1-exercise

Xavier Bryant

30/03/2021

Question 1

Problem Description

Assess S claims of battery failure from temperatures from a sample of previous batteries that experienced failure and from a sample of past battery temperatures in general.

a. Perform a kernel density estimation for temps-7 and temps-other using what you consider is the most adequate bandwidth. Since the temperatures are positive, is it required to perform any transformation?.

b. Is there any important difference on the results from considering the LSCV selector over the DPI selector?

c. It seems that in temps-7 there is a secondary mode. Compute a kernel derivative estimation for temps-7 and temps-other using what you consider are the most adequate bandwidths.

d. Precisely determine the location of the extreme points

e. Check with a kernel second derivative that the extreme points are actually modes.

Preliminary Work

We see the summary of the "Problematic phone" series, phones that have burned due to excessive temperature, and then the "Past phone" series of the working temperature of general phones in the past. We notice that the problematic phones have a much higher maximum value but a similar median. This leads us to suspect that phones that do burn up, do more often, have higher temperatures that lead to their failure.

```
## [1] "Problematic phones"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.85   12.51   15.68   17.34   19.52   62.84
```

```
## [1] "Past phones"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.39   12.69   17.22   17.76   22.01   46.86
```

We can see comparing the histograms below that, as we saw in the summaries, the problematic phones have much more disperse temperatures and a longer tail in the direction of higher temperatures. This indeed leads us to suspect that phones that do fail have a section of their series that have higher temperatures. However, phones that fail have temperatures that are normally, for most of the sample, in line with those of the general past phones.

We see in out initial kernel densities with default bandwidths that there is a tail on the problematic series with a second mode for the high temperatures. There is also some distortion at the mode of the past battery curve, but the tail of high temperatures is not present.

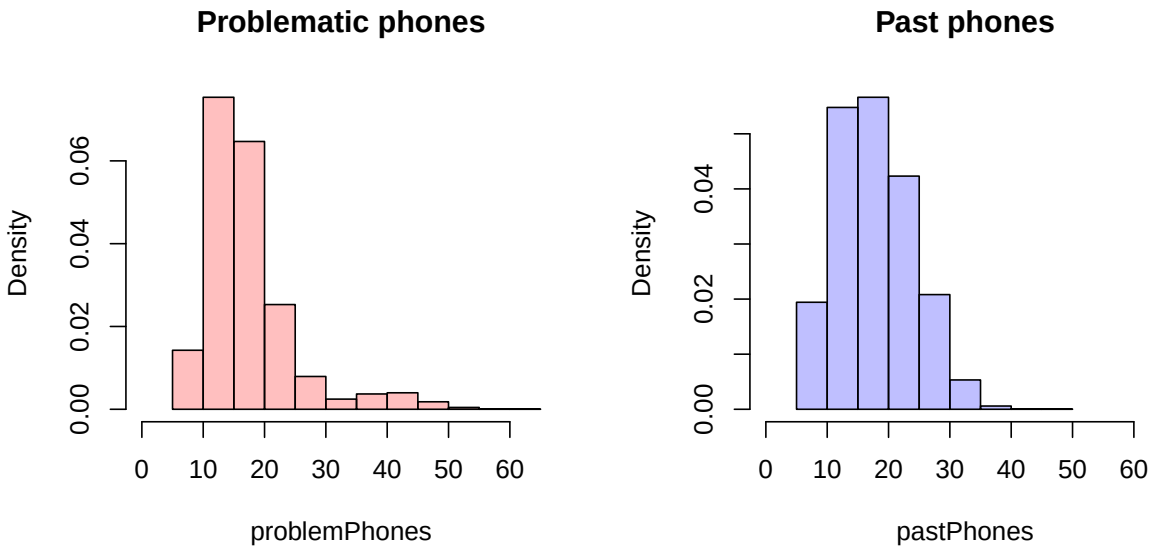

Figure 1: Comparison of histograms: Default bandwidths

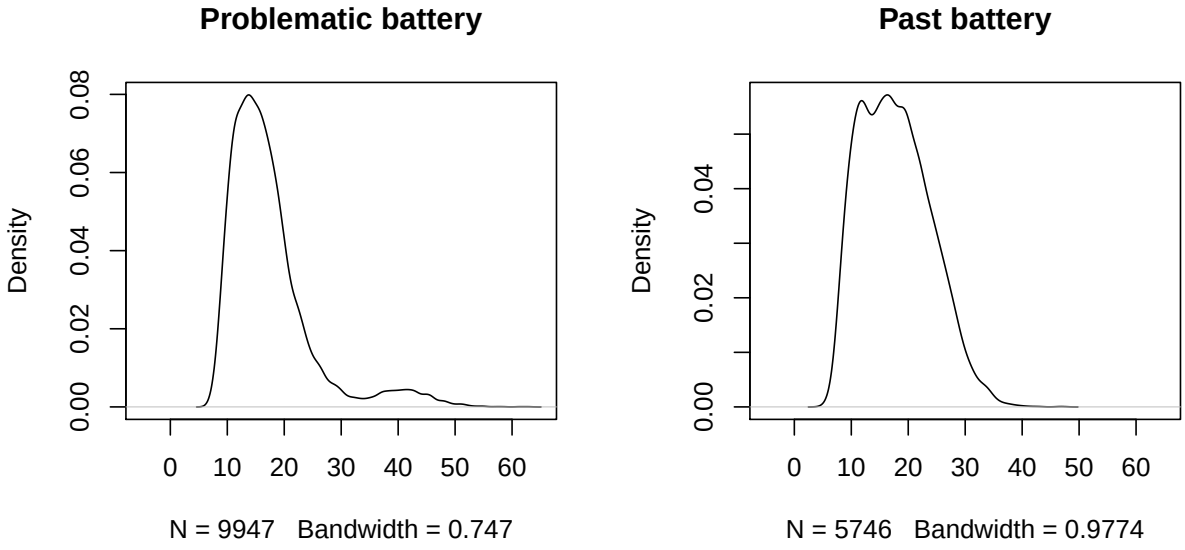

Figure 2: Comparison of densities; Default bandwidths

Part a

Problem Description

Perform a kernel density estimation for temps-7 and temps-other using what you consider is the most adequate bandwidth. Since the temperatures are positive, is it required to perform any transformation?.

Results

We can see these are the default bandwidths that are provided for the density estimates shown above.

```
## [1] "Problematic Temperatures"
## [1] 0.7469943
## [1] "Past Temperatures"
## [1] 0.9773738
```

Transformations Transformations must be performed in order to avoid boundary bias, which is to assign probability, with the function mapping, when there is no real density to support it or can be bias towards zero. We are looking at temperatures that can take values close to zero or very high values, therefore this would make us consider a log transformation. Below in Figure 3 and Figure 4, we do not see the shape of the curve change substantially - there are not sections that are mapped by the function but not supported by any density - nor do we see much boundary bias. However, the log transformation does further smooth our data - which for our analysis comparing the overall distribution of batteries - is helpful. This is due to the smaller bandwidth that is enabled due to the transformation, which also advantageous due to asymptotic properties. However, the untransformed series functions as well for our purposes, as a more precise curve is not necessarily important for our general type of analysis, and we don't see significant bias. Therefore, we proceed with the regular series.

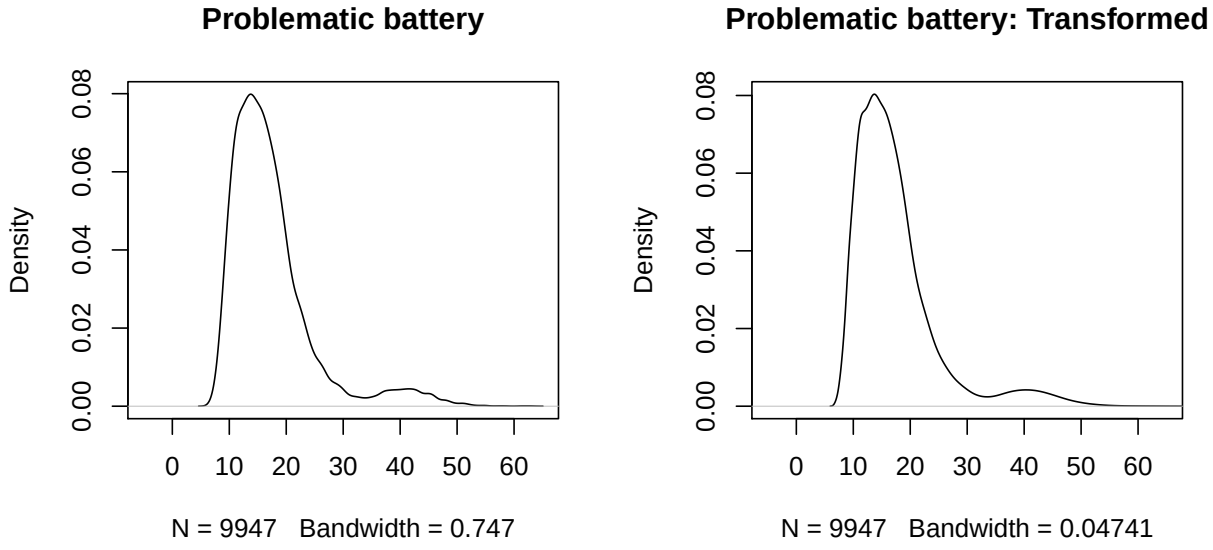

Figure 3: Comparisono default and transformed densities for Problematic series

Bandwidth Selection We will use several techniques to select the optimal bandwidth. The first technique is "rule of thumb" bandwidth selection derived from minimizing the AMISE with a normal parametric assumption for the unknown $R(f'')$, in zero stage part of the AMISE (asympotic mean integrated squared

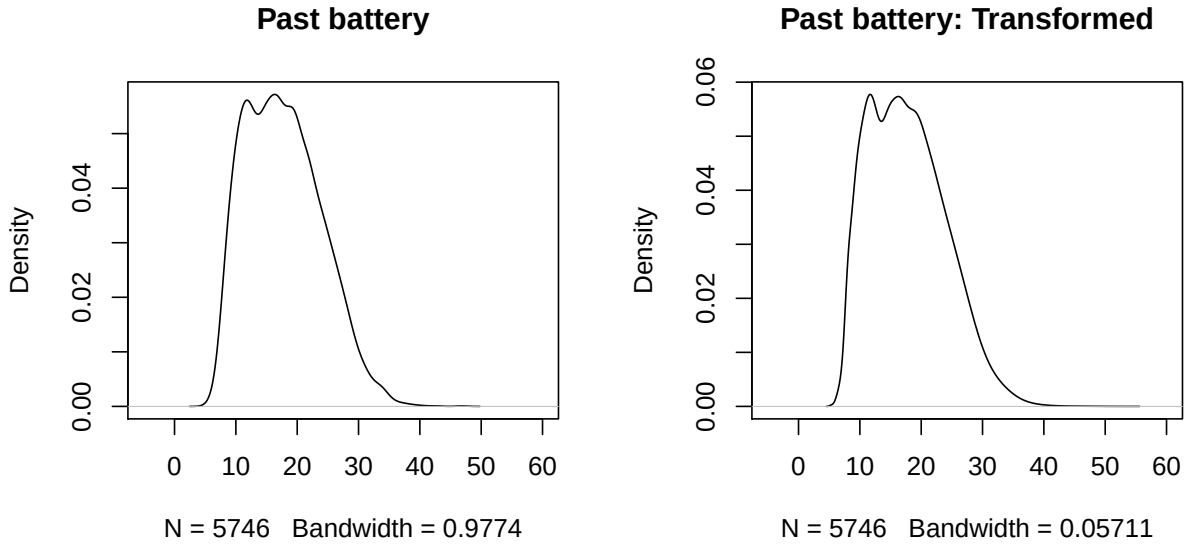

Figure 4: Comparison of default and transformed densities for Past series

error) derivation: $h_{AMISE} = [\frac{R(K)}{\mu_2^2(K)R(f'')n}]^{\frac{1}{5}}$. We then look at the Direct Plug In (DPI) selector, which instead of using a normal parameterization in the zero stage, the DPI buries the parametric assumption, usually, with two stages which balances between bias (lower with number of stages) and variance (higher with the number of stages). We then look at two cross-validation techniques, where we attempt to minimize the MISE (mean integrated squared error), through means of cross validation utilizing leave-one-out techniques. The two types of CV analysis are: the Least Squared Cross-Validation (or Unbiased Cross Validation) and the Biased Cross Validation. Numerical optimization is required for the LSCV and therefore we can get trapped in spurious solutions, necessitating the limiting of the bandwidth grid for the $\hat{h}_{LSCV}$. The BCV adapts a hybrid strategy estimating a modification of $R(f'')$ with leave-out-diagonals. It is important that the plot or the grid is reviewed in this selector as well as we are looking for the local minimizer not the global as h, the bandwidth , goes to infinity. We ensure to limit the bandwidth grid for both these functions, when necessary, from the function shown in class.

First we see the bandwidths of the "rule of thumb" selector:

```
## [1] "Past temperatures"
## [1] 1.151129
## [1] "Problematic temperatures"
## [1] 0.8797933
```

Second we see the DPI bandwidths:

```
## [1] "Past temperatures"
## [1] 0.8413892
## [1] "Problematic temperatures"
## [1] 0.6307256
```

Third, we have the LSCV bandwidths:

```
## [1] "Past temperatures"
```

```
## [1] 0.64052
```

```
## [1] "Problem temperatures"
```

```
## [1] 0.6555154
```

Lastly we have the BCV bandwidths:

```
## [1] "Past temperatures"
```

```
## [1] 0.8766243
```

```
## [1] "Problematic temperatures"
```

```
## [1] 0.6149622
```

To summarize we have:

Past temperature bandwidths: - RT: 1.15 - DPI: 0.84 - LSCV: 0.64 - BCV: 0.88

Problematic temperature bandwidths: - RT: 0.87 - DPI: 0.63 - LSCV: 0.66 - BCV: 0.61

Reviewing our results, we see that that the RT gives us bandwidths that are quite large in comparison with the other bandwidth selectors, which is common for non-normal data like our own. From the literature, we know from the DPI selector has a convergence rate that is much faster than the cross-validation technique, and therefore is dominant among academics. We also note that the BCV tends to be more bias, but have substantially less variance than the LSCV, tending to have larger bandwidths. However, we notice that the default values for bandwidth of 0.75 for the problematic series and, for the past series, 0.98 are fairly similar and, considering part b, we decide to use the default bandwidths as, for our purpose of comparing the series at its entirety, the differences are not important.

Part b

Problem Description

Is there any important difference on the results from considering the LSCV selector over the DPI selector?

Results

Bandwidths for LSCV and BCV:

Past temperature bandwidths: - DPI: 0.84 - LSCV: 0.64

Problematic temperature bandwidths: - DPI: 0.63 - LSCV: 0.66

We can see that the bandwidths of LSCV and DPI for the past temperatures differ significantly more than those for the problematic temperatures. The difference between the bandwidths is only 0.03 for the problematic temperatures. It is 0.2 for the past temperatures. In the literature, the DPI has much higher convergence rate than cross-validation methods, like LSCV, however, with very volatile or non-normal data the DPI may over smooth. We may be seeing this in practice as the LSCV does have a smaller bandwidth that is capturing more of disruption around the mode in the past temperatures series. The sample size is also considerably larger in the problematic temperature series so this may allow the two techniques to more closely approach one another. For functional purposes, they are fairly equivalent to us as we are comparing the overall past and problematic series, and both bandwidths more or less find similar curves.

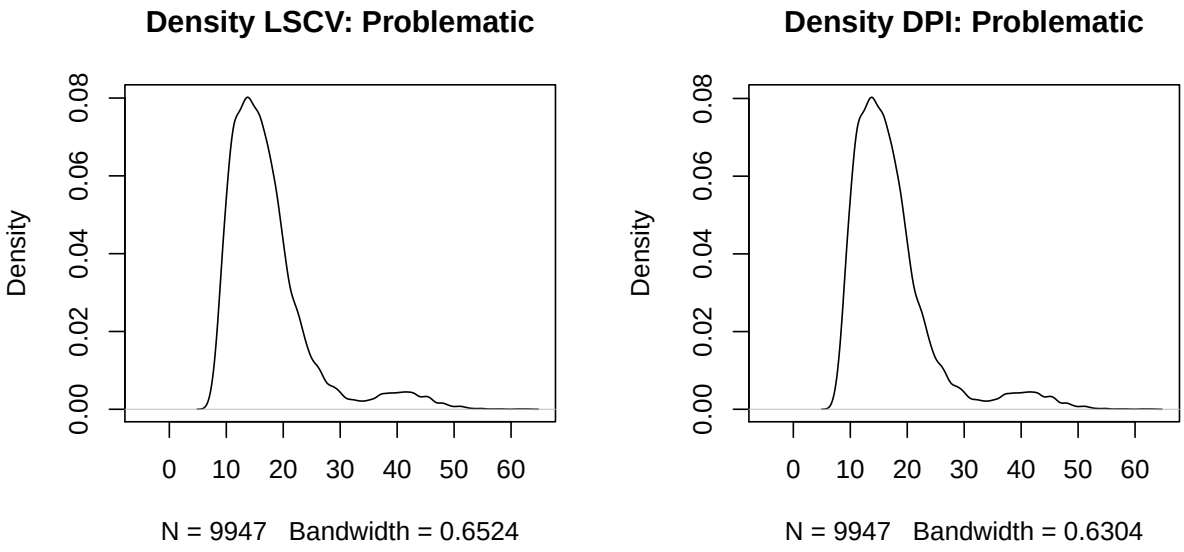

Figure 5: LSCV and DPI densities problematic series

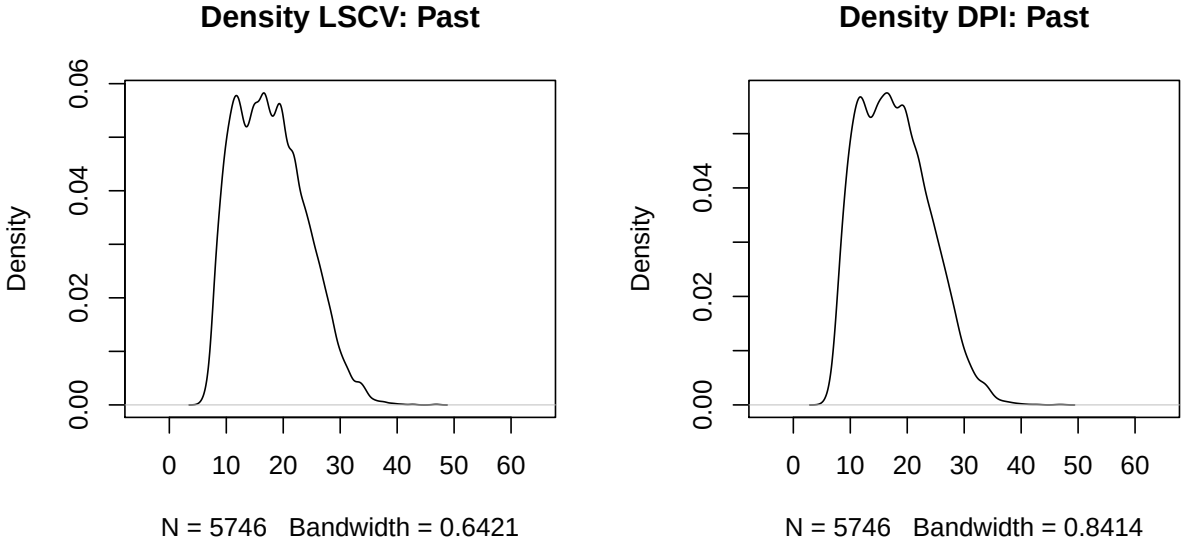

Figure 6: LSCV and DPI densities past series

Part c

Problem Description

It seems that in temps-7 there is a secondary mode. Compute a kernel derivative estimation for temps-7 and temps-other using what you consider are the most adequate bandwidths.

Results

We can see indeed that there is two modes in the problematic series. This second mode is key for our analysis as it likely shows that this mode of high temperatures could be largely responsible for the failure of these cellphones rather than those in general, or the past temperature series. There is further evidence they are modes as the the first derivative crosses the x-axis at these points. We also see a minimum in red.

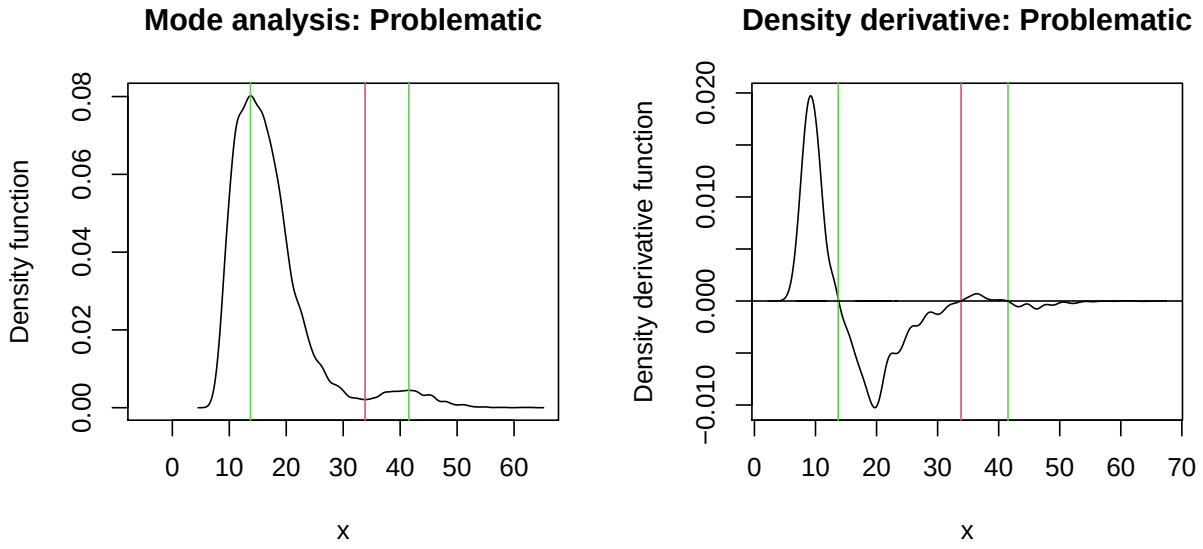

Figure 7: Mode and derivative estimation analysis: Problematic series

We can see the mode identified in the past phone series with the green line. We can see that the first derivative crosses the x-axis several times reflecting the disturbance we have around the mode of the series. However, we can still see that a majority of the data falls in the region around 10 and 20. This general idea is suitable for our analysis.

Part d

Problem Description

Precisely determine the location of the extreme points.

Results

We can see the extreme points of the problematic series for the two modes (13.72, 0.08) and (41.53, 0.01) as well as the minimum point (33.83, 0.002). For the past series, we can see the mode at (31.36, 0.007).

```
## [1] "Problematic temperatures"
```

```
## [1] "Mode 1: x-value"
```

```
## [1] 13.72123
```

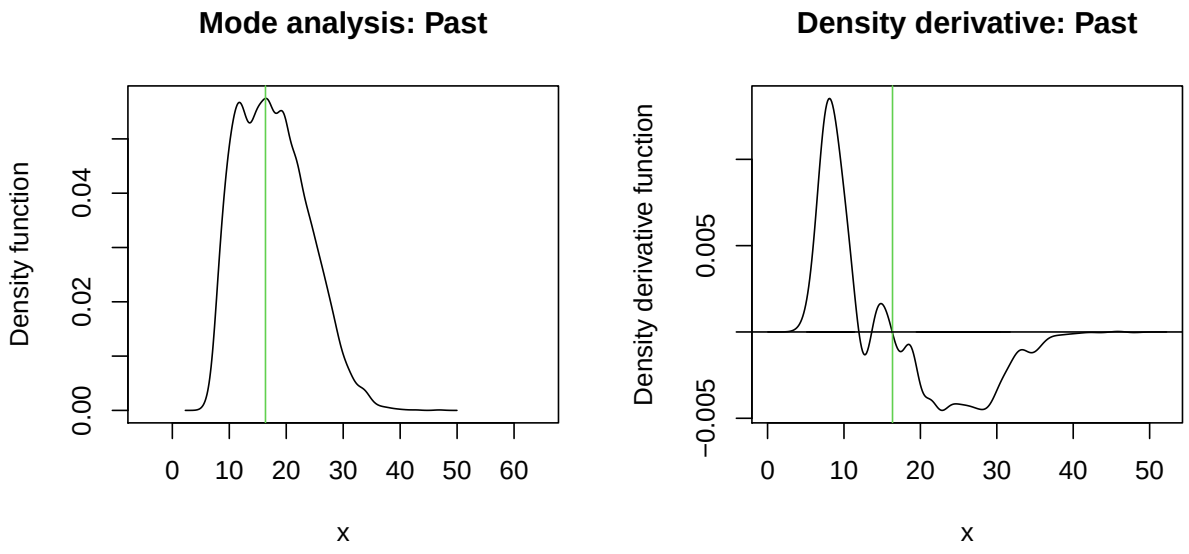

Figure 8: Mode and derivative estimation analysis: Past series

```
## [1] "Mode 2: x-value"
## [1] 41.53123
## [1] "Minimum: x-value"
## [1] 33.83911
## [1] "Mode 1: y-value"
## [1] 0.0798685
## [1] "Mode 2: y-value"
## [1] 0.004438786
## [1] "Minimum: y-value"
## [1] 0.002133435
## [1] "Past temperatures"
## [1] "Mode 1: x-value"
## [1] 31.35863
## [1] "Mode 1: y-value"
## [1] 0.006814199
```

Part e

Problem Description

Check with a kernel second derivative that the extreme points are actually modes.

Results

We can confirm that the points of the modes for the problematic series are indeed the modes as the they remain on the opposite side of the second derivative estimator. A mode should be on the negative side of the second derivative and a minimum should be on the positive side. We see this is true for both the modes and minimum of the problematic series.

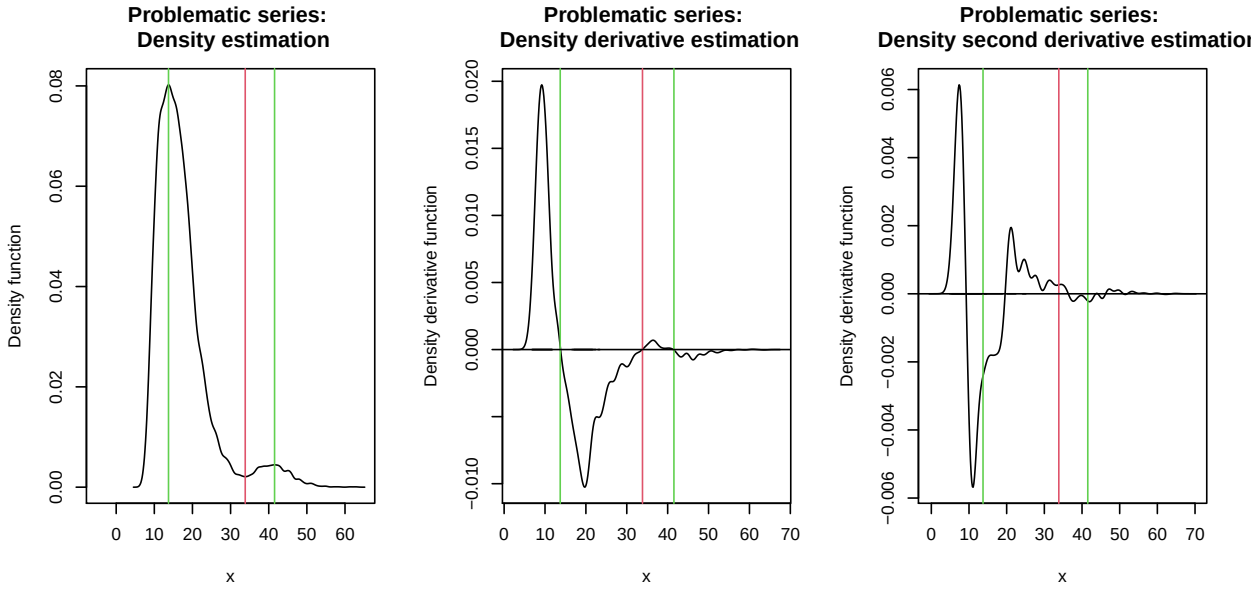

Figure 9: Second derivative analysis: Problematic series

We see this is also true for the past series, where the mode is on the negative side of the second derivative estimator.

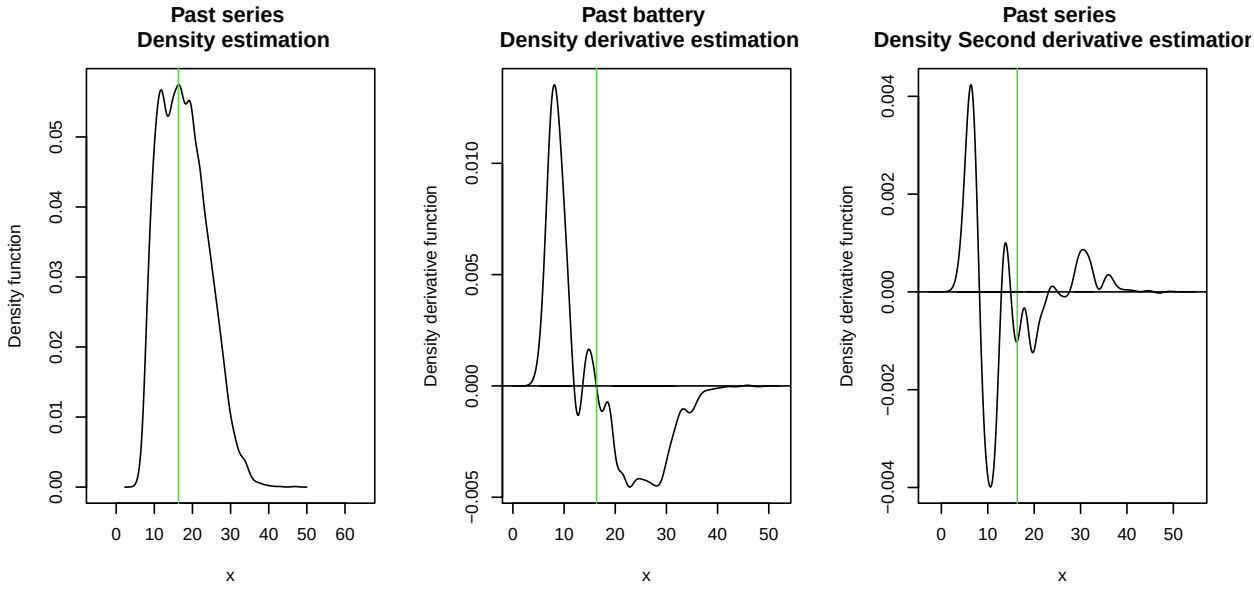

Figure 10: Second derivative analysis: Past series