algorithms. The method's first stage involves sequential iterations of a changepoint detection algorithm in order to identify the largest statistically significant set of changepoints. The method's second stage involves trimming false positives within linear trends and seasonal patterns. Each stage uses easily understood parameters. We demonstrate the utility of the method both on simulated data and real mobile health data collected through the Precision VISSTA mobile health study.

***K*eywords** Mobile health · Changepoint selection · Offline changepoint detection · Empirical p-values · Trimming · Regression · Time series

1 Introduction

In recent years, mobile health has taken on a growing importance in medicine and public health, among other fields [1, 2, 3]. The data collected through mobile devices can often be interpreted as time series with variables, such as heart rate and number of steps, recorded at regular intervals (e.g. hourly, daily, etc.). Studying these time series can lend important insights in how health changes over time. In particular, sudden changes in these data may be due to an external factor that alters an individual's behavior. For example, an individual who was previously an avid walker might effectively stop walking after a leg injury. The times at which such sudden changes in the distribution of the data occur are called "changepoints". Unfortunately, mHealth data are also subject to technological artifacts, such as firmware updates and glitches, which are introduced by the devices themselves. These can be very difficult to distinguish from behaviorally-driven changes, obscuring patterns of interest. It is necessary to identify and correct for these technological changepoints before proceeding with downstream analysis.

While an investigator could potentially monitor a device manufacturer's release notes to determine when updates are pushed or manually inspect every variable for every different device to ascertain where changepoints are located, it is difficult to scale monitoring of these types of updates across a wide range of manufacturers and the various devices each one produces. Some devices produced by the same manufacturer (e.g. Fitbit) may have an algorithm change while others do not. Additionally, these updates are often not publicized and some require the user to update the associated smartphone apps to implement the changes. Furthermore, what one investigator considers a significant changepoint may be defined differently by another.

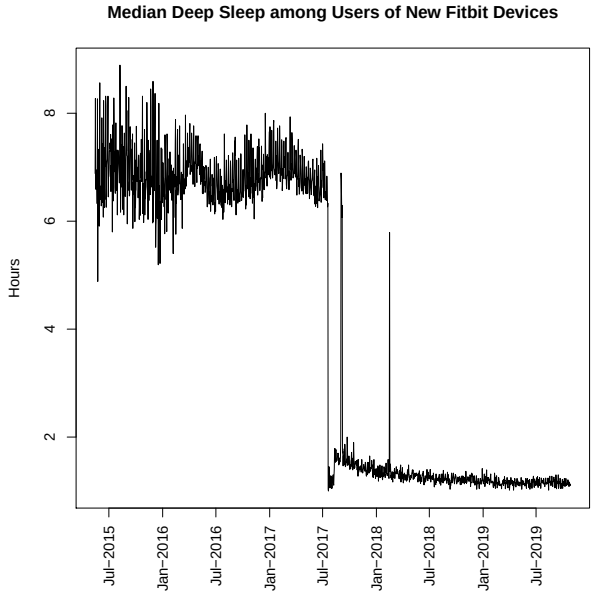


Figure 1: The median deep sleep among participants in the Precision VISSTA study who use a relatively new Fitbit device (i.e. the Alta HR, Blaze, Charge 2, Charge 3, Inspire HR, Ionic, Versa, or Versa 2), multiple Fitbit devices, or an unknown Fitbit device.

To address these issues, we propose a procedure, called Automated Selection of Changepoints using Empirical P-values and Trimming (ASCEPT), as a rigorous method to identify changepoints in mHealth data. We specifically consider offline analysis, which is performed on a fixed data set. This contrasts with online analysis, which is performed on streaming data and entails identifying changepoints in real time. Additionally, we distinguish our approach for offline changepoint "selection" from those that already exist for offline changepoint "detection". That is, there already exist multiple approaches that will take in a series of data and detect changepoints by solving an optimization problem. We consider the current state-of-the-art detection approach to be Pruned Exact Linear Time (PELT) [4]. However, using PELT generally entails having to specify a relatively non-intuitive optimization penalty, which is difficult to do in practice. Changepoints for a Range of PenaltieS (CROPS) [5] allows one to efficiently run PELT under various penalties, but still does not select a final optimal set of changepoints. Rather, it presents the results for each of the multiple runs of PELT. Thus, instead of proposing another method for changepoint "detection", we propose ASCEPT as a form of changepoint "selection", which will consider the results from multiple runs of PELT and yield a single, final optimal set of changepoints. For a more detailed review of offline changepoint detection and the specific motivation for ASCEPT, please refer to Section A of the Supplemental Material.

We apply this method to mHealth data from the Precision VISSTA study on individuals with inflammatory bowel disease (IBD) [6], which has clear evidence of technological changepoints. The study collects data on individuals' sleep and daily activity habits, among other characteristics. In Figure 1, we present one such variable from the Precision VISSTA study: the daily median amount of deep sleep across a subset of individuals using Fitbit devices. We focus on a single time series of the median, rather than studying many individual time series, to make the challenge of differentiating between behaviorally-driven and technologically-driven changepoints more tractable. While a single individual could reasonably experience large changes in deep sleep due to injury or illness, it is unlikely that the median amount of deep sleep obtained by these Fitbit users truly decreased by 5-6 hours after July 19, 2017, only for it to rebound multiple times in subsequent months. Instead, these shifts are likely due to changes in Fitbit's calculation of deep sleep. It is critical to control for these technological changepoints in order to distinguish between them and any behaviorally-driven chances that are relevant to health and disease.

In Section 2, we review the ASCEPT procedure, which is broken down into two stages. In Section 3, we present results from running ASCEPT on data from the Precision VISSTA study. We conclude with a discussion in Section 4.

2 Methods

We now describe the two stages of ASCEPT. The first stage is an iterative process of obtaining empirical p-values for increasingly larger sets of changepoints. This stage will retain changepoints found to be significant and terminate upon finding a statistically insignificant result. The second stage is a process for trimming false positives that arise within linear trends and seasonality. That is, we are generally interested in identifying changepoints that correspond with a mean-shift, in which the average value of the observations suddenly changes. A linear trend, in which observations increase or decrease linearly with time, implicitly includes many mean-shifts. The same is true for seasonal patterns, in which observations systematically change depending on some period. For instance, individuals may take more steps during summer months than winter months every year. In both cases, it is more appropriate to identify the linear trend or seasonal pattern as a whole rather than breaking it down into arbitrarily many mean-shifts. Therefore, we consider changepoints that arise within a seasonal pattern to be false positives.

In practice, when changepoints, trends, and seasonality appear in real data, they may not be well-defined enough to serve as a gold standard. That is, where shifts, trends, and seasonality begin and end may not be particularly clear. Therefore, in order to demonstrate the stages of ASCEPT, we will use a simulated series of data, displayed at the top of Figure 2. This simulated time series of 800 observations has standard mean-shift changepoints at indices 49, 60, 600, 699, and 700, along with an upwards trend between indices 201 and 400 inclusive and a seasonal pattern between indices 401 and 600 inclusive. Using a simulated time series with these known properties will be helpful for demonstrating the processes ASCEPT uses. ASCEPT's workflow is displayed in Figure 2.

2.1 Empirical P-values for Candidate Changepoint Selection

The goal of the empirical p-value stage of ASCEPT is to incrementally include more changepoints detected by PELT until the newly proposed changepoints do not offer a statistically significant improvement in goodness-of-fit. Analogously, we are checking for a statistically significant decrease in a cost function. We will let τ_k indicate the cumulative set of changepoints detected by step k. We consider $\tau_0 = \emptyset$. That is, we initialize to the case where no changepoints have been detected. This corresponds with a scenario where a very large penalty has been imposed with PELT.

We will use Figure 3 to illustrate the process when iterating from step k to step $k + 1$.

At step k, we can consider having already detected changepoints τ_k. Decreasing the penalty associated with PELT, we can find the next set of changepoints. We will choose to denote this subsequent set of changepoints as τ^*_{k+1} since we have not yet determined whether or not we should reject τ^*_{k+1} as offering a statistically significant improvement in goodness-of-fit relative to τ_k. In Figures 3a and 3b, we consider the scenario where we have detected changepoints $\tau_k = \{305, 600\}$ and are evaluating $\tau^*_{k+1} = \{49, 60, 305, 600\}$. τ_k will generally be a subset of τ^*_{k+1}, but this doesn't necessarily have to be the case.

In order to empirically assess whether or not τ^*_{k+1} offers a significant improvement in goodness-of-fit, we must choose both a measure for goodness-of-fit and generate an empirical null distribution of this measure. While, in theory, one can choose any goodness-of-fit measure, we will specifically consider the log-likelihood under the assumption that the data are normally distributed. More specifically, between any two changepoints, or between a changepoint and the start or end of the series, the observations form a "segment". Each segment is assumed to consist of independent and identically distributed (iid) normal observations. While observations within a single segment follow the same normal distribution, observations across segments may follow different normal distributions. This parametric assumption is largely keeping in line with the implementation of PELT in `R`'s `changepoint` package [7], created by the authors of [4]. Additionally, the normality assumption is appropriate for a wide variety of scenarios, such as when using the mean or median of a variable given some simple assumptions [8]. It may also be reasonably assumed outright for many continuous variables, such as types of sleep, and count variables with large values, such as steps taken.

In order to generate an empirical null distribution for the change in the log-likelihood, we first generate a random sample corresponding with the observed data under the null. If at step k, there are $|\tau_k|$ changepoints, then there are $|\tau_k| + 1$ corresponding segments. For each segment, we randomly draw from a normal distribution with a mean and standard deviation corresponding to the sample mean and sample standard deviation of the segment in the observed data. For instance, in Figure 3a, the simulated data set has been split into three segments by the two changepoints at indices 305 and 600. In Figure 3c, we have gone segment-by-segment, and randomly sampled from the normal distribution that best fits each of the three segments in Figure 3a. Note that this random sample has been generated under the null in which there are no additional changepoints to be detected. We record the goodness-of-fit (i.e. the log-likelihood) for the random sample under the normality assumption by segment.

We then impose the changepoints in τ^*_{k+1} onto this same random sample and calculate the new log-likelihood, accounting for the segments imposed by τ^*_{k+1}. This is depicted in Figure 3d for the simulated data. The change in the log-likelihood under the null is then recorded.

This process of generating random samples under the null and calculating the change in the log-likelihood is repeated for some large number of trials/simulations, generally on the order of thousands or tens of thousands. An empirical p-value for the observed change in the log-likelihood is recorded. This p-value reflects the evidence against the null for the set τ^*_{k+1} as a whole relative to τ_k. That is, we are not assessing the significance of individual points within τ^*_{k+1} and then combining them. If the observed change is statistically significant at some chosen significance level, α, then we reject the null that τ^*_{k+1} does not offer an improved fit to the observed data relative to τ_k and retain τ^*_{k+1} as the current set of changepoints, τ_{k+1}. The procedure continues, comparing τ_{k+1} to τ^*_{k+2} and so forth, until we obtain a statistically insignificant result, as depicted in the second row of Figure 2.

The procedure of performing a subsequent hypothesis test if the current one yields a significant result is sometimes called a "fixed-sequence" procedure, or a stepwise "gatekeeping" procedure involving only one hypothesis test per family. Such an approach appears in other fields, such as clinical trials, and controls the family-wise error rate (FWER) at the nominal significance level chosen by the investigator [9, 10], thereby addressing the multiple testing issue.

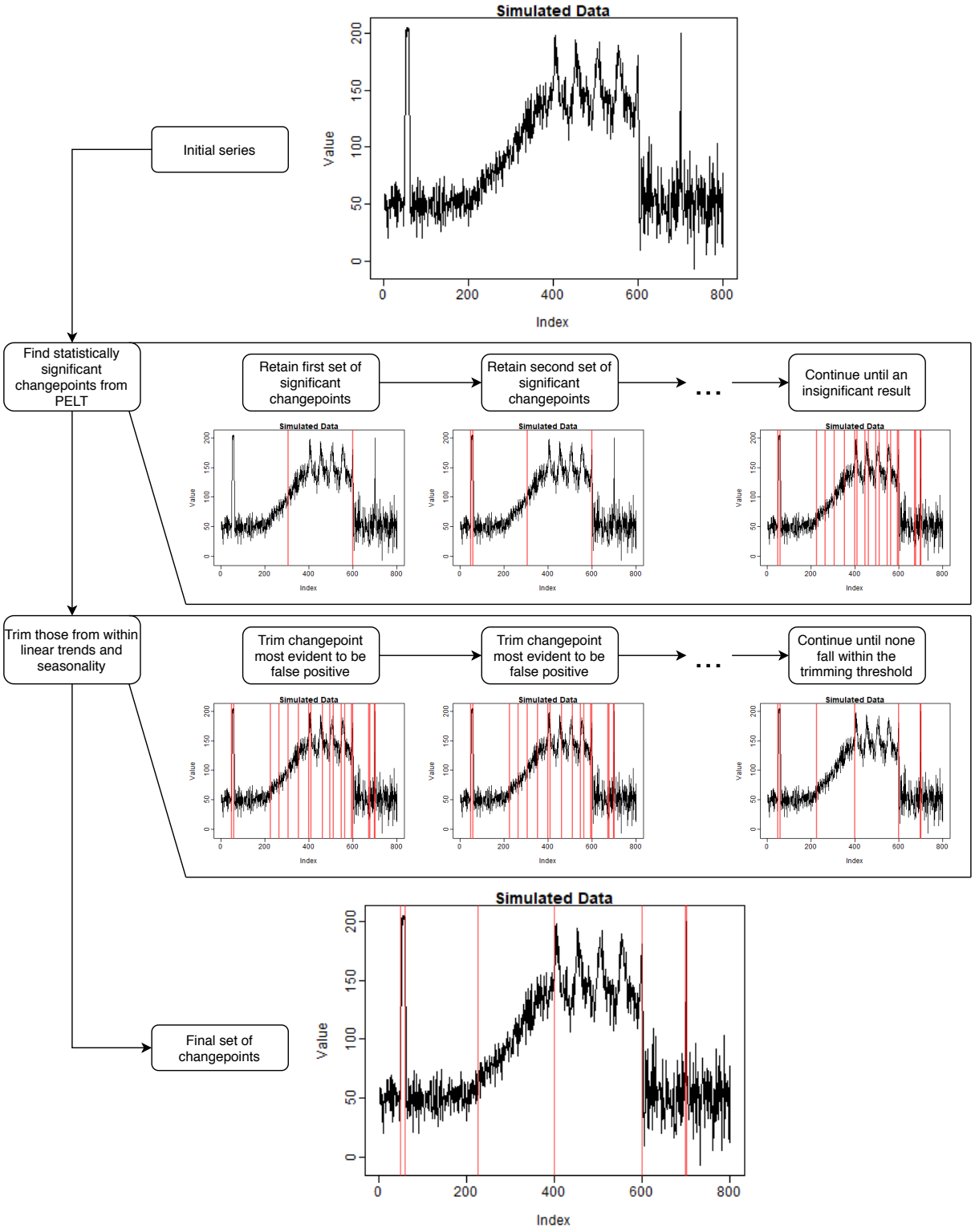
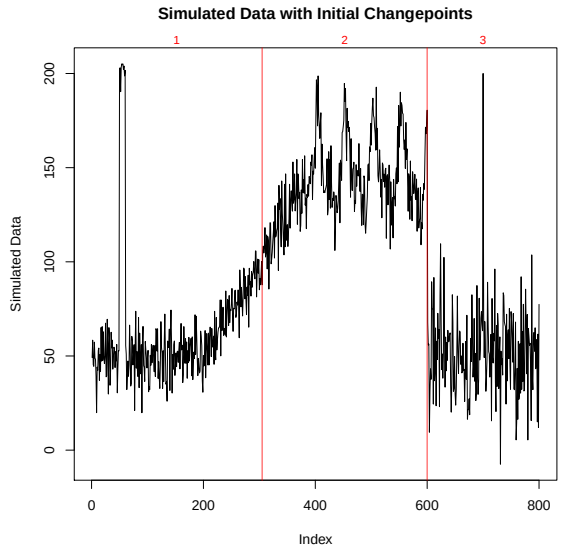
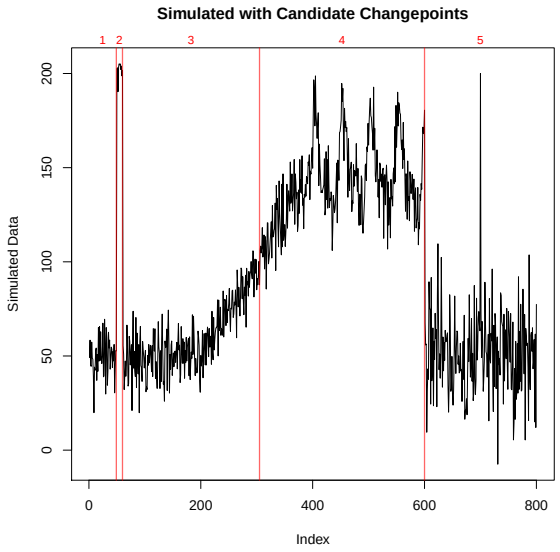


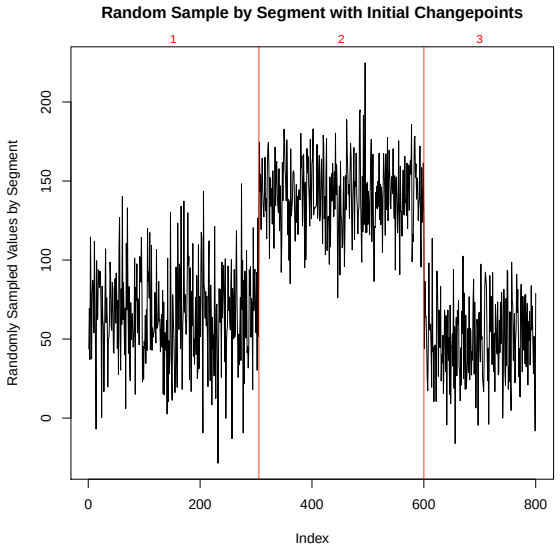
Figure 2: The ASCEPT workflow. The first row depicts a simulated data set with true changepoints at indices 49, 60, 600, 699, and 700. An upwards trend exists between indices 201 and 400 inclusive. A seasonal pattern exists between indices 401 and 600 inclusive. The second row corresponds with the empirical p-value procedure described in Section 2.1. The third row corresponds with the trimming procedure described in 2.2. The final set of identified changepoints, as a result of running ASCEPT, is shown in the fourth row.



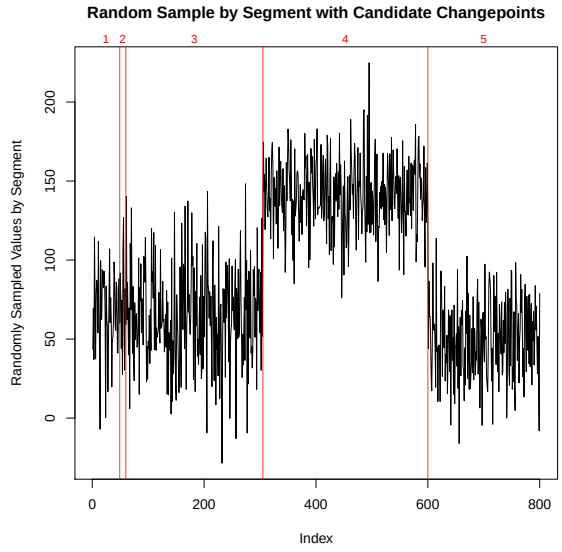
(a) Simulated data set with initial changepoints at indices 305 and 600. The log-likelihood assuming a normal distribution for each segment is -3727.3



(b) Simulated data set with the next set of candidate changepoints, {49, 60, 305, 600}, yielding 5 segments total. The log-likelihood assuming a normal distribution for each segment is -3512.3. The observed change in log-likelihood is therefore 215.0.



(c) A Monte Carlo sample and the changepoints associated with subfigure a. Each segment is randomly sampled from a normal with a mean and standard deviation equal to the sample mean and sample standard deviation of the corresponding segment from the original data. The log-likelihood assuming a normal distribution for each segment is -3746.6.



(d) The same Monte Carlo sample from subfigure c, but with proposed changepoints from subfigure b: {49, 60, 305, 600}. The log-likelihood assuming a normal distribution for each segment is -3745.0. The change in the log-likelihood for this random sample under the null is therefore 1.6.

Figure 3: The process by which it is determined whether or not a set of new candidate changepoints is significant. The process depicted in subfigures c and d is repeated a large number of times to generate an empirical null distribution of the change in the log-likelihood.

2.2 Trimming False Positives

Upon completing the process described in Section 2.1, it will be common to have false positives to the extent that many mean-shift changepoints may be detected within linear trends or seasonal patterns. That is, we may be interested in identifying when a trend or seasonal pattern begins or ends within a time series, but we do not want to identify changepoints within the trend or seasonal pattern. We will refer to this process of removing false positives as "trimming", but it is of the same principle as "pruning" used by methods such as CBS [11]. We purposefully avoid the term "prune" because it is overloaded in the context of changepoint detection. While it is used to refer to the process we describe here, it is also used by algorithms, including PELT [4], for describing a component of the optimization process.

We illustrate the trimming process in Figure 4 using the simulated data from Figure 2. To start, we can consider having some initial set of changepoints that are in need of trimming. In Figure 4a, we show such a set of changepoints for the simulated data. These changepoints are not all of those found after running the first stage of ASCEPT described in Section 2.1, but they are a large subset of those changepoints useful for illustrative purposes.

For every changepoint in consideration, we will perform two sets of model fits in order to assess whether or not it appears to be a false positive due to an ongoing linear trend or seasonality. The first set consists of piecewise linear regression and harmonic regression fits to each segment on either side of a candidate changepoint. The second set consists of overall linear regression and harmonic regression fits across the two segments, effectively ignoring the candidate changepoint.

For each set of fits, we calculate the root mean square error (RMSE). In cases when a recorded changepoint reflects a true mean-shift in the data, the piecewise fits should outperform the cross-segment fits by a relatively large margin. This is demonstrated in Figure 4b when considering a true changepoint at index 60 in the simulated data. In such a case, the best piecewise fit outperforms the best overall fit by approximately a factor of three. This large discrepancy in performance suggests that the candidate changepoint is not a false positive due to an ongoing trend or seasonal pattern.

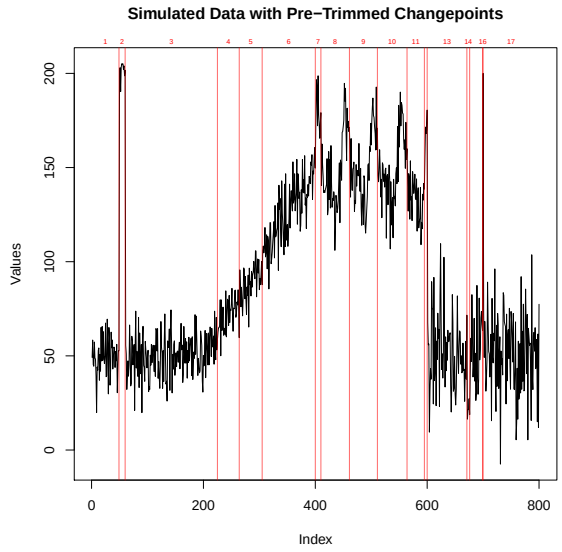
However, in cases when a candidate changepoint is actually a false positive within a linear trend, the RMSE associated with the best cross-segment fit will be relatively close to that for the best piecewise fit. For instance, in Figure 4c, the candidate changepoint is a false positive due to an ongoing trend. As a result, a linear regression fit across both segments on either side of the changepoint performs only marginally worse than the best piecewise fit to the segments. Likewise, in Figure 4d, the candidate changepoint is a false positive due to an ongoing seasonal pattern. As a result, a harmonic regression fit across both segments on either side of the candidate changepoint performs only marginally worse than the best piecewise fit to the segments.

This process of fitting piecewise and cross-segment linear and harmonic regressions is done for every candidate changepoint. In each case, the ratio of RMSE for the best cross-segment fit to the RMSE for the best piecewise fit is recorded. If one or more of these ratios fall below a given threshold, then the changepoint corresponding to the smallest ratio is trimmed and the process repeats until none of these ratios fall below the threshold, as depicted in the third row of Figure 2.

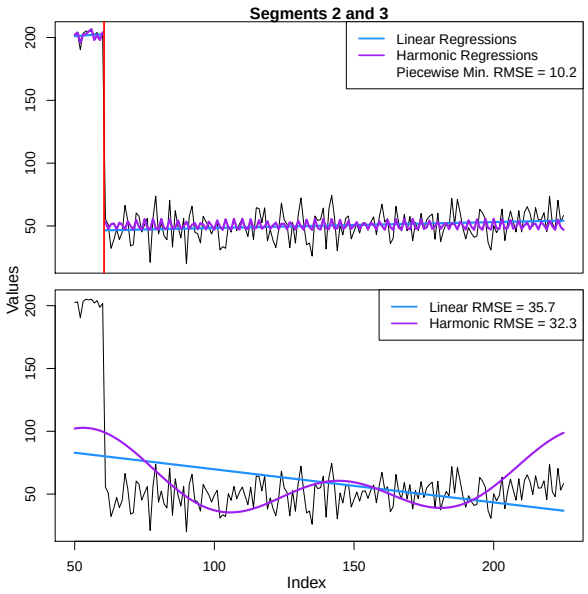
2.3 Precision VISSTA Data

Before evaluating the performance of ASCEPT, we describe the data from the Precision VISSTA study [6] on which ASCEPT is run in Section 3. The full pre-processed data set **[cite pre-processing manuscript]** includes users of many different devices, such as those from Fitbit, Garmin, and Withings, among others. Due to their prevalence, we choose to focus on individuals who use a relatively new Fitbit device (i.e. the Alta HR, Blaze, Charge 2, Charge 3, Inspire HR, Ionic, Versa, or Versa 2), multiple Fitbit devices, or an uknown Fitbit device. This subset includes 203,351 observations on 298 individuals recorded between May 15th, 2015 and October 27, 2019.

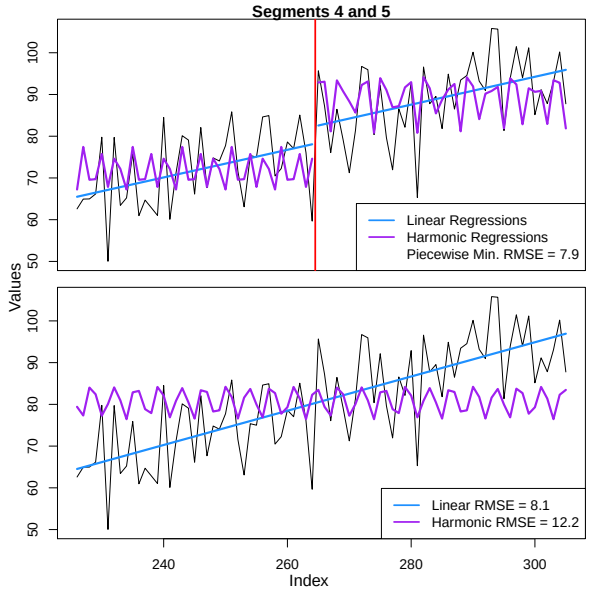
There are 12 activity and sleep variables available for analysis: steps taken, distance walked, floors climbed, elevation climbed, calories burned, time spent active, amount of total sleep, amount of deep sleep, amount of light sleep, amount of REM sleep, time spent awake at night, and times woken during the night. We focus on identifying changepoints based on the median of each variable on a daily basis. However, we exclude floors and elevation climbed as their median values stay within narrow ranges over time near zero. Likewise, we exclude REM sleep due to a lack of any non-missing values between May 20, 2016 and March 26, 2017. Missingness also differs across variables more generally. Following **[cite pre-processing manuscript]**, we can define the usage interval to be the time between an individual's first recorded non-missing value for a given variable and their last non-missing recorded value. Percent coverage refers to the proportion of these days for which a non-missing value was recorded. Steps, distance, calories burned, and active duration all have a median usage interval of 792.5 days with a median percent coverage of 92.0%. Deep sleep, light sleep, awake time, and times woken all have a median usage interval of 735.0 days with a median



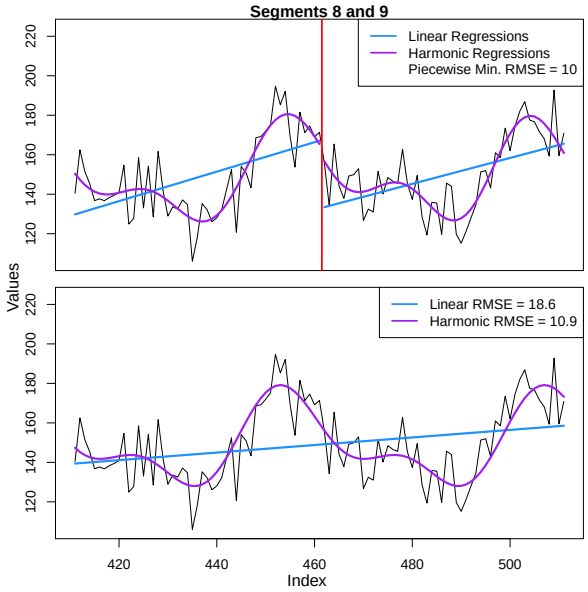
(a) Simulated data set with an initial set of changepoints to be trimmed. This set of changepoints is selected for illustrative purposes and does not include all pre-trimmed changepoints from Figure 2.



(b) Determination of whether to trim the changepoint between segments 2 and 3, which is a true changepoint. The cross-segment fits are more than 3 times worse than the best piecewise fit such that this changepoint would likely not be trimmed.



(c) Determination of whether to trim the changepoint between segments 4 and 5, which is a false positive due to a linear trend. The cross-segment linear fit is only about 3% worse than the best piecewise fit, such that we would likely choose to trim this changepoint.



(d) Determination of whether to trim the changepoint between segments 8 and 9, which is a false positive due to seasonality. The cross-segment harmonic fit is only about 9% worse than the best piecewise fit, such that we would likely choose to trim this changepoint.

Figure 4: The process by which candidate changepoints are trimmed. We consider each pair of consecutive segments and fit linear and harmonic regressions to each segment individually, as well as together. If the cross-segment fit is not more than some multiple (e.g. 1.1, 1.25, 1.5) times worse than the best piecewise fit, then the changepoint is removed. Alternatively, a threshold of $1 + p$ indicates that we trim if the best cross-segment fit is no more than $p \times 100\%$ worse than the best piecewise fit.

percent coverage of 74.5%. Total sleep has a median usage interval of 739.0 days with a median percent coverage of 75.1%.

3 Results

We now present some results from running ASCEPT as described in Section 2. We will present results for the simulated data from Figure 2, along with results for data from the Precision VISSTA study. Specifically, we show results for median deep sleep, median light sleep, and median total sleep obtained by users of newer Fitbit devices (i.e. the Alta HR, Blaze, Charge 2, Charge 3, Inspire HR, Ionic, Versa, or Versa 2), multiple Fitbit devices, or an uknown Fitbit device [**cite pre-processing manuscript**]. These results are displayed in Figure 5. For each data set, the empirical p-value process described in Section 2.1 was run using an $\alpha = 0.01$ significance level and 10,000 random samples/simulations generated under the null at each step. Additionally, the thresholds used to trim, as described in Section 2.2, in each case were 1.2. That is, changepoints whose best cross-segment fit had an RMSE no more than 1.2 times that for the best piecewise fit were subject to being removed.

For the simulated data, we see that the empirical p-value process finds both true mean-shift changepoints at indices 49, 60, 600, 699, and 700, and false positives within linear trends and seasonality, as shown in 5b. However, trimming the changepoints readily removes false positives and leaves detected changepoints at indices 49, 60, 225, 400, 600, 699, and 700, as shown in Figure 5c. Five of these correspond to standard mean-shift changepoints, while the other two effectively segment off the trend and seasonal pattern. In this case, we would interpret the final results as indicating that starting immediately after indices 49, 60, 225, 400, 600, 699, and 700, the time series experiences a mean-shift in its distribution statistically significant at the 0.01 level and that these changes are not attributable to an ongoing trend or seasonal pattern at a trimming threshold of 1.2. One would need to visually investigate whether or not a particular changepoint corresponds purely to a sudden mean-shift or the start/end of a trend or seasonal pattern, but this is easily assessed with a plot. In practice, it should be rare for a trend or seasonal pattern to start or end within a time series, as opposed to existing throughout the time series, such that most changepoints should correspond to a sudden mean-shift.

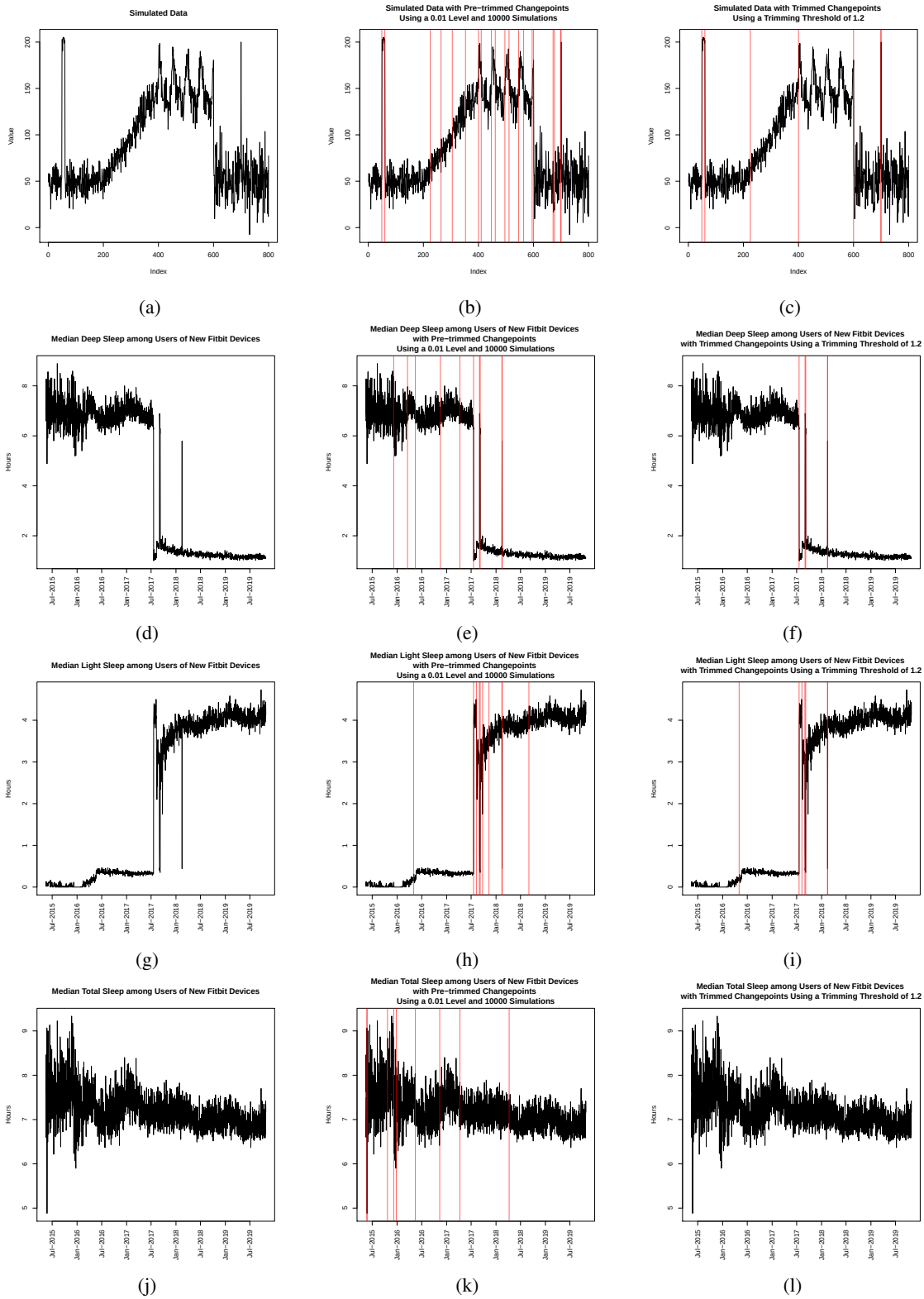


Figure 5: Overall results for simulated data (a-c), median deep sleep (d-f), median light sleep (g-i), and median total sleep (j-l) for users of newer Fitbit devices, multiple Fitbit devices, or an unknown Fitbit device. The first column of figures reflects the original data, the second column the changepoints initially detected with a significance level of 0.01 and 10,000 trials, and the third column the changepoints after trimming with a threshold of 1.2.

Running ASCEPT on median deep sleep and median light sleep among users of new Fitbit devices yields fairly comparable results. In Figures 5d and 5g, the original data visually demonstrate how deep sleep and light sleep are complements of one another. In running the empirical p-value process, there are clear false positives due to trends and seasonality shown in Figures 5e and 5h. However, trimming again primarily yields detection of true mean-shifts in the data, shown in Figures 5f and 5i. For median deep sleep, the final results indicate that starting immediately after July 19th, 2017, September 1st, 2017, September 6th, 2017, February 14th, 2018, and February 15th, 2018, the time series experiences mean-shifts in its distribution statistically significant at the 0.01 level that are not attributable to an ongoing trend or seasonal pattern at a trimming threshold of 1.2. These are days immediately after which, based on the chosen significance level of 0.01 and trimming threshold of 1.2, we suspect Fitbit changed the calculation of deep sleep. For median light sleep, the final results indicate that starting immediately after May 2nd, 2016, July 19th, 2017, August 9th, 2017, August 31st, 2017, September 6th, 2017, February 14th, 2018, and February 15th, 2018, the time series experiences mean-shifts in its distribution statistically significant at the 0.01 level that are not attributable to an ongoing trend or seasonal pattern at a trimming threshold of 1.2. These are days immediately after which, based on the chosen significance level of 0.01 and trimming threshold of 1.2, we suspect Fitbit changed the calculation of light sleep.

It is worth noting that the two changepoints in February 2018 correspond with a mean-shift lasting only a single day. Therefore, it is important to be able to detect such short changes in the context of mobile health. However, some comparable changepoint detection and selection approaches, such as CBS [11] are not designed for this context, an issue that we discuss further in Section 3.1.

We can also consider verifying these changepoint dates with information available online. While Fitbit lists previous firmware versions, it does not readily provide release dates or specific notes regarding each [12]. One needs to comb through community forums [13] to find more details. For instance, some firmware updates correspond with the general timeline of the changepoints observed here, such as the Charge 2 receiving firmware update 22.54.6 between July 24th, 2017 and August 1st, 2017 [14] and the Alta HR receiving firmware update 26.62.6 between August 1st, 2017 and August 10th, 2017 [15]. Likewise, Fitbit overhauled its calculation of sleep by introducing "Sleep Stages", starting on March 6th, 2017 [16]. However, this feature was rolled out with some difficulty. Glitches with Sleep Stages were reported from April 2017 through late July 2017 for Alta HR, Blaze, and Charge 2 devices [17]. Alta HR was again subject to glitches in September 2017 [18]. Thus, while we are not able to find posts corresponding with the exact dates found above, this only further emphasizes the need for an automated process by which changepoints are identified. Updates, roll-outs, and glitches may occur over the course of days or months, making pinning down a particular date difficult based on an online search. It is also possible that some sudden changes occur on Fitbit's end without formal notice on a blog or forum, rendering an online search effectively useless.

We use the daily median total sleep achieved by users of new Fitbit devices, depicted in Figure 5j, as a negative control. The final results indicate that the time series does not experience any mean-shifts in its distribution that are both statistically significant at the 0.01 level and not attributable to an ongoing trend or seasonal pattern at a trimming threshold of 1.2. Thus, we do not suspect that Fitbit changed the calculation of total sleep during the period of the study in any way that is significant at the 0.01 level that is also not attributable to an ongoing linear trend or seasonal pattern at a 1.2 tirmming threshold. The only potential false negative appears at the very start of the series on May 24th, 2015. This is a day near the beginning of the Precision VISSTA study on which only 7 individuals contributed data for those using newer, multiple, or unknown Fitbit devices. Therefore, this shift may be a true signal and we may not want to segment it off as we have done with other changepoints. However, if one did want to detect such a case, simply adjusting the threshold for trimming would enable the investigator to retain the changepoints initially detected on May 23rd and 24th of 2015, as shown in Figure 5k.

The final results from running ASCEPT for other variables in the Precision VISSTA study are shown in Section B of the Supplemental Material.

While we present the results when using specific values for the significance level and trimming thresholds, it's important to note that there will be some variation depending on how the parameters are set. As an example, we depict the variation in how changepoints are retained or trimmed in response to the trimming threshold in Figure 6 for the simulated time series.

3.1 Comparison with Circular Binary Segmentation

ASCEPT shares some of the same principles as previous methods for changepoint detection and selection, which are discussed in greater detail in Section A of the Supplemental Material. In particular, there have been earlier approaches that recursively run changepoint detection on subsegments until some stopping criterion is met. Of particular importance is CBS, which uses an empirical p-value approach for the initial detection of changepoints and a pruning, or trimming,

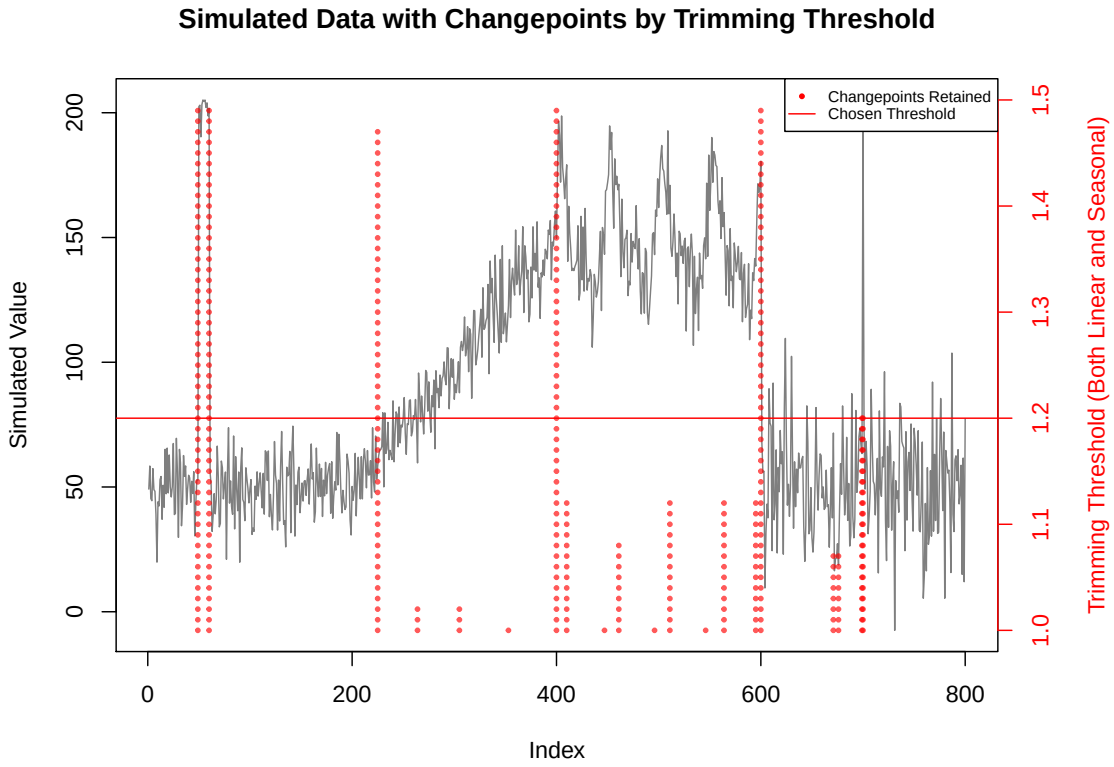


Figure 6: The simulated time series with changepoints initially detected using a 0.01 significance level and 10,000 Monte Carlo simulations, as indicated in Figure 5b, trimmed at various thresholds. Changepoints that are retained after the trimming process are indicated by a red point. At trimming thresholds equal to 1, all of the initially detected changepoints are retained. As the trimming threshold increases, changepoints are removed. At a trimming threshold of 1.5, all initially detected changepoints would be removed. In this case, thresholds between 1.13 and 1.2 inclusive would all yield the same final results as a threshold of 1.2, depicted in Figure 5c.

approach for removing false positives [11]. While ASCEPT follows the same general outline, we now discuss the differences in implementations and why they are important for the mobile health context.

Foremost, it is important to recall that CBS is an approximate method, one that only guarantees a locally optimal solution to changepoint detection. In contrast, ASCEPT is built upon PELT, which is an exact method that guarantees a globally optimal solution. Thus, in theory, the optimization offered by PELT itself should always be at least as good as that offered by CBS in an equivalent setting. In practice, equivalent settings (e.g. parameters) do not necessarily exist so some discrepancy is possible.

Regarding the use of an empirical p-value, the most important distinction between CBS and ASCEPT is the use of a permutation method versus random sampling from an assumed distribution. CBS effectively works by permuting the data within a segment, re-detecting changepoints in the permuted data, recording a test statistic, and comparing the empirical distribution of these test statistics under the null to the observed results. If the observed result is significant, the proposed changepoints are retained and if insignificant, they are rejected. What is important to note is that this process does not allow for the detection of single-point shifts in a series. If a segment has a mean-shift that only lasts for one observation, then permuting that segment and re-calculating a test statistic will yield nearly the same result each time, thereby preventing a significant observed result. In contrast, ASCEPT's empirical p-value approach allows for the detection of single-point shifts in a series, as shown with the simulated data, median deep sleep, and median light sleep recordings in the Precision VISSTA study in Section 3. This comes at the cost of greater reliance on a parametric assumption, in this case that the data within a segment are normally distributed.

In trimming, CBS uses a sum of squares measure to calculate the smallest subset of changepoints that can be retained without increasing the sum of squares measure beyond a certain threshold [11]. In contrast, ASCEPT uses an approach

that more directly targets linear trends and seasonal patterns by using linear regression and harmonic regression. As a result, the trimming done by ASCEPT is generally more favorable than that done by CBS when considering simulated data and data from the Precision VISSTA study, as we will show now.

We compare our implementation of ASCEPT with that for CBS provided in R's `DNAcopy` package [19], each using a significance level of $\alpha = 0.01$ and 10,000 random samples or permutations respectively. The trimming with ASCEPT was done with a threshold of 1.2 while the pruning with CBS was done with a threshold of 0.5 using the `undo.prune` argument for the `segment()` function in the `DNAcopy` package. For the particular data referred to here, these two thresholds appear comparable in the total number of changepoints they yield per data series.

In comparing the results for the simulated data in Figures 7b and 7c, we find that CBS cannot capture the single-point segment at index 700 and that it fails to segment off the trend and seasonality of the simulated data, instead effectively splitting the trend itself multiple times. Likewise, CBS fails to detect the single-day shift corresponding to February 15th, 2018 for median deep sleep among users of new, multiple, or unknown Fitbit devices, as shown in Figure 7f. As a third example, we see that CBS misses multiple changepoints, while also failing to trim or prune false positives for the median number of times woken during the night, as shown in Figure 7i. ASCEPT fares better in all three cases. Comparisons of ASCEPT and CBS for the remaining major variables from the Precision VISSTA study are provided in the Section B of the Supplemental Material.

4 Discussion and Conclusion

We have presented an approach for identifying changepoints that combines principles previously used with approximate methods, such as stopping criteria, with the current state-of-the-art exact method, PELT. ASCEPT begins by considering progressively larger sets of changepoints proposed by PELT under different optimization penalties. If a given set of changepoints is found to be significant based on a Monte Carlo procedure, they are retained and the next set of changepoints is assessed. This continues until a set is not found to be significant. Retained changepoints are then trimmed if it appears that they arose due to an ongoing trend or seasonal pattern.

ASCEPT offers several advantages over comparable methods. The first is that, because it currently uses PELT, it is guaranteed to select changepoints corresponding to a globally optimal solution, up to the constraint imposed by selecting a particular significance level. Second, as opposed to using PELT outright, which requires choice of a penalty for an optimization problem, ASCEPT allows an investigator to use a significance level, a largely more intuitive parameter. The choice of a penalty for optimization would tend to be highly specific to the particular data being analyzed, while significance levels offer a more universal measure that can be used across different data sets and variables. Third, ASCEPT's trimming process is specifically designed to consider linear trends and seasonality as they arise in mobile health data and time series at large. The thresholds associated with this trimming approach are also intuitive and will be easy to grasp for investigators. Thus, ASCEPT aims to utilize many of the advantages offered by PELT and CROPS, while presenting them in a more accessible manner tailored towards processing mobile health data.

Lastly, while the current presentation of ASCEPT uses PELT, the processes described here could, in theory, be applied to other changepoint detection algorithms. ASCEPT really only requires that consecutive sets of changepoints are proposed by an underlying detection algorithm. Likewise, assumptions made by ASCEPT, namely the normality of segments, could be adjusted to allow for other distributional assumptions as necessary.

However, there are also a couple of potential limitations associated with ASCEPT. The first is that because it involves a Monte Carlo method, ASCEPT is not necessarily guaranteed to give the same results over repeated runs. However, setting the number of simulations used to a relatively large number mitigates this issue. A second limitation is that it is necessary for the investigator to pick multiple thresholds in order to use ASCEPT, both a significance level and threshold for trimming. While these are relatively intuitive parameters, there is some subjectivity to selecting them and investigators may want to consider different threshold values that seem appropriate for their data. In particular, choosing a single trimming threshold may offer varying performance depending on the particular data series. An investigator may want to consider a few different trimming thresholds depending on the scale of changepoints relative to other observations in a given data series.

ASCEPT has been designed with the intention of establishing a formal process for selecting changepoints from those proposed by PELT, while also accounting for common false positives. This identification of changepoints will often be a necessary step for those hoping to analyze mobile health data, which are vulnerable to sudden changes in propriety algorithms used to record measurements. With ASCEPT, this process should be more approachable and effective than when using an alternative method, which may not provide as intuitive a procedure or may not appropriately account for false positives.

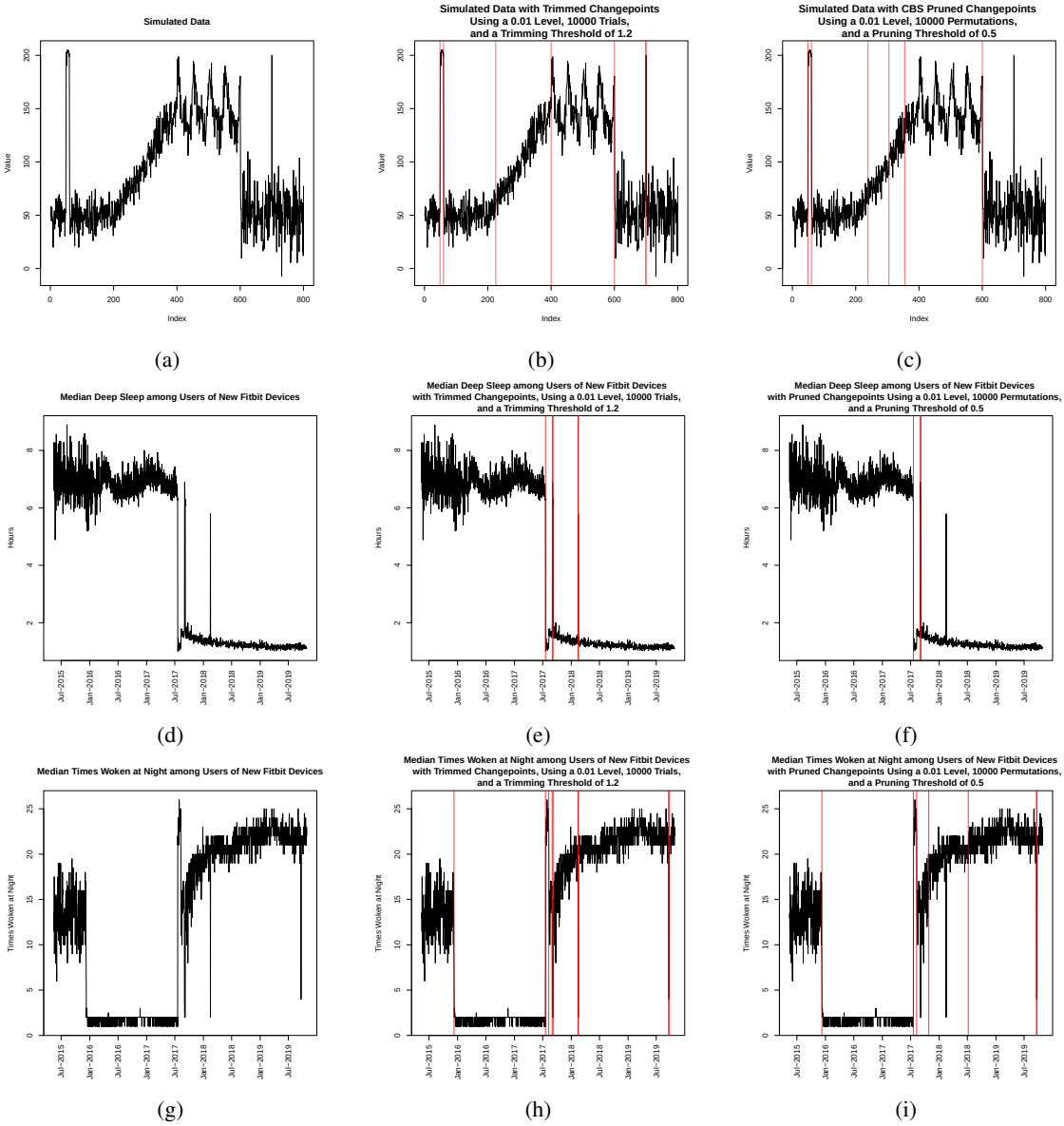


Figure 7: Comparison of results from ASCEPT with those from CBS for the simulated data (a-c), median deep sleep (d-f), and median times woken during the night (g-i) for users of new Fitbit devices, multiple Fitbit devices, or unknown Fitbit devices. Original time series are shown in the first column of subfigures, results from ASCEPT in the second column, and results from CBS in the third column. Each set of results used a significance level of $\alpha = 0.01$ with 10,000 simulations or permutations respectively. The trimming threshold for ASCEPT was 1.2 while the pruning threshold for CBS (`undo.prune` in the `DNAcopy` package in `R`) was 0.5.

4.1 R Package

ASCEPT is currently implemented in an `R` package, named `changepointSelect`, hosted on GitHub at `https://github.com/matthewquinn1/changepointSelect`.

References

[1] Santosh Kumar, Wendy J. Nilsen, Amy Abernethy, Audie Atienza, Kevin Patrick, Misha Pavel, William T. Riley, Albert Shar, Bonnie Spring, Donna Spruijt-Metz, Donald Hedeker, Vasant Honavar, Richard Kravitz, R. Craig Lefebvre, David C. Mohr, Susan A. Murphy, Charlene Quinn, Vladimir Shusterman, and Dallas Swendeman.

Mobile health technology evaluation: The mhealth evidence workshop. *American Journal of Preventive Medicine*, 45(2):228 – 236, 2013.

[2] Bruno M.C. Silva, Joel J.P.C. Rodrigues, Isabel de la Torre Díez, Miguel López-Coronado, and Kashif Saleem. Mobile-health: A review of current state in 2015. *Journal of Biomedical Informatics*, 56:265 – 272, 2015.

[3] Heval Mohamed Kelli, Bradley Witbrodt, and Amit Shah. The future of mobile health applications and devices in cardiovascular health. *European Medical Journal Innovations*, pages 92–97, January 2017.

[4] Rebecca Killick, Paul Fearnhead, and Idris Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[5] Kaylea Haynes, Idris A. Eckley, and Paul Fearnhead. Computationally efficient changepoint detection for a range of penalties. *Journal of Computational and Graphical Statistics*, 26(1):134–143, 2017.

[6] Arlene Chung, David Gotz, Michael Kappelman, Luca Mentch, Kimberly Glass, and Nils Gehlenborg. Precision VISSTA: Enabling Precision Medicine through the Development of Quantitative and Visualization Methods. `http://precisionvissta.web.unc.edu/`. Accessed: 2020-3-14.

[7] Rebecca Killick and Idris A. Eckley. changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19, 2014.

[8] Paul R. Rider. Variance of the median of small samples from several special populations. *Journal of the American Statistical Association*, 55(289):148–150, 1960.

[9] A. Dmitrienko, A.C. Tamhane, and F. Bretz. *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman & Hall/CRC Biostatistics Series. CRC Press, 2009.

[10] Alex Dmitrienko and Ajit C. Tamhane. Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics*, 6(3):171–180, 2007.

[11] Adam Olshen, E. S. Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, October 2004.

[12] What's changed in the latest Fitbit device update? `https://help.fitbit.com/articles/en_US/Help_article/1372`. Accessed: 2020-06-06.

[13] Community. `https://community.fitbit.com/`. Accessed: 2020-06-06.

[14] Charge 2 Firmware Release - 22.54.6. `https://community.fitbit.com/t5/Charge-2/Charge-2-Firmware-Release-22-54-6/td-p/2106845`, 2017. Accessed: 2020-06-06.

[15] Alta HR Firmware Release - 26.62.6. `https://community.fitbit.com/t5/Alta-HR/Alta-HR-Firmware-Release-26-62-6/td-p/2119538`, 2017. Accessed: 2020-06-06.

[16] Danielle Kosecki. New Fitbit Features Deliver Data Previously Only Available Through a Sleep Lab. `https://blog.fitbit.com/sleep-stages-and-sleep-insights-announcement/`, 2017. Accessed: 2020-06-11.

[17] Charge 2 Sleep Stages. `https://community.fitbit.com/t5/Charge-2/Charge-2-Sleep-Stages/td-p/1907433`, 2017. Accessed: 2020-06-11.

[18] Received Classic Sleep rather than Sleep Stages. `https://community.fitbit.com/t5/Alta-HR/RESOLVED-9-3-Received-Classic-Sleep-rather-than-Sleep-Stages/td-p/2174232`, 2017. Accessed: 2020-06-11.

[19] Venkatraman E. Seshan and Adam Olshen. *DNAcopy: DNA copy number data analysis*, 2019. R package version 1.60.0.

[20] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[21] Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, 42(6):2243–2281, December 2014.

[22] Ivan Auger and Charles Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, pages 39–54, 1989.

[23] Brad Jackson, Jeffrey D. Scargle, David Barnes, Sundararajan Arabhi, Alina Alt, Peter Gioumousis, Elyus Gwin, Paungkaew Sangtrakulcharoen, Linda Tan, and Tun Tao Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108, February 2005.

A Review of Offline Changepoint Detection

Various methods for offline changepoint detection have been created over the years, and while more extensive reviews exist [20], we briefly review the most relevant approaches here. In particular, we are concerned with the challenge of offline changepoint detection for an unknown number of changepoints and focus primarily on changepoints that reflect mean-shifts in a time series. This is a very common scenario for changepoint analysis and appears reasonable for mobile health research in particular.

In offline changepoint detection, the goal is typically to perform an optimization. Often, one will make some parametric assumption about the data, such as being normally distributed. All observations between two changepoints, which form a "segment", are often assumed to follow the same distribution, while those in segments separated by changepoints are allowed to follow different distributions, such as normal distributions with different means. Many detection algorithms will identify changepoints as to minimize a cost function (e.g. negative log-likelihood) subject to a penalty for introducing additional changepoints that prevents overfitting. There are methods that provide approximate, or locally optimal, results as well as those that provide exact, or globally optimal, results. While approximate methods do not guarantee a globally optimal result, they typically offer lower computational complexities.

Arguably the most popular approximate method is binary segmentation. Binary segmentation effectively considers splitting a time series of observations, $y_1, \ldots, y_T$ for times $t = 1, \ldots, T$ into two subsegments by identifying a changepoint at time τ. To do so, one first defines a cost function, $\mathcal{C}(\cdot)$, and sets $\tau = \text{argmin}_{t \in \{1,\ldots,T\}} \mathcal{C}(y_1, \ldots, y_t) + \mathcal{C}(y_{t+1}, \ldots, y_T)$. Here, the cost function may be something such as the negative log-likelihood, if assuming a parametric model. If one wishes to detect multiple changepoints, then one can run this minimization again on each subsegment, one from $t = 1$ to $t = t^*$ and the other from $t = t^* + 1$ to $t = T$. This process repeats until some stopping criterion is met. The primary advantage of this approach is its relatively low computational complexity of $\mathcal{O}(n \log n)$ when considering a series of n observations [4].

Other approximate approaches have built off of binary segmentation, such as Circular Binary Segmentation (CBS) [11], which allows for detection of two changepoints at a time, and Wild Binary Segmentation (WBS) [21], which randomly draws and checks segments. Though CBS is approximate, we will demonstrate that ASCEPT actually utilizes similar principles to it. For instance, CBS generates empirical p-values to iteratively assess potential changepoints, retaining those found to be significant. It then prunes, or trims, the final set of changepoints found to remove false positives. ASCEPT follows a comparable procedure, but uses different implementations at each step. We discuss this comparison further in Section 3.1.

As with approximate methods, a number of exact methods for multiple changepoint detection have been proposed. However, these have historically suffered from relatively poor computational complexities. For instance, the Segment Neighbourhood method [22] has $\mathcal{O}(mn^2)$ complexity for a time series of length n and m changepoints. Likewise, the Optimal Partitioning algorithm has $\mathcal{O}(n^2)$ computational complexity [23]. The method that we will consider to be the state-of-the-art is Pruned Exact Linear Time (PELT), a modified version of the Optimal Partitioning algorithm that is capable of running in $\mathcal{O}(n)$ time under certain assumptions [4]. Consider detecting m changepoints, $\tau_1, \ldots, \tau_m$, with $1 \leq \tau_1 \leq \cdots \leq \tau_m \leq n - 1$. We define $\tau_0 = 0, \tau_{m+1} = n$ for the purpose of segmenting all of the data. For a cost function, $\mathcal{C}(\cdot)$, PELT performs the minimization:

$$\min_{m, \tau_1, \ldots, \tau_m} \sum_{i=1}^{m+1} \left[\mathcal{C}(y_{\tau_{i-1}+1}, \ldots, y_{\tau_i}) \right] + \beta f(m) \tag{1}$$

where $f(m)$ is a penalty based on the number of changepoints and β is a multiplier on the penalty. PELT is primarily intended for use with a penalty linear in the number of changepoints, $\beta f(m) = \beta m$. Under this condition, we can re-expression Equation 1 as:

$$\min_{m, \tau_1, \ldots, \tau_m} \sum_{i=1}^{m+1} \left[\mathcal{C}(y_{\tau_{i-1}+1}, \ldots, y_{\tau_i}) + \beta \right] \tag{2}$$

PELT solves this optimization problem using dynamic programming in a similar manner to Optimal Partitioning [23], but is able to obtain its considerable speed-up by pruning the space over which it searches for changepoints. Namely, consider the scenario where the cost function is defined to be the negative log-likelihood associated with a segment. Likewise consider indices t and s where $t < s < T$, letting $\mathcal{T}_t$ denote the set of possible changepoints to be detected over indices $1, \ldots, t$ and likewise for $\mathcal{T}_s$. Then, if:

$$\left[\min_{m, \mathcal{T}_t} \sum_{i=1}^{m+1} \left[\mathcal{C}(y_{\tau_{i-1}+1}, \ldots, y_{\tau_i}) + \beta \right] \right] + \mathcal{C}(y_t, \ldots, y_s) \geq \left[\min_{m, \mathcal{T}_s} \sum_{i=1}^{m+1} \left[\mathcal{C}(y_{\tau_{i-1}+1}, \ldots, y_{\tau_i}) + \beta \right] \right] \tag{3}$$

holds, then t cannot be the last optimal change point prior to T [4]. Under certain regularity conditions, notably that the expected number of changepoints increases linearly with n, this approach can achieve a complexity of $\mathcal{O}(n)$. In the worst cases, PELT has the same computational complexity as Optimal Partitioning, $\mathcal{O}(n^2)$.

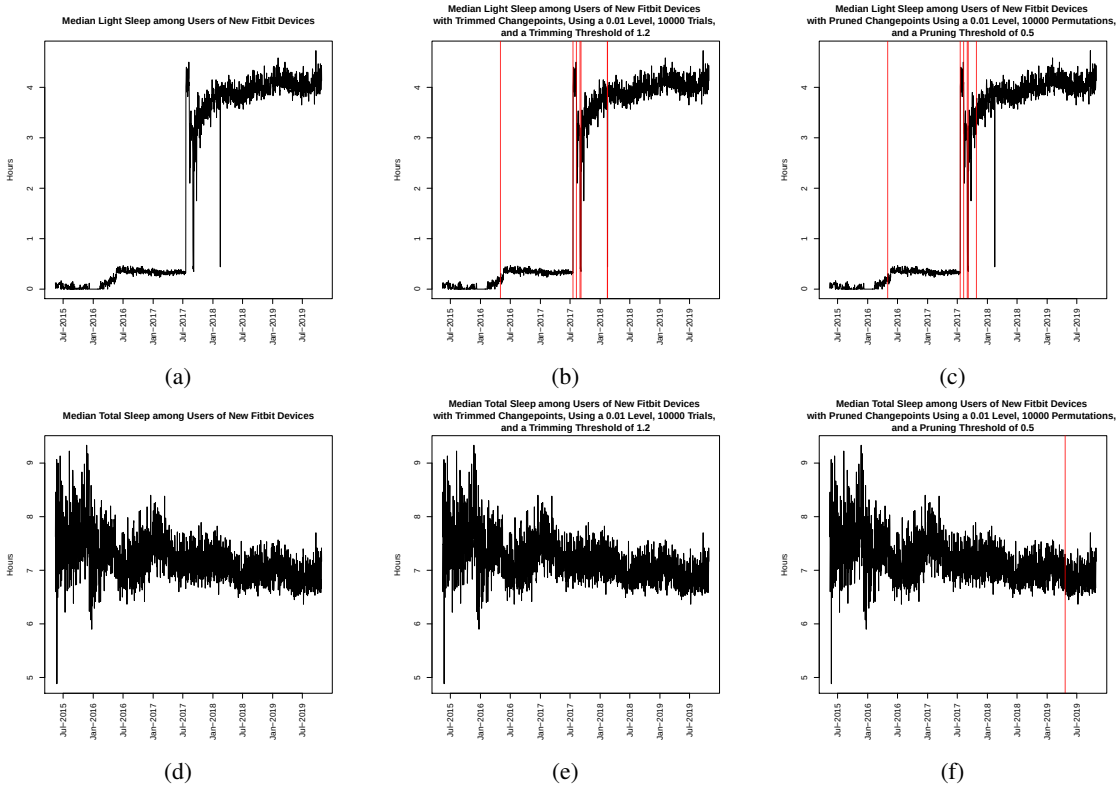
The main difficulty in using PELT is the specification of the penalty constant, β. Choosing such a β is a non-intuitive choice. To assist with this, the Changepoints for a Range of PenaltieS (CROPS) algorithm offers an efficient approach for running PELT for many different βs, namely all of those between some chosen $\beta_{\min}$ and $\beta_{\max}$ [5]. For instance, CROPS takes advantage of the fact that many different penalty constants will yield the same results under PELT. That is, if a chosen β yields the set of changepoints $\mathcal{T}$, then increasing or decreasing β by a marginal amount will not lead to fewer or more changepoints ultimately being detected by PELT. Using CROPS, PELT needs to be run a maximum of $m(\beta_{\min}) - m(\beta_{\max}) + 2$ times where $m(\beta)$ refers to the number of changepoints detected under penalty constant β.

Running CROPS on PELT allows an investigator to explore the results from PELT under many different penalties. However, this approach still suffers from some practical challenges from the investigator's perspective. While picking a penalty value is not as difficult a choice when using CROPS, since one simply chooses a large range of penalties, the results do not provide an indication of which set of changepoints is actually optimal. The investigator is simply given the results under many runs, not the "best" set from those many runs. The investigator will have to manually find which set of changepoints yields the best fit. Implicitly, we need an approach for selecting an optimal penalty value among those considered by CROPS. This, again, will be tedious for those with many series to investigate, but also presents some issues with formality. Different investigators may have different approaches for determining the truly optimal set of changepoints or optimal pen. Thus, there is a clear need for a formal and more rigorous approach for selecting a final set of changepoints.

B Additional Results for Precision VISSTA

We now present the results from running ASCEPT on different data series from the Precision VISSTA mobile health study, excluding those that were presented in Section 3. In Supplemental Figures S1, S2, and S3, we present the results for both ASCEPT and CBS, using comparable parameters. In general, we find that the results from ASCEPT are favorable both in identifying true shifts in the data, especially those lasting only one day, and in trimming false positives.

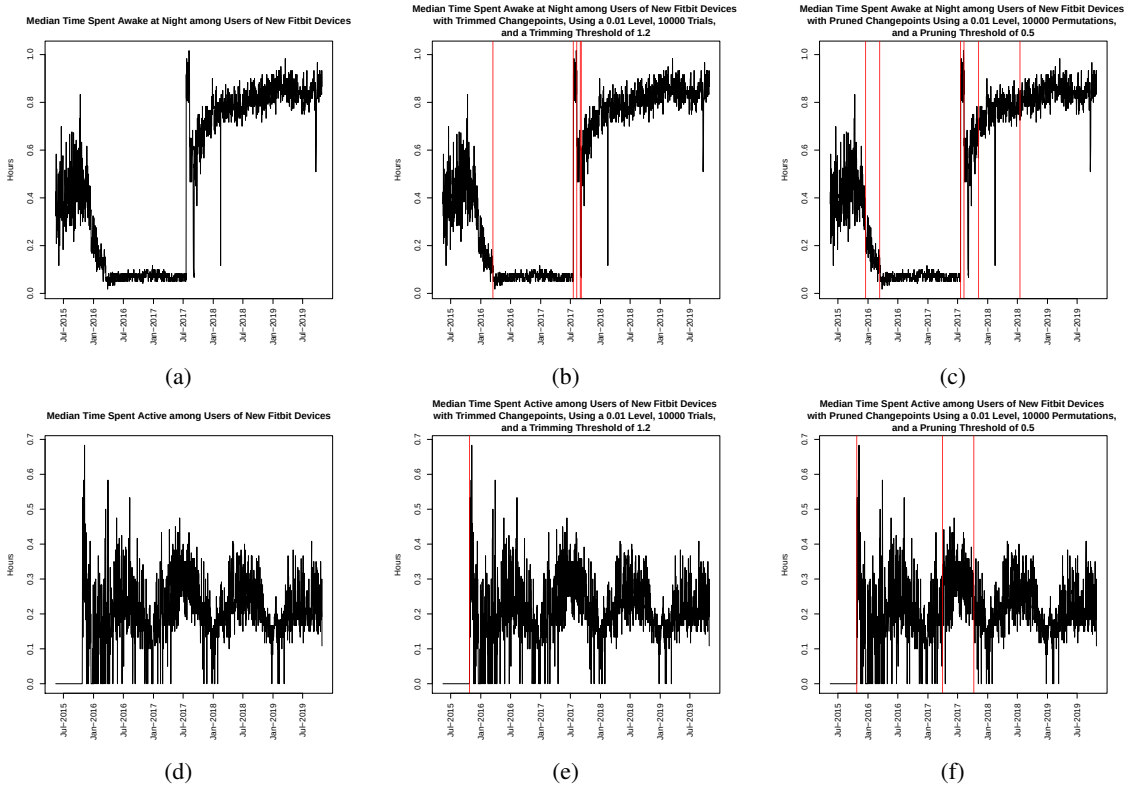
While ASCEPT performs very well on these various time series in general, the one major exception arises when investing the median time spent awake at night, depicted in Supplemental Figures S2a-S2c. Here, both ASCEPT and CBS miss two true changepoints for the given parameters. In the case of ASCEPT, reducing the trimming threshold to 1.15 will allow capture of one of these changepoints. Interestingly, this variable is nearly identical to times woken during the night, on which ASCEPT performed very well, shown in Supplemental Figure 7h. Thus, it is clear that small changes in a series can still yield fairly different results in the final changepoints identified. Additionally, changing the trimming threshold to 1.15 would also introduce a couple false positives in series of median times woken during the night. This again emphasizes the need to consider multiple trimming thresholds and the trade-off between identifying true positives and false positives.



(a) (b) (c)

(d) (e) (f)

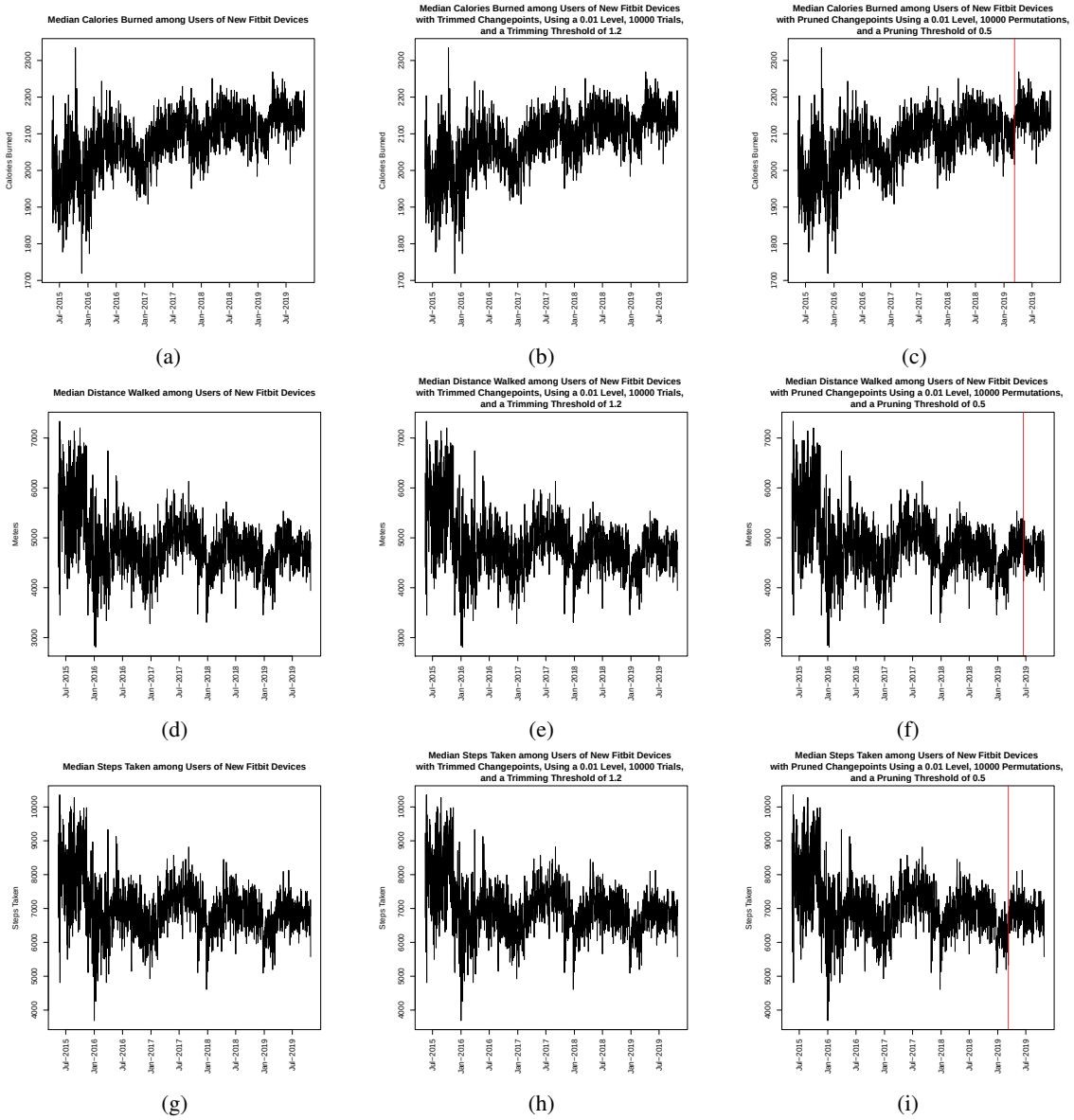
Supplemental Figure S1: Comparison of results from ASCEPT with those from CBS for the median light sleep (a-c) and median total sleep (d-f) for users of new Fitbit devices, multiple Fitbit devices, or unknown Fitbit devices. Original time series are shown in the first column of subfigures, results from ASCEPT in the second column, and results from CBS in the third column. Each set of results used a significance level of $\alpha = 0.01$ with 10,000 simulations or permutations respectively. The trimming threshold for ASCEPT was 1.2 while the pruning threshold for CBS (`undo.prune` in the `DNAcopy` package in `R`) was 0.5.



(a) (b) (c)

(d) (e) (f)

Supplemental Figure S2: Comparison of results from ASCEPT with those from CBS for the median time spent awake at night (a-c) and median time spent active during the day (d-f) for users of new Fitbit devices, multiple Fitbit devices, or unknown Fitbit devices. Original time series are shown in the first column of subfigures, results from ASCEPT in the second column, and results from CBS in the third column. Each set of results used a significance level of $\alpha = 0.01$ with 10,000 simulations or permutations respectively. The trimming threshold for ASCEPT was 1.2 while the pruning threshold for CBS (`undo.prune` in the `DNAcopy` package in `R`) was 0.5.



Supplemental Figure S3: Comparison of results from ASCEPT with those from CBS for the median calories burned (a-c), median distance walked (d-f), and median steps taken (g-i) for users of new Fitbit devices, multiple Fitbit devices, or unknown Fitbit devices. Original time series are shown in the first column of subfigures, results from ASCEPT in the second column, and results from CBS in the third column. Each set of results used a significance level of $\alpha = 0.01$ with 10,000 simulations or permutations respectively. The trimming threshold for ASCEPT was 1.2 while the pruning threshold for CBS (`undo.prune` in the `DNAcopy` package in `R`) was 0.5.