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

February 8, 2021

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 (mHealth) 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 to 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". These changepoints frequently correspond with a change in the mean of the data, making "mean-shift" changepoints our primary focus. 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.

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. However, it is difficult to scale these monitoring approaches 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.

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. The ASCEPT procedure consists of two main stages. 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 removing, or "trimming", false positives. In some cases, ASCEPT may initially select changepoints within a pattern, such as a linear trend or seasonality. Though these patterns implicitly include many mean-shifts, they are often behaviorally-driven and will be interesting phenomena for the investigator to discuss, rather than technological artifacts to adjust for. We therefore prefer to identify these patterns as a whole, rather than arbitrarily splitting them up with identified changepoints. Trimming addresses this concern.

ASCEPT is designed for 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 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 describe ASCEPT's procedure, decomposing it into its two main stages. We then present results from running ASCEPT on both simulated data and real mHealth data from the Precision VISSTA study on individuals with inflammatory bowel disease (IBD) [6]. This study collects data on individuals' sleep and daily activity habits, among other characteristics. We then compare the performance of ASCEPT on these data sets to that of a comparable method, Circular Binary Segmentation (CBS) [7]. We find that ASCEPT appropriately identifies changepoints on the simulated data and most of the Precision VISSTA variables considered. Additionally, ASCEPT appears to provide better changepoint detection and selection on these data than CBS under comparable settings. We find these differences to consequentially have a clear impact when attempting to correct the simulated data for these changepoints. ASCEPT provides an effective manner by which mHealth researchers can objectively identify mean-shift changepoints before continuing with downstream analysis.

2 Methods

2.1 Data

2.1.1 Precision VISSTA Data

We were primarily interested in developing a changepoint selection procedure for mHealth data. Therefore, we evaluated the performance on ASCEPT on real data from the Precision VISSTA study [6]. This study's 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 unknown Fitbit device. This subset of the data 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, distance, floors, elevation, calories, time active, total sleep, deep sleep, light sleep, REM sleep, time awake at night, and times woken up. However, we excluded floors and elevation climbed as their median values stay within narrow ranges near zero over the study period. Likewise, we excluded 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 defined 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

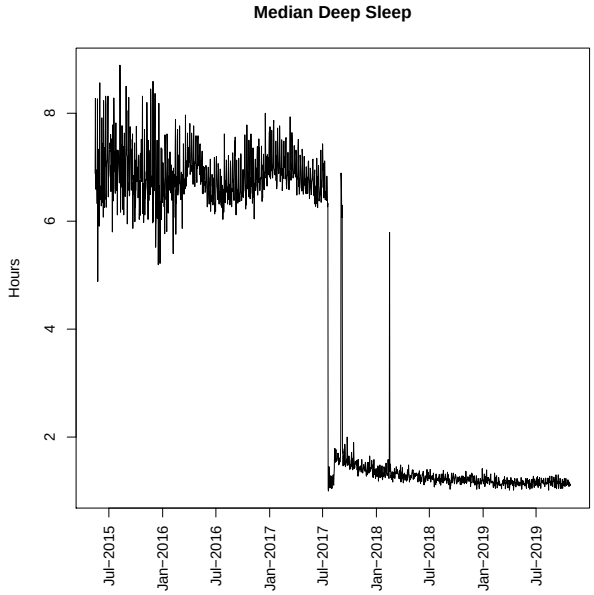


Figure 1: The median deep sleep among participants in the Precision VISSTA study who use a relatively new Fitbit device (i.e. the Alta HR, Blaze, Charge 2, Charge 3, Inspire HR, Ionic, Versa, or Versa 2), multiple Fitbit devices, or an unknown Fitbit device.

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

In Figure 1, we present one variable from the Precision VISSTA study: the daily median amount of deep sleep in our subset of users of relatively new Fitbit devices, multiple Fitbit devices, or an unknown Fitbit device. 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 then identify true human behavior that is relevant to health and disease.

2.1.2 Simulated Data

In practice, when changepoints and patterns, such as linear trends and seasonlity, appear in real data, they are often not defined well enough to serve as a gold standard for illustrative purposes. In Figure 1, there seems to be seasonality prior to July 19, 2017, but it does not appear to be consistent. The peak of this pattern in 2016 is considerably sharper than in 2017 and it's not clear that the same seasonality occurred in late 2015. Likewise, different investigators may dispute whether some points in 2015 and early 2016 constitute behaviorally-driven mean-shift changepoints. From May 15, 2015 to May 15, 2016, the study included anywhere from only 7 to 54 unique users of relatively new Fitbit devices, multiple Fitbit devices, or an unknown Fitbit device contributing deep sleep observations. Large behaviorally-driven fluctuations were therefore more likely during this time compared with later in the study, when up to 160 unique users contributed deep sleep observations.

Therefore, in order to describe 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 sudden mean-shift changepoints at indices 49, 60, 600, 699, and 700, along with an increasing linear 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 by avoiding any ambiguity. ASCEPT's workflow is displayed in Figure 2. We return to Precision VISSTA Data in Section 3 in order to present the results of running ASCEPT on real mHealth data.

2.2 Stage 1: Empirical P-values for Changepoint Selection

The goal of the first stage of ASCEPT is to incrementally include more mean-shift changepoints detected by PELT until the newly proposed changepoints do not offer a statistically significant improvement in goodness-of-fit. We will let τ_k indicate the cumulative set of changepoints detected by step k, where $\tau_0 = \emptyset$. That is, we initialize to the case where no changepoints have been detected. This corresponds with a scenario where a very large optimization 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 $\tau^*_{k+1} = \{49, 60, 305, 600\}$ for significance. τ_k will generally be a subset of τ^*_{k+1}, but this is not always the case.

In order to empirically assess whether or not τ^*_{k+1} offers a significant improvement in goodness-of-fit, we must both choose 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 [8], 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 [9]. Normality 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.

The null to be assessed is that the changepoints contained in τ_k are all of the true mean-shift changepoints in the time series. In order to generate an empirical null distribution, 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 are split into three segments by the two changepoints at indices 305 and 600. In Figure 3c, we randomly sample from the normal distributions that best fit each of the three segments in Figure 3a. Note that this random sample has been generated under the null in which the changepoints in τ_k are all of the true mean-shift changepoints. We record the goodness-of-fit (i.e. the log-likelihood) for this random sample.

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 represents the goodness-of-fit under the null if the proposed changepoints in τ^*_{k+1} were were used instead. It is depicted in Figure 3d for the simulated data. The change in the log-likelihood under the null, when using τ^*_{k+1} compared with τ_k, is then recorded. One could consider re-detecting changepoints on this random sample, instead of using those already found in τ^*_{k+1}, in order to generate results under the null. We choose not to do this for multiple reasons. In particular, this would greatly increase the computational complexity of the procedure, but it would not yield appreciably different results from the current design. Since the Monte Carlo sample truly only contains noise outside of changepoints in τ_k, the change in the log-likelihood whether re-detecting changepoints or using τ^*_{k+1} will tend to only slightly differ. While this could still result in some differences regarding which changepoints are ultimately detected, most of these detected changepoints associated with small changes in the log-likelihood would be trimmed in Stage 2 regardless. Thus, we choose the approach that offers magnitudes of improvement in run-time.

This process of generating random samples under the null and calculating the change in the log-likelihood is repeated for a large number of simulations, generally on the order of thousands or tens of thousands. Based on these simulations, 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 contains all the true mean-shift changepoints for the time series, and instead select 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 [10, 11], thereby addressing the multiple testing issue.

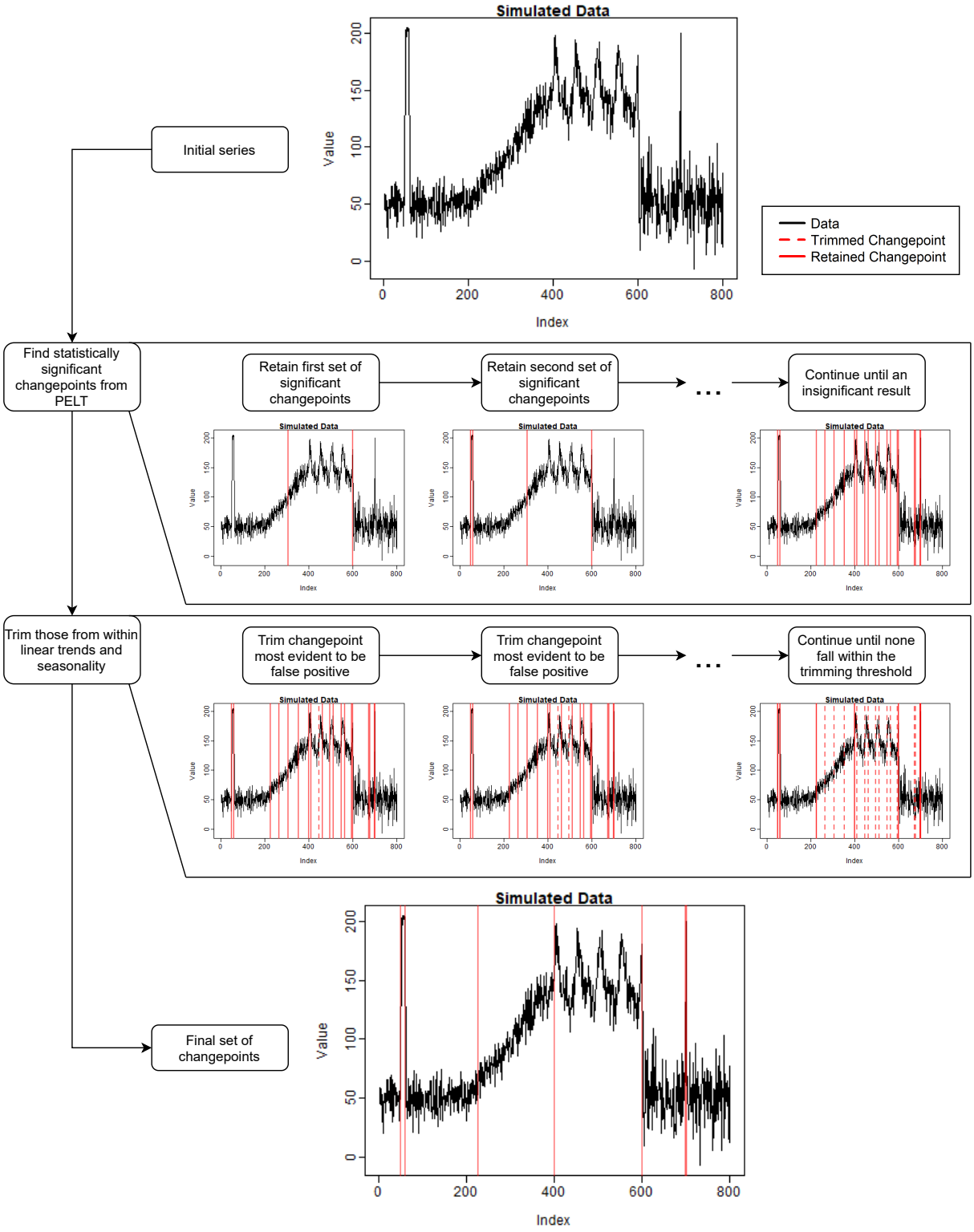


Figure 2: The ASCEPT workflow with proposed changepoints indicated in red. The first row depicts a simulated data set with sudden mean-shift changepoints at indices 49, 60, 600, 699, and 700. An upwards trend exists between indices 201 and 400 inclusive. A seasonal pattern exists between indices 401 and 600 inclusive. The second row corresponds with the empirical p-value procedure described in Section 2.2. The third row corresponds with the trimming procedure described in 2.3. The final set of identified changepoints, as a result of running ASCEPT, is shown in the fourth row.

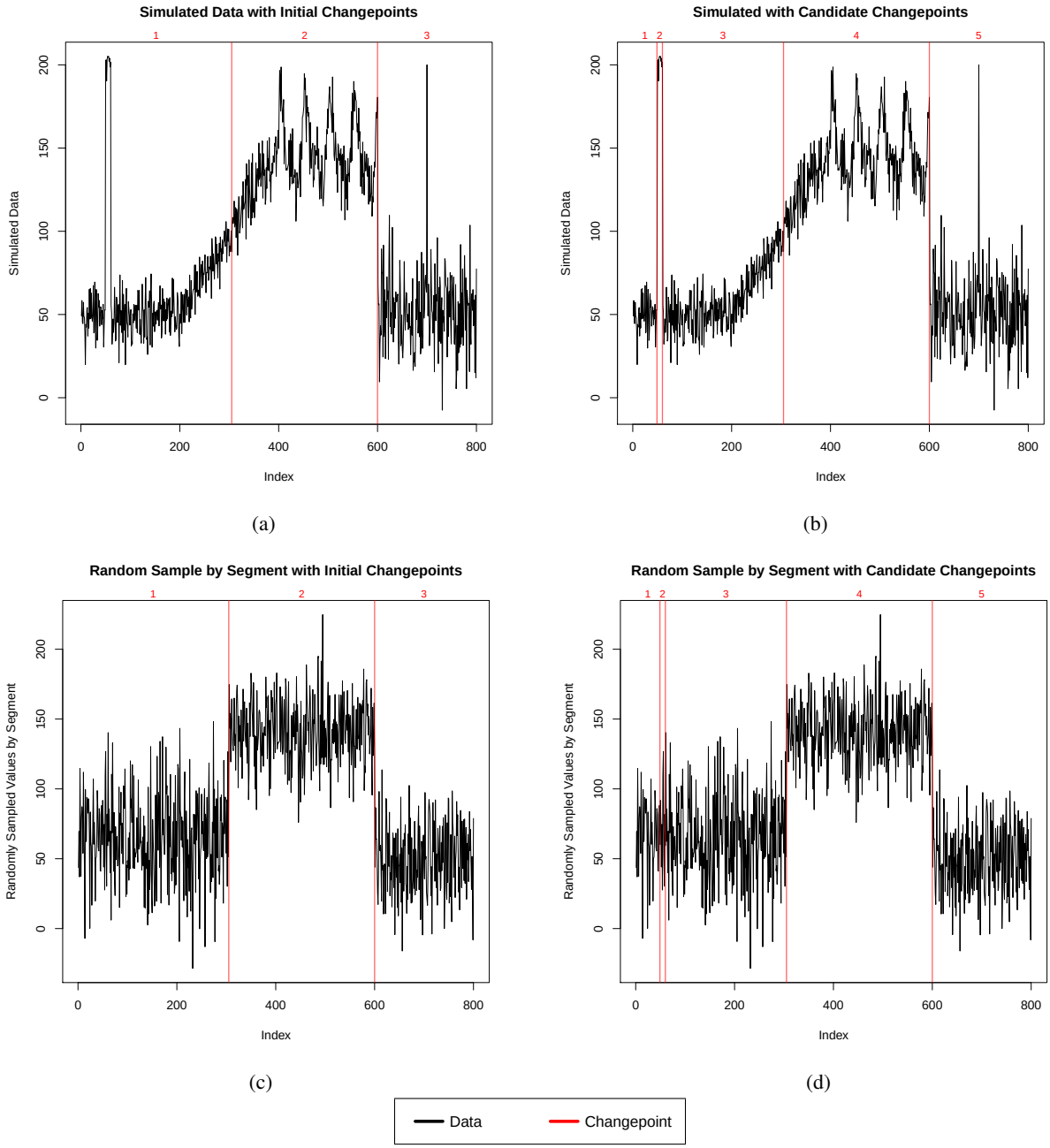


Figure 3: The process by which it is determined whether or not a new set of changepoints is significant. (a) We start with the simulated data with initial changepoints at indices 305 and 600. The log-likelihood, assuming a normal distribution for each segment, is -3727.3. (b) We then consider the simulated data set with the next set of 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. We must assess the significance of this change under the null, where the changepoints at 305 and 600 are all of the true changepoints in the series. To do this, we repeat the following steps many times. (c) We generate a Monte Carlo sample and impose the null changepoints at 305 and 600. Each segment's observations in this sample are randomly generated from a normal with a mean and standard deviation equal to the sample mean and sample standard deviation of the corresponding segment in subfigure a. The log-likelihood, assuming a normal distribution for each segment, is -3746.6. (d) We take the same Monte Carlo sample as from subfigure c, but now impose the 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 Monte Carlo sample under the null is therefore 1.6. This process of generating Monte Carlo samples is repeated a large number of times to generate an empirical null distribution of the change in the log-likelihood.

2.3 Stage 2: Trimming False Positives

Upon completing Stage 1 as described in Section 2.2, it is possible to have false positives. That is, we are primarily interested in adjusting for technological changepoints that induce a mean-shift in our data. However, in some cases, Stage 1 will identify changepoints within behaviorally-driven patterns, such as linear trends and seasonality, that implicitly includes many mean-shifts. For instance, a new government program that encourages the public to walk more may cause steps walked per day to increase linearly with time for multiple weeks across many individuals. This trend implicitly contains many mean-shifts, but they are behaviorally-driven and the trend will likely serve as a point of discussion for the investigator. In this case, the investigator would want to identify the trend as a whole for further study, rather than studying pieces of the same trend separately. 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. Such a phenomenon again implicitly contains many mean-shifts, but it should be studied as a whole and reported by the investigator. Needlessly breaking down an annual pattern by months would deprive the researcher of all the data comprising the pattern. Even if a pattern is a technological artifact, such as a slow roll-out of an update that induces a linear trend, correcting the pattern as a whole would utilize more data compared to correcting pieces of the pattern. Therefore, it is more appropriate to identify linear trends and seasonal patterns as a whole rather than decomposing them arbitrarily by their many mean-shifts. These changepoints that are initially identified by Stage 1 within linear trends and seasonal patterns are thus considered false positives. Other mean-shift changepoints, corresponding with a sudden mean-shift, or the start or end of a linear trend or seasonal pattern are our true changepoints of interest.

We will refer to this process of removing false positives as "trimming", but it is the same principle as "pruning" used by methods such as CBS [7]. 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 2.2, 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 changepoint. The second set consists of overall linear regression and harmonic regression fits across the two segments, effectively ignoring the changepoint. The linear model regresses the values in the time series against their indices in a simple linear regression. For harmonic regressions, we first estimate a segment's period using the frequency associated with the peak of the periodogram. The harmonic regression is then fit with a linear model taking this estimate to reflect the true period.

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

However, in cases when a 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 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 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 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 iteratively done for every 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.

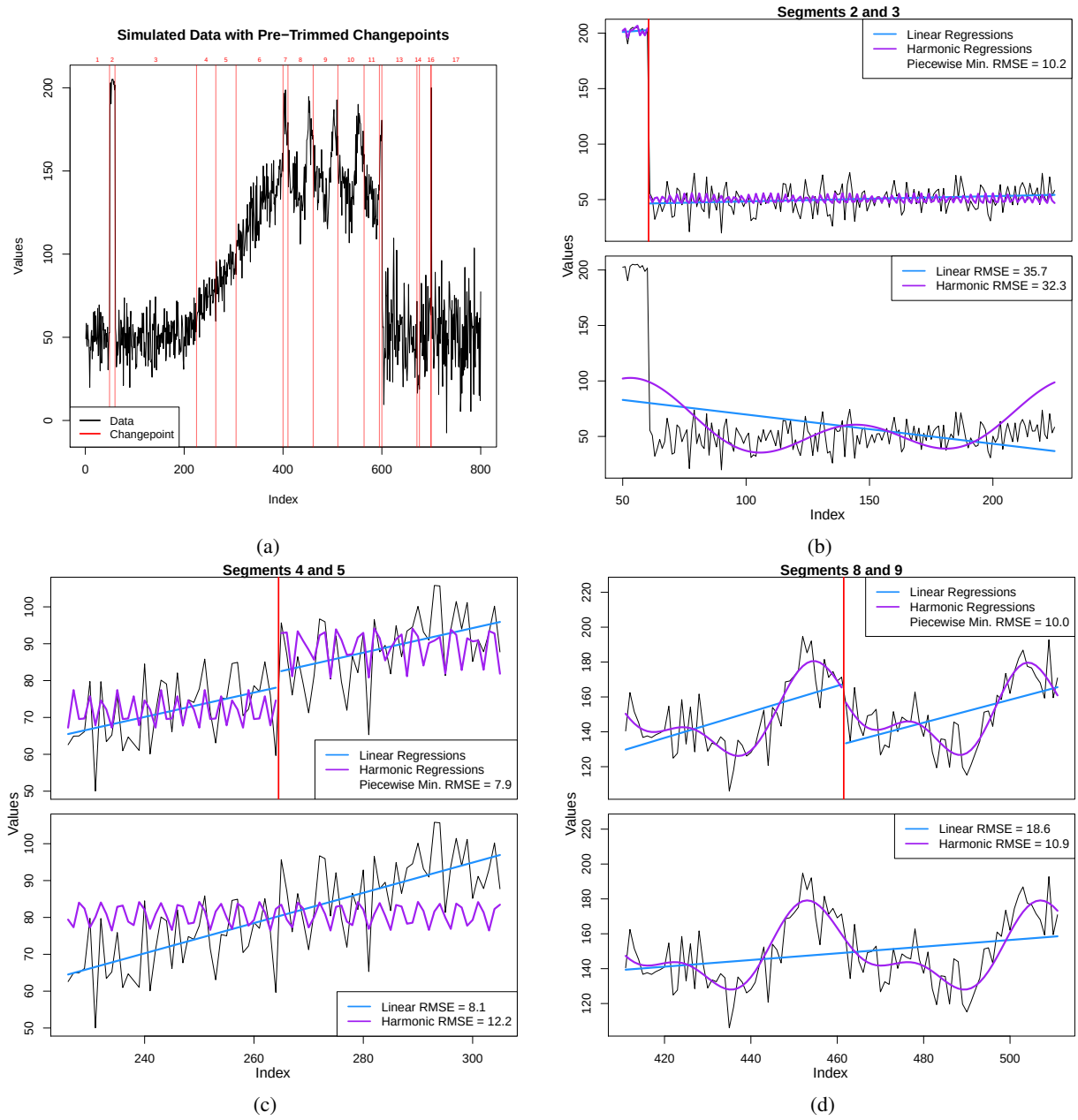


Figure 4: The process by which changepoints are trimmed. (a) We start with the simulated data set with an initial set of changepoints to be trimmed. This set of changepoints is selected for illustrative purposes and does not include all pre-trimmed changepoints from Figure 2. (b) We determine whether to trim the changepoint between segments 2 and 3, which is a true changepoint of interest. 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) In contrast, we can consider 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) Likewise, the changepoint between segments 8 and 9 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. In general, 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.

2.4 Comparison with Circular Binary Segmentation (CBS)

It is worth noting that ASCEPT shares some of the same principles as previous methods for changepoint detection and selection, such as CBS (for more information, see Section A of the Supplemental Material). However, there are also key differences important for mHealth research. For example, CBS uses a permutation test to obtain empirical p-values for changepoints [7]. However, using this permutation test, it is very difficult for CBS to capture segments containing only one observation. In fact, in the context of DNA copy numbers, which CBS was designed to analyze, single-point jumps are typically considered outliers that get smoothed. In contrast, ASCEPT's Monte Carlo procedure does not run into this same problem. While this Monte Carlo approach comes at the cost of a parametric assumption, mHealth data can truly exhibit single-point segments, as shown in Figure 1, highlighting the importance of capturing such segments.

ASCEPT shares a similar trimming principle to CBS as well, but they again differ in an important manner for mHealth. CBS trims, or "prunes", changepoints using 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 [7]. 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 for mHealth data, which are prone to these particular types of patterns.

To provide a benchmark comparison, we compare the results of ASCEPT on the simulated data and precision VISSTA data against those of CBS in Section 3.3.

2.5 Segment Correction

After identifying technological mean-shift changepoints, an investigator will need to make an adjustment or correction to account for these changepoints when performing downstream analysis. We perform a simple correction to the simulated data that demonstrates the importance of accurately identifying changepoints. A relatively simple approach to segment correction is to de-trend or de-seasonalize each segment and then transform each segment separately to a common scale. That is, we can iterate over segments, fitting constant, linear, and harmonic regressions to each. If we determine that the best fit is a linear or harmonic regression, we can de-trend or de-seasonalize that segment. Then, all segments are separately shifted and scaled to match the mean and standard deviation of a chosen reference segment. After doing so, if the true changepoints were accurately captured, then the resulting transformed series should appear to just be noise without any mean-shifts present, including trends and seasonality.

In our approach, a linear or harmonic regression is deemed the best fit to a segment only if the ratio of the constant fit's RMSE to the best corresponding linear regression or harmonic regression's RMSE is greater than a given threshold. This "fitting threshold" is similar to the trimming threshold discussed in Section 2.3. However, instead of comparing cross-segment linear and harmonic models to piecewise linear and harmonic models, we are now comparing linear and harmonic models for a single segment to the constant (i.e. mean) model for that segment. We choose this approach rather than using an overall F statistic because one will often find significant harmonic regressions on relatively long segments, even if those segments are solely noise. This likely arises in this context because we first estimate the segment's period using the frequency associated with the peak of the periodogram before fitting a harmonic regression. The harmonic regression is then fit with a linear model taking this estimate of the period to reflect the true period. In cases where the segment is purely noise, there is no true period. This can result in an overfit model that typically yields only a small improvement in RMSE over a mean model, despite being statistically significant.

After de-trending and de-seasonalizing those segments whose best fits are linear regressions and harmonic regressions, we can transform the residuals to the location and scale of a chosen reference segment. The location is taken to be the mean of the reference segment before any correction was performed and the scale is taken to be the standard deviation of that segment as estimated by the residual standard error for the best fitting model. We use this approach, rather than the sample standard deviation before correction, to avoid having trends and seasonality influence the standard deviation estimate.

2.6 Parameters

For all subsequent runs of ASCEPT on the simulated data and Precision VISSTA data in Section 3, Stage 1 of ASCEPT was run using an $\alpha = 0.01$ significance level and 10,000 random samples/simulations generated under the null at each step. Stage 2 was run using a trimming threshold of 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.

We show the results comparing ASCEPT to CBS in Section 3.3. We used a 0.01 significance level and 10,000 permutations for CBS as implemented in `R`'s `DNAcopy` package [12]. CBS's pruning threshold was set to 0.5 using the `undo.prune` argument within that package. For the particular data referred to here, this pruning threshold for CBS

appears comparable to the 1.2 trimming threshold for ASCEPT in terms of the total number of changepoints they yield per time series.

We also performed simple segment correction on the simulated data and present the results in Section 3.4. We used a fitting threshold of 1.75. That is, a linear trend or harmonic regression was accepted as the best fit if its RMSE was at least 75% higher than the constant fit's RMSE for that segment. Additionally, shifting and scaling was performed with respect to the seasonal segment from indices 401 to 600 as the reference segment.

2.7 R Package

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

3 Results

3.1 Results of ASCEPT on the Simulated Data

We first applied ASCEPT to the simulated data and found that Stage 1's empirical p-value process detected both true mean-shift changepoints of interest at indices 49, 60, 600, 699, and 700, and false positives within linear trends and seasonality, as shown in Figure 5a. However, Stage 2's trimming readily removed false positives and left detected changepoints at indices 49, 60, 225, 400, 600, 699, and 700, as shown in Figure 5a. Five of these correspond to sudden 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 simulated data experienced a mean-shift in their 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. A visual inspection of the plot can then be done in order to determine if a changepoint corresponds to a sudden mean-shift or the start/end of a trend or seasonal pattern

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. Any trimming threshold between 1.13 and 1.2 inclusive would yield the same final results on the simulated data. In contrast, using a trimming threshold of 1.21 or higher would trim out the changepoints initially detected at indices 699 and 700 during Stage 1 of ASCEPT, thereby introducing false negatives. Decreasing the trimming threshold to 1.12 or below would avoid trimming multiple changepoints initially detected within the seasonal pattern between indices 401 and 600 inclusive, thereby introducing false positives.

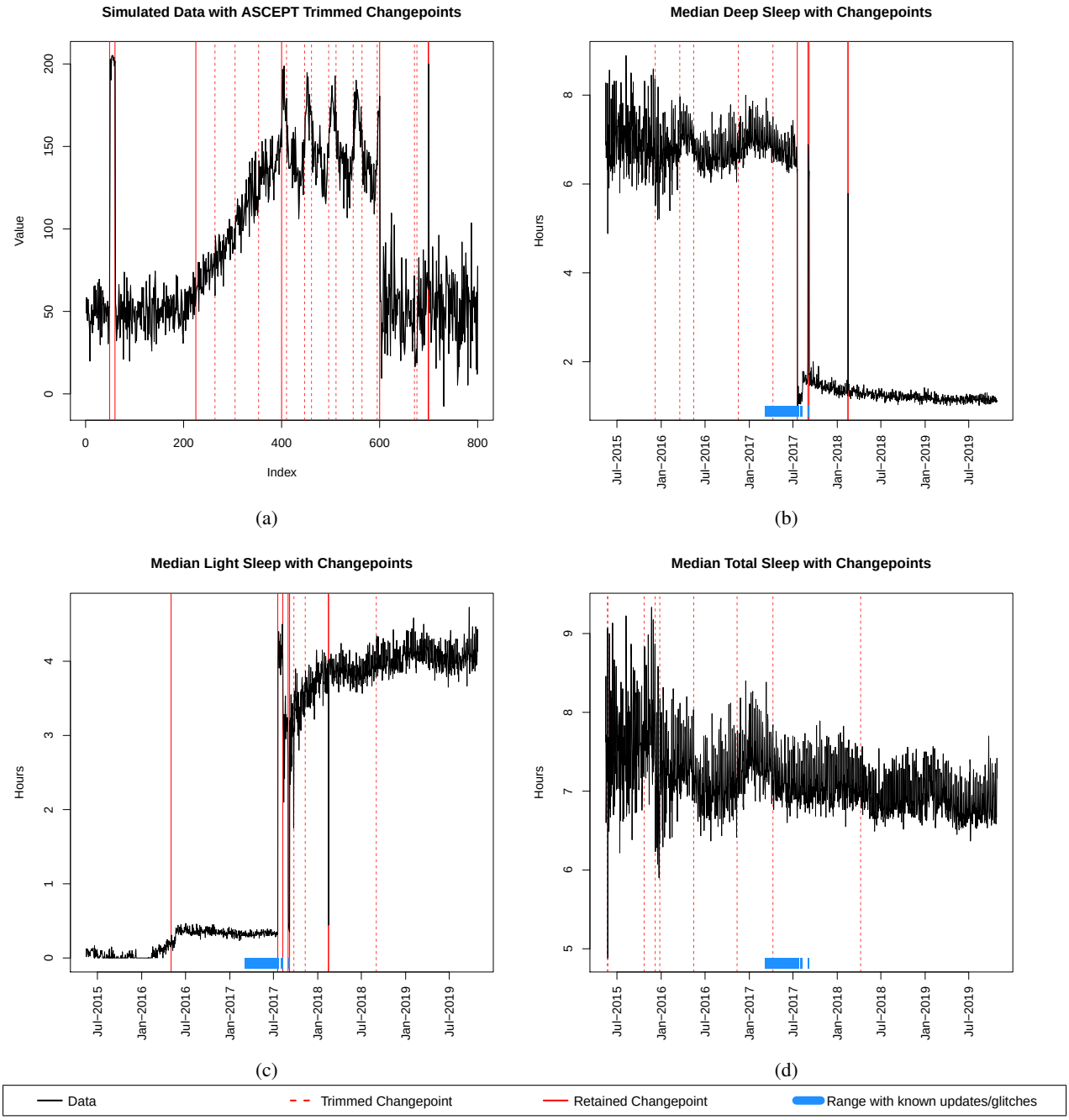


Figure 5: Overall results for (a) the simulated data, (b) median deep sleep, (c) median light sleep, and (d) median total sleep for users of newer Fitbit devices, multiple Fitbit devices, or an unknown Fitbit device. We used a 0.01 significance level, 10,000 Monte Carlo trials, and a 1.2 trimming threshold.

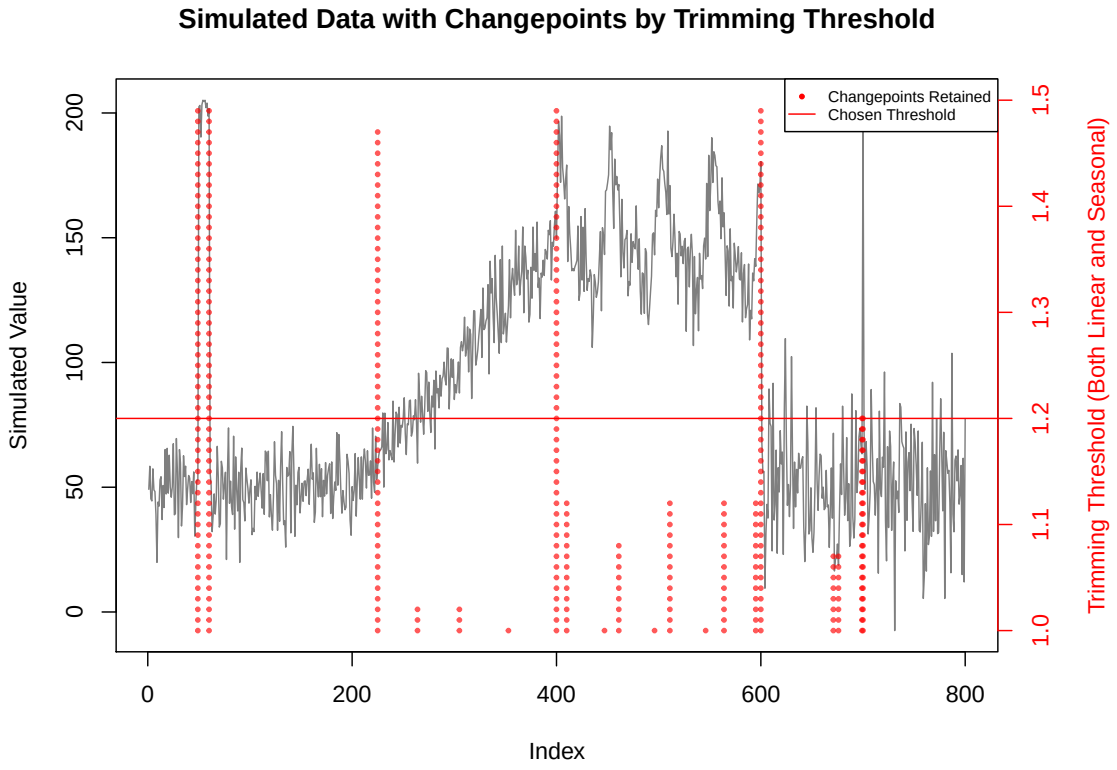


Figure 6: The simulated time series with changepoints initially detected using a 0.01 significance level and 10,000 Monte Carlo simulations, as indicated in Figure 5a, trimmed at various thresholds. Changepoints that are retained after the trimming process are indicated by a red point. At trimming thresholds 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 5a.

3.2 Results of ASCEPT on the Precision VISSTA Data

Next, we applied ASCEPT to real-world mHealth data from the Precision VISSTA study. In Figures 5b and 5c, the original data visually demonstrate how deep sleep and light sleep are complements of one another. Therefore, as we would hope, we found that running ASCEPT on median deep sleep and median light sleep yielded fairly comparable results. The first stage of ASCEPT yielded clear false positives due to trends and seasonality. However, trimming again primarily retained changepoints corresponding to true mean-shifts of interest. Both sets are changepoints are shown in Figures 5b and 5c. 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 experienced 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 experienced 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. It is important that we are able to detect such short changes in the context of mHealth. However, some comparable changepoint detection and selection approaches, such as CBS [7] are not designed for this context (for more information, see Sections 2.4 and 3.3).

We further assessed these changepoints by cross-referencing with online information. In particular, some firmware updates and glitches correspond with the changepoints observed here. Alta HR received firmware update 26.62.6 between August 1st, 2017 and August 10th, 2017 [13], corresponding with the August 10th, 2017 changepoint identified when analyzing light sleep. Likewise, Fitbit overhauled its calculation of sleep by introducing "Sleep Stages", starting on March 6th, 2017 [14]. Glitches with Sleep Stages were reported from within a week of release through July 24th, 2017 for Alta HR, Blaze, and Charge 2 devices [15], encompassing the mean-shift observed immediately after July 19, 2017 for both deep and light sleep. Blaze was again subject to glitches between September 3rd, 2017 and September 7th, 2017 [16], corresponding with the mean-shift immediately after September 6th, 2017 for both deep and light sleep. Compared to the ranges of dates found when searching for information online, ASCEPT provides a relatively effective manner of pinpointing exact times at which such changes occur.

Next, we used the daily median total sleep, depicted in Figure 5d, as a negative control. While there were some large changes in median total sleep during 2015, this was a period consisting of a large variance since there were as few as 7 individuals contributing data. Accordingly, ASCEPT did not identify any changepoints in this time series after trimming. 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 trimming threshold.

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

3.3 Results Comparison between ASCEPT and CBS

Since ASCEPT shares some of the same principles as CBS, we used CBS as a benchmark. We first compared ASCEPT and CBS on the simulated data and found that ASCEPT performed favorably. The results from each procedure are shown in Figure 7a. CBS failed to capture the single-point segment at index 700, while ASCEPT successfully did. ASCEPT also successfully segmented off the linear trend and seasonal pattern by selecting changepoints at indices 225, 400, and 600. In contrast, CBS roughly split the linear trend into fourths, segmenting the last portion of this linear trend with the seasonal pattern.

Next, we applied both ASCEPT and CBS to variables from the Precision VISSTA study. The results from running ASCEPT and CBS on deep sleep are shown in Figure 7b. The two procedures yielded nearly the same results, except that CBS failed to detect the single-day shift corresponding to February 15th, 2018. ASCEPT successfully identified this segment. Similar was true for most variables from the Precision VISSTA study. However, one variable on which the two procedures differed considerably was times woken. We show the results for the median number of times woken during the night in Figure 7c. CBS failed to capture multiple changepoints from late 2017 into early 2018 and failed to prune two false positives during the same time period. ASCEPT successfully captured the major changepoints and trimmed false positives within what appear to be trends and seasonality. 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.

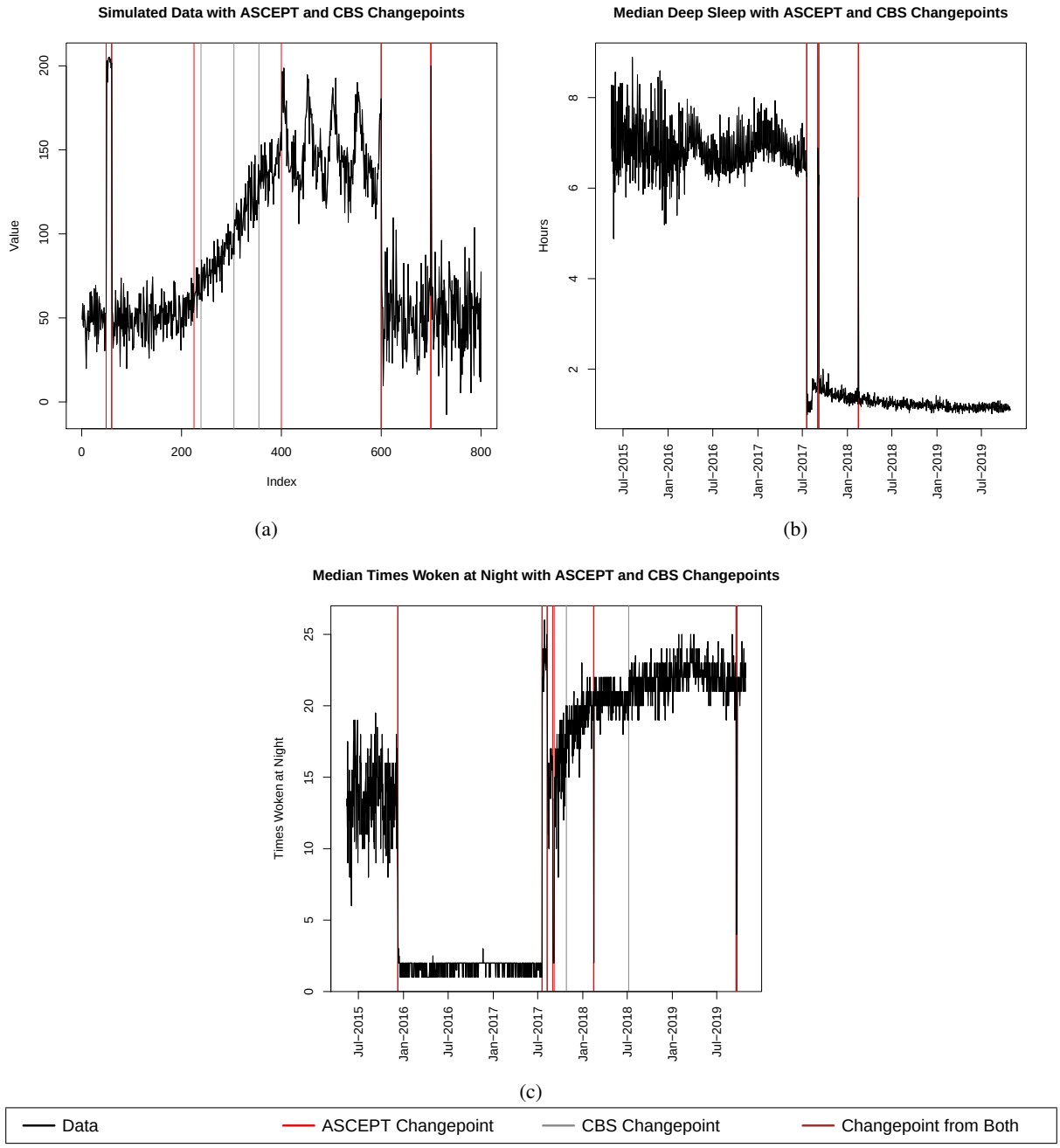


Figure 7: Comparison of results from ASCEPT with those from CBS for (a) the simulated data, (b) median deep sleep, and (c) median times woken during the night for users of new Fitbit devices, multiple Fitbit devices, or unknown Fitbit devices. We used a 0.01 significance level and 10,000 simulations or permutations for ASCEPT or CBS respectively. The trimming threshold for ASCEPT was 1.2 and the pruning threshold for CBS was 0.5.

3.4 Segment Correction Results for ASCEPT and CBS

Having performed changepoint detection and selection on the simulated data, we then adjusted the data for these identified changepoints. In Figure 8, we present the results after performing the simple segment correction described in Section 2.5 for the changepoints identified by each ASCEPT and CBS.

In this case, the ASCEPT segment-corrected series in Figure 8c largely appears to only contain noise with no clear mean-shifts, except for some residual trends around index 200. The CBS segment-corrected series in Figure 8d still contains clear trends, seasonality, and other mean-shifts due to the less accurate identification of changepoints. These incorrect changepoints made it more difficult to assess whether a linear or harmonic regression was preferred to a constant model on the corresponding segments. For instance, CBS identified a changepoint at index 355 rather than 400, causing part of the simulated data's linear trend to be segmented with its seasonal pattern. This led to a relatively poor fit for a harmonic regression, causing the constant fit's RMSE to be less than 1.75 times that of the harmonic regression. We therefore took the constant fit to be the best model for this segment, rather than a harmonic regression, at the 1.75 fitting threshold. We present the results for some additional fitting thresholds in Section C of the Supplemental Material.

4 Discussion and Conclusion

We have presented an approach for identifying changepoints that combines principles previously used with methods, such as a stopping criterion, with the current state-of-the-art 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 they arise within a linear 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 in its first stage, up to the constraint imposed by selecting a particular significance level. This is because PELT is an exact changepoint detection method, as opposed to an approximate method, like CBS, which only guarantees a locally optimal solution. 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 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. Thus, ASCEPT aims to utilize many of the advantages offered by PELT, while presenting them in a more accessible manner tailored towards processing mobile health data.

Additionally, ASCEPT offers advantages over researching technological changepoints based on information available online. For example, while Fitbit lists previous firmware versions online, it does not readily provide release dates or specific notes regarding each one [17]. A researcher needs to comb through community forums [18] to find more details. While we were not able to find posts corresponding with all of the exact dates of changepoints found in Section 3.2, 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.

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.

Despite these limitations, ASCEPT is generalizable to a wide variety of contexts moving forward. For instance, while the data studied here are from the Precision VISSTA study specifically, ASCEPT is applicable to mean-shift changepoint selection in any univariate time series for which linear trends and seasonality may give rise to false positives. This may apply to many different mobile health data sets, along with other time series in public health and medical contexts at large. Likewise, while the current presentation of ASCEPT uses PELT, the processes described here could, in theory, be applied to other changepoint detection algorithms. ASCEPT 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.

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

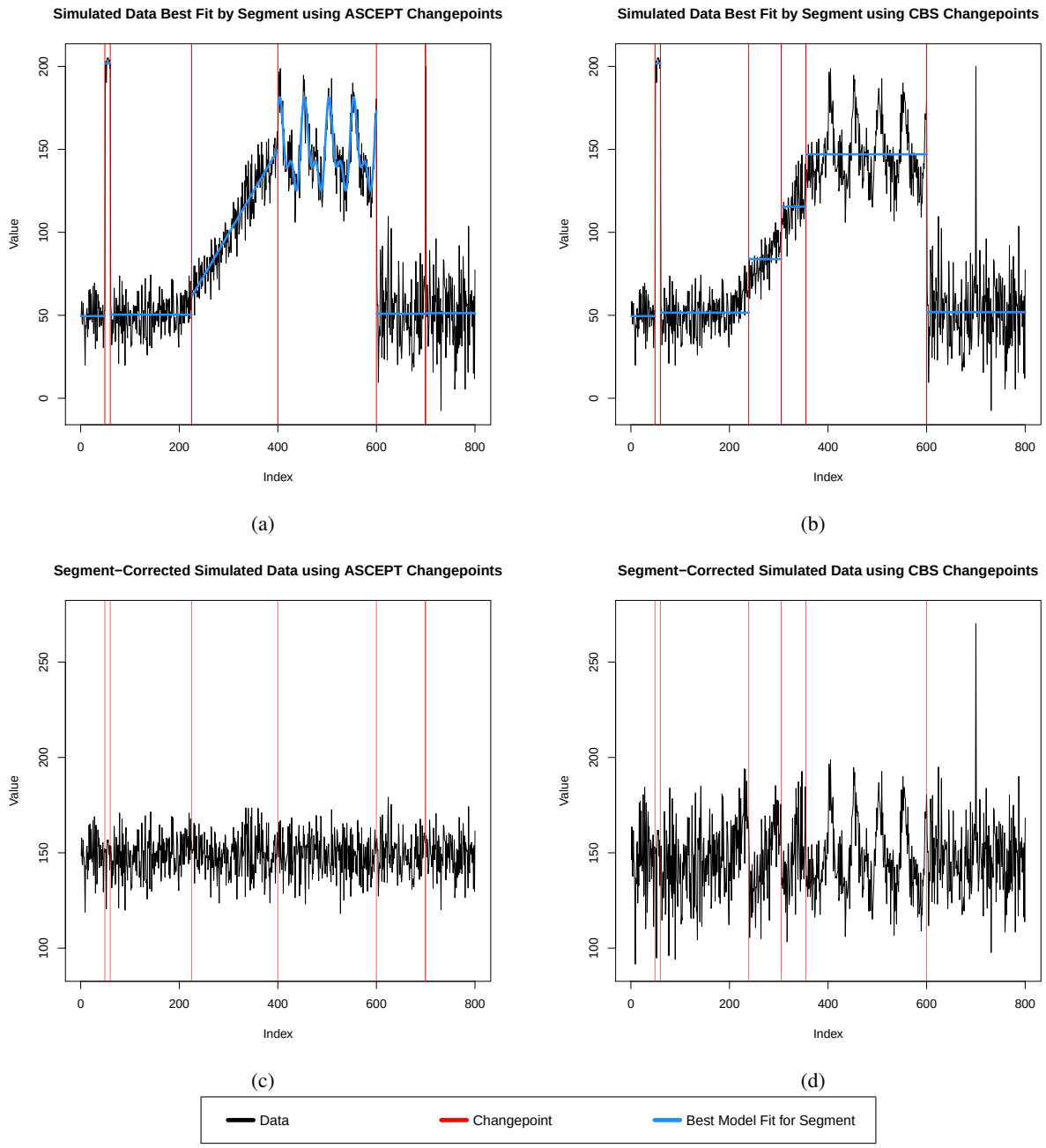


Figure 8: A demonstration of a simple correction process after identifying changepoints. Each segment is assessed for its best constant, linear, or harmonic fit using the changepoints from either ASCEPT or CBS, as shown in subfigures (a) and (b) respectively. If the best fit is a linear trend, it is then de-trended. If the best fit is a harmonic regression, it is then de-seasonalized. All segments are then shifted and scaled to match the mean and residual standard error associated with the seasonal segment from indices 401 to 600. The results of this process are shown in subfigures (c) and (d) for ASCEPT and CBS respectively. If changepoints were captured properly, there should be no remaining mean-shifts, including any trends or seasonality. We used a 0.01 significance level and 10,000 simulations or permutations for ASCEPT or CBS respectively. The trimming threshold for ASCEPT was 1.2 and the pruning threshold for CBS was 0.5. We used a 1.75 fitting threshold for segment correction for both.

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.

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] 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.

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

[9] 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.

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

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

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

[13] 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.

[14] 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.

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

[16] Received Classic Sleep rather than Sleep Stages. `https://community.fitbit.com/t5/Blaze/RESOLVED-9-3-Received-Classic-Sleep-rather-than-Sleep-Stages/td-p/2174227`, 2017. Accessed: 2020-06-11.

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

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

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

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

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

[22] 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 [19], 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 a 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 $t = \tau$ and the other from $t = \tau + 1$ 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) [7], which allows for detection of two changepoints at a time, and Wild Binary Segmentation (WBS) [20], which randomly draws and checks segments. Though CBS is approximate, we discuss how ASCEPT utilizes similar principles to it in Section 2. 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 comparable principles, but uses different implementations at each step.

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 [21] has $\mathcal{O}(mn^2)$ complexity for a time series of length n with m changepoints. Likewise, the Optimal Partitioning algorithm has $\mathcal{O}(n^2)$ computational complexity [22]. The method that we 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 [22], 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 value for the penalty 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 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. This is the primary motivation for ASCEPT.

B Additional Results for Precision VISSTA and CBS

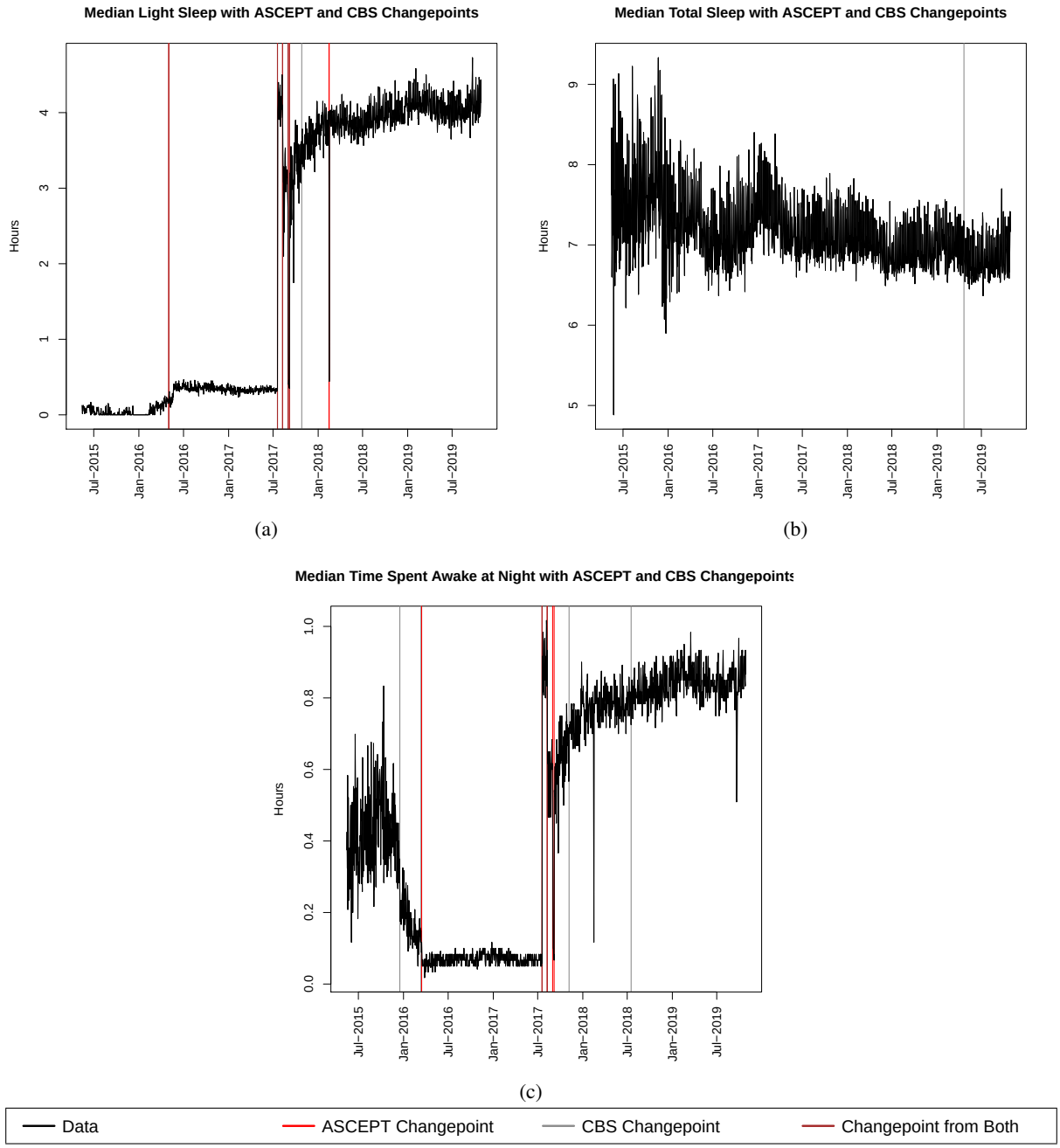
We now present the results from running ASCEPT on different data series from the Precision VISSTA study, excluding those that were presented in Section 3.3. In Supplemental Figures S1 and S2, we present the results for both ASCEPT and CBS, using the parameters specified in Section 2.6. We found that the results from ASCEPT are favorable both in identifying mean-shifts in the data, especially those lasting only one day, and in trimming false positives.

While ASCEPT performed very well on these various time series in general, the one major exception arose when investigating the median time spent awake at night, depicted in Supplemental Figure S1c. Here, both ASCEPT and CBS missed four true changepoints of interest. In the case of ASCEPT, reducing the trimming threshold to 1.15 would capture two of these changepoints. Interestingly, this variable is nearly identical to times woken during the night, on which ASCEPT performed very well, shown in Figure 7c. 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.

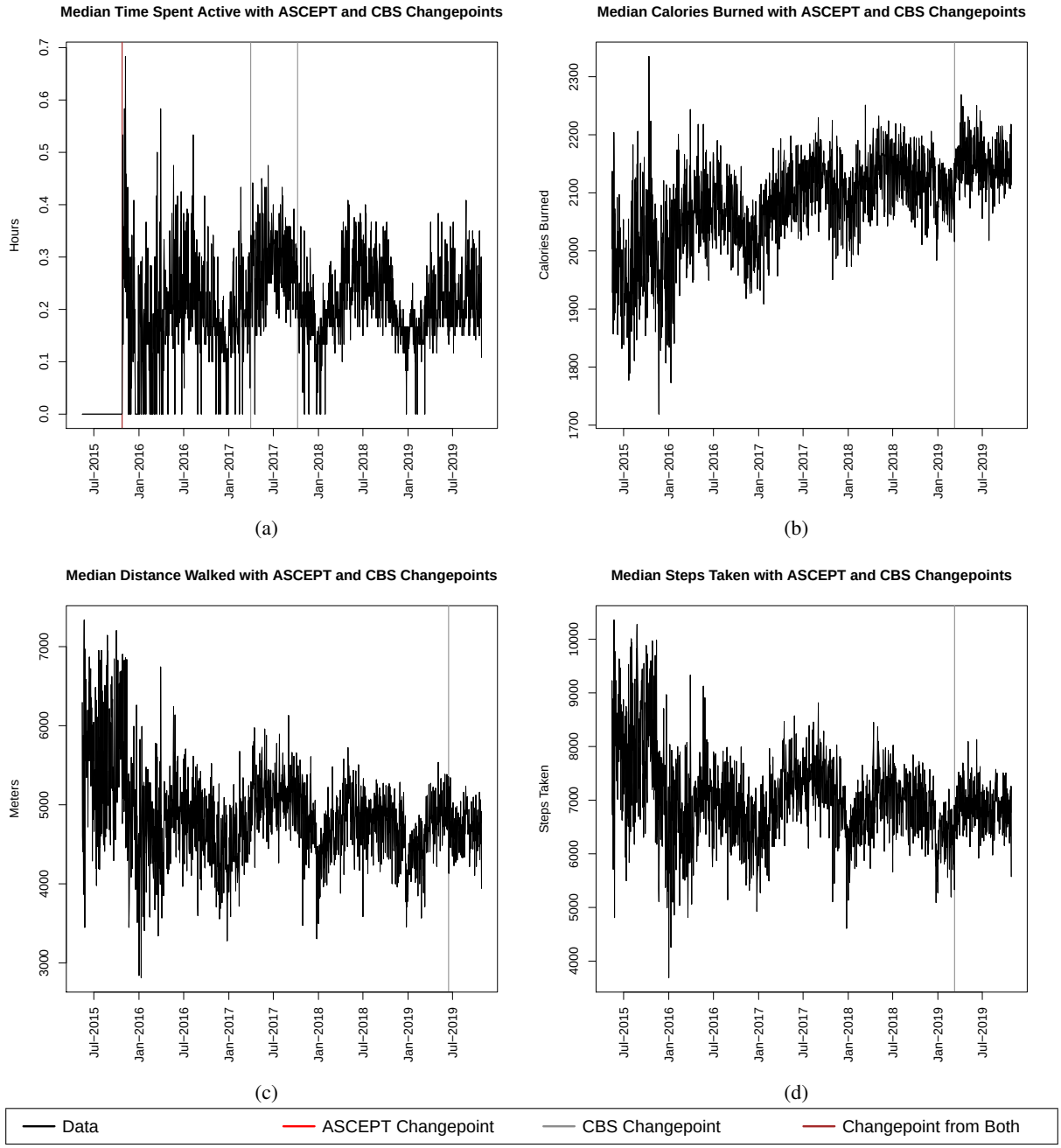
C Additional Results for Segment Correction

We now show additional results pertaining to Section 3.4 on segment correction. In Section 3.4, we used a fitting threshold of 1.75. A linear or harmonic regression was deemed the best fit to a segment only if the ratio of the constant fit's RMSE to the best corresponding linear regression or harmonic regression's RMSE was greater than this fitting threshold. We present the results when using 1.50 and 1.25 as fitting thresholds instead in Supplemental Figures S3 and S4 respectively. Using the ASCEPT changepoints, the results did not change appreciably. For CBS, the trends and seasonality were more appropriately identified for smaller fitting thresholds, but there are still clear residual mean-shifts, including trends between indices 201 and 400, and the single-point segment at index 700.

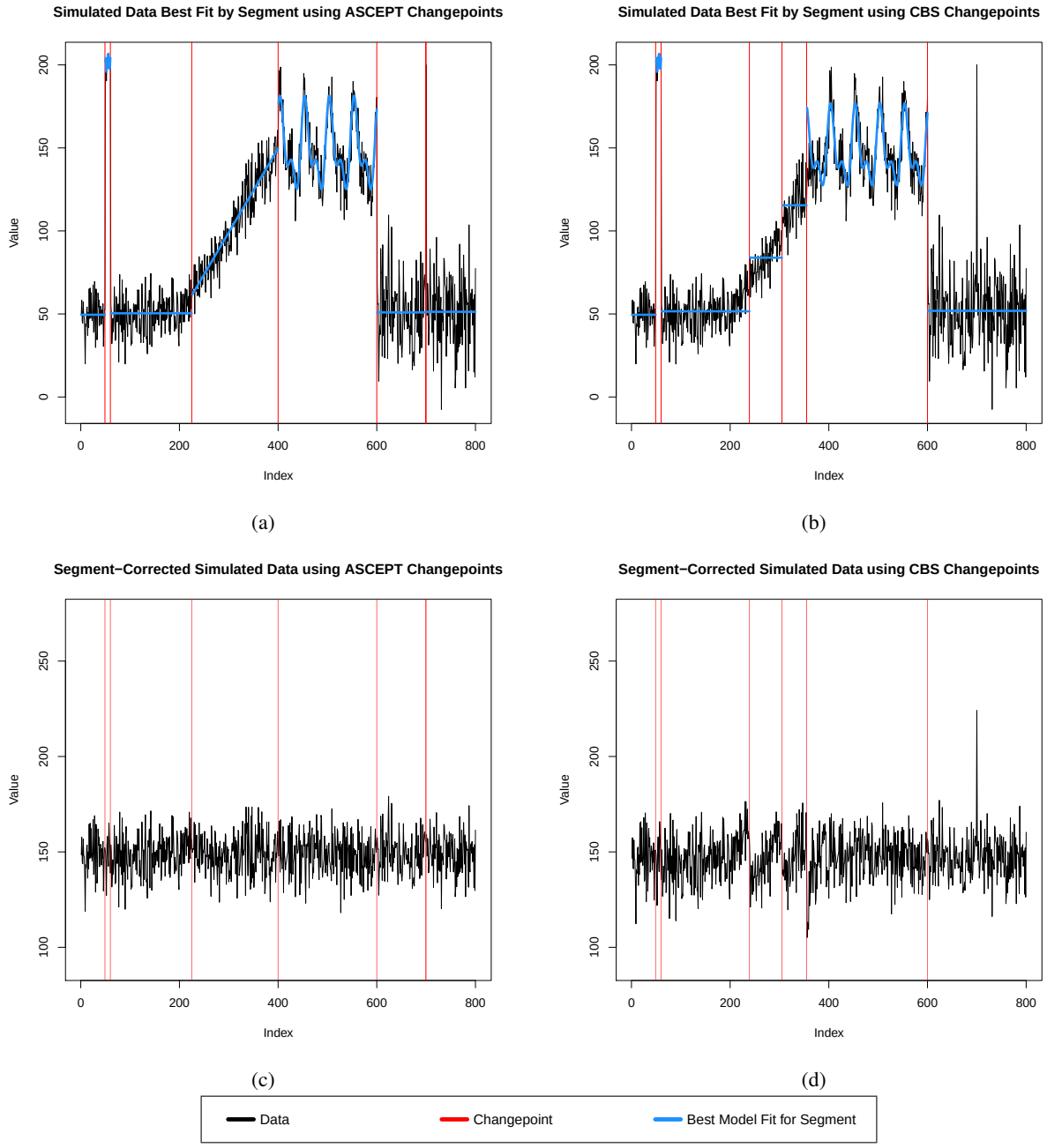
It is also worth noting that for these fitting thresholds of 1.50 and 1.25, the best fitting model for the seasonal segment from indices 401 to 600 is now a harmonic regression using CBS changepoints. Since all segments are scaled to match the residual standard error of this chosen reference segment, the corrected series based on CBS changepoints in Supplemental Figures S3d and S4d have smaller spreads than that shown in Figure 8d, where the best fitting regression was constant for that seasonal segment.



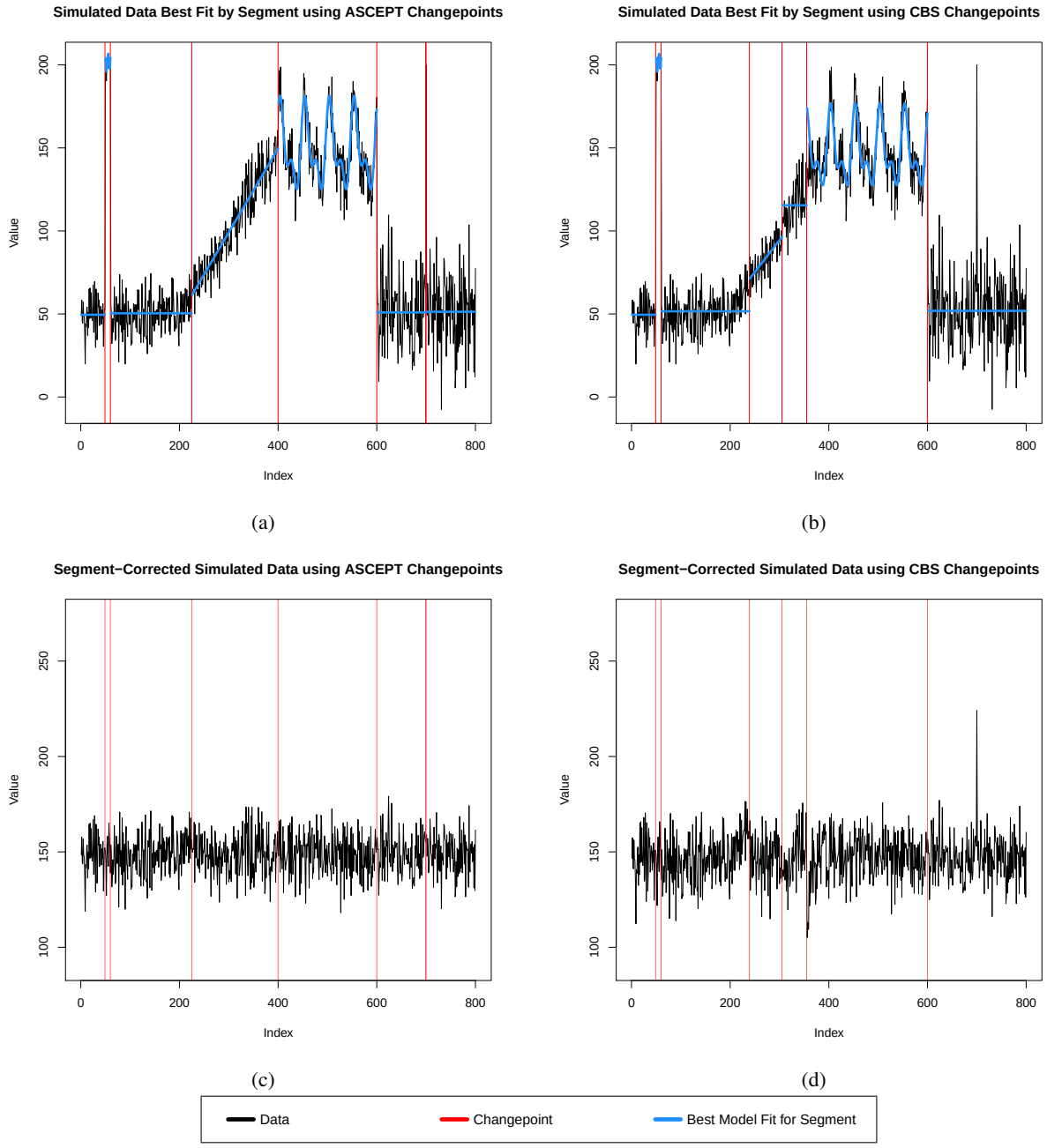
Supplemental Figure S1: Comparison of results from ASCEPT with those from CBS for (a) median light sleep, (b) median total sleep, and (c) median time spent awake at night for users of new Fitbit devices, multiple Fitbit devices, or unknown Fitbit devices. We used a 0.01 significance level and 10,000 simulations or permutations for ASCEPT or CBS respectively. The trimming threshold for ASCEPT was 1.2 and the pruning threshold for CBS was 0.5.



Supplemental Figure S2: Comparison of results from ASCEPT with those from CBS for (a) median time spent active, (b) median calories burned, (c) median distance walked, and (d) median steps for users of new Fitbit devices, multiple Fitbit devices, or unknown Fitbit devices. We used a 0.01 significance level and 10,000 simulations or permutations for ASCEPT or CBS respectively. The trimming threshold for ASCEPT was 1.2 and the pruning threshold for CBS was 0.5.



Supplemental Figure S3: A demonstration of a simple correction process after identifying changepoints. Each segment is assessed for its best constant, linear, or harmonic fit using the changepoints from either ASCEPT or CBS, as shown in subfigures (a) and (b) respectively. If the best fit is a linear trend, it is then de-trended. If the best fit is a harmonic regression, it is then de-seasonalized. All segments are then shifted and scaled to match the mean and residual standard error associated with the seasonal segment from indices 401 to 600. The results of this process are shown in subfigures (c) and (d) for ASCEPT and CBS respectively. If changepoints were captured properly, there should be no remaining mean-shifts, including any trends or seasonality. We used a 0.01 significance level and 10,000 simulations or permutations for ASCEPT or CBS respectively. The trimming threshold for ASCEPT was 1.2 and the pruning threshold for CBS was 0.5. We used a 1.50 fitting threshold for segment correction for both.



Supplemental Figure S4: A demonstration of a simple correction process after identifying changepoints. Each segment is assessed for its best constant, linear, or harmonic fit using the changepoints from either ASCEPT or CBS, as shown in subfigures (a) and (b) respectively. If the best fit is a linear trend, it is then de-trended. If the best fit is a harmonic regression, it is then de-seasonalized. All segments are then shifted and scaled to match the mean and residual standard error associated with the seasonal segment from indices 401 to 600. The results of this process are shown in subfigures (c) and (d) for ASCEPT and CBS respectively. If changepoints were captured properly, there should be no remaining mean-shifts, including any trends or seasonality. We used a 0.01 significance level and 10,000 simulations or permutations for ASCEPT or CBS respectively. The trimming threshold for ASCEPT was 1.2 and the pruning threshold for CBS was 0.5. We used a 1.25 fitting threshold for segment correction for both.