Rolling median

Outliers

Contents

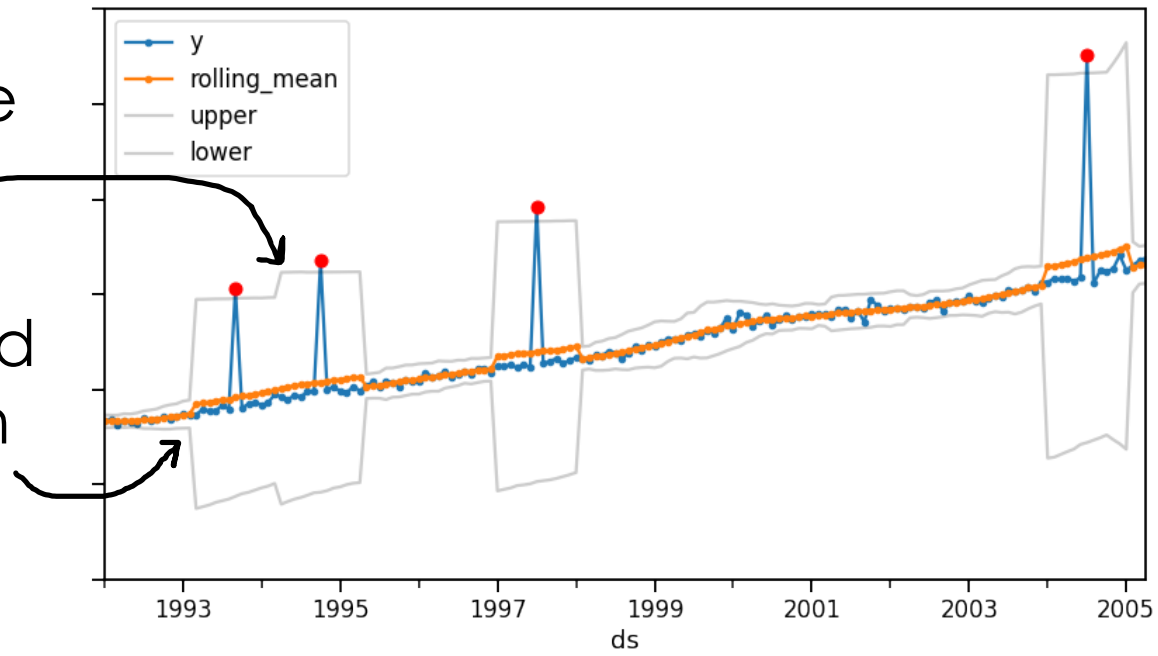
RE-VISIT LIMITATIONS OF ROLLING MEAN

MEDIAN ABSOLUTE DEVIATION

ROLLING MEDIAN FOR OUTLIER DETECTION

Mean and std not robust to outliers

- Mean and standard deviation change a lot in presence of outlier
- Hence outlier detection less sensitive as thresholds become large when window includes an outlier
- The rolling mean and rolling standard deviation change abruptly when an outlier enters the window
- So outlier detection becomes very sensitive to the choice of threshold and window size



Rolling Median: Motivation

- Median is robust to outliers, can use instead of the mean
- What is an outlier robust alternative to standard deviation?
- The Median Absolute Deviation (MAD):

$$MAD = \text{median}(|\mathrm{y_t} - \text{median}(\mathrm{y})|)$$

- A value can be considered an outlier if it lies outside median ± 3.5 × median absolute deviations[1]

[1] Boris Iglewicz and David Hoaglin (1993), "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor.

Median absolute deviation

```
median_absolute_deviation = lambda y: np.median(np.abs(y - np.median(y)))
```

```
# Example with an outlier
data_with_outlier = [1,1,1,2,2,2,1e6]

print(f"Dataset: {data_with_outlier}")
print(f"Median: {np.median(data_with_outlier)}")
print(f"Median absolute deviation: {median_absolute_deviation(data_with_outlier)}")
print(f"Mean: {np.mean(data_with_outlier)}")
print(f"Standard deviation: {np.std(data_with_outlier)}")
```

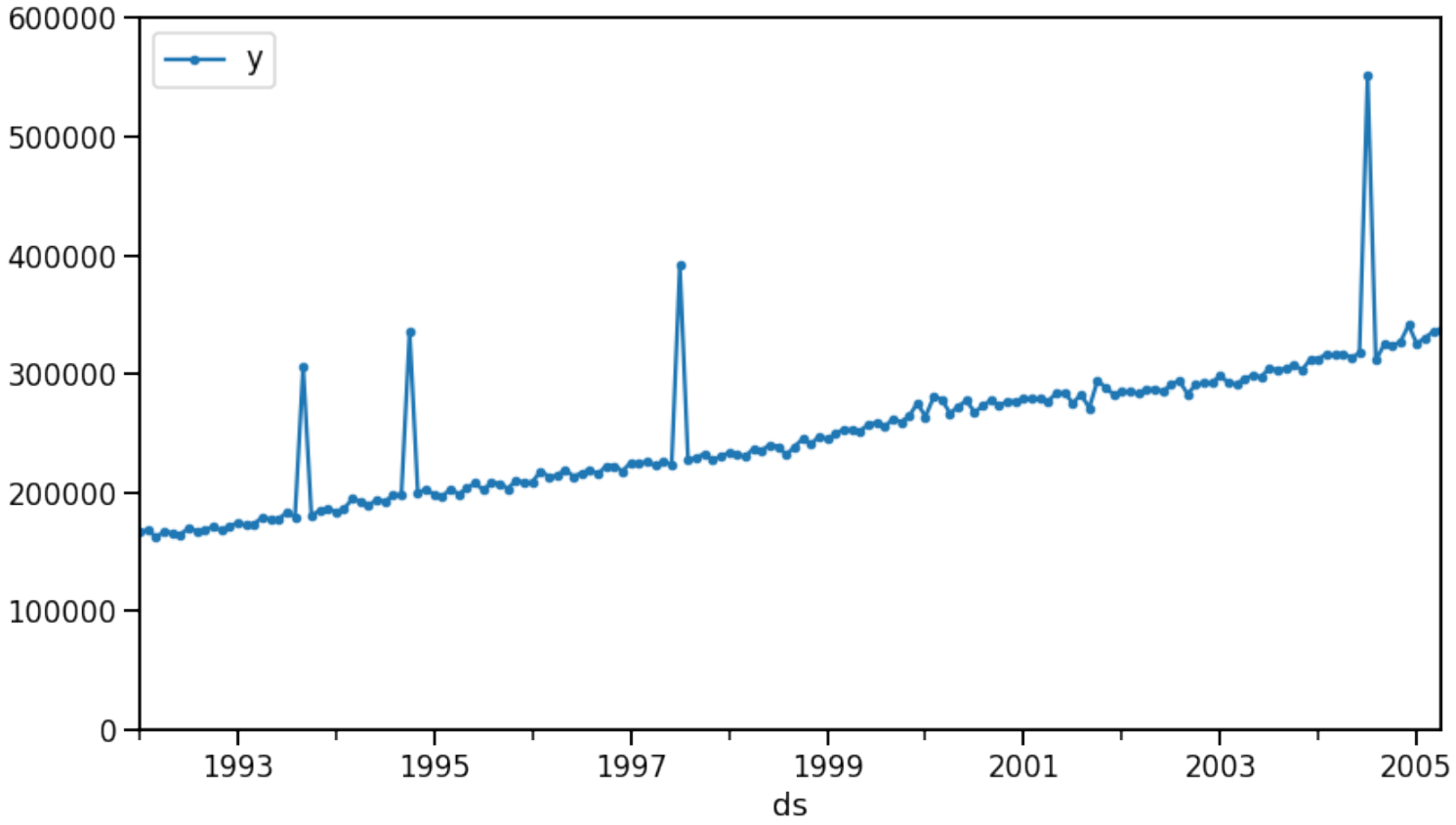
```
Dataset: [1, 1, 1, 2, 2, 2, 1000000.0]
Median: 2.0
Median absolute deviation: 1.0
Mean: 142858.42857142858
Standard deviation: 349926.5812215296
```

Rolling median as estimation method

- $|y_t - \hat{y}_t| > \delta$
- Use rolling median and MAD
- $\hat{y}_t = median(y_{t-T}, \dots, y_{t-1}, y_t, y_{t+1}, \dots, y_{t+T})$
- $\delta_t = \alpha \times MAD(y_{t-T}, \dots, y_{t-1}, y_t, y_{t+1}, \dots, y_{t+T});\ \alpha = 3.5$[1]

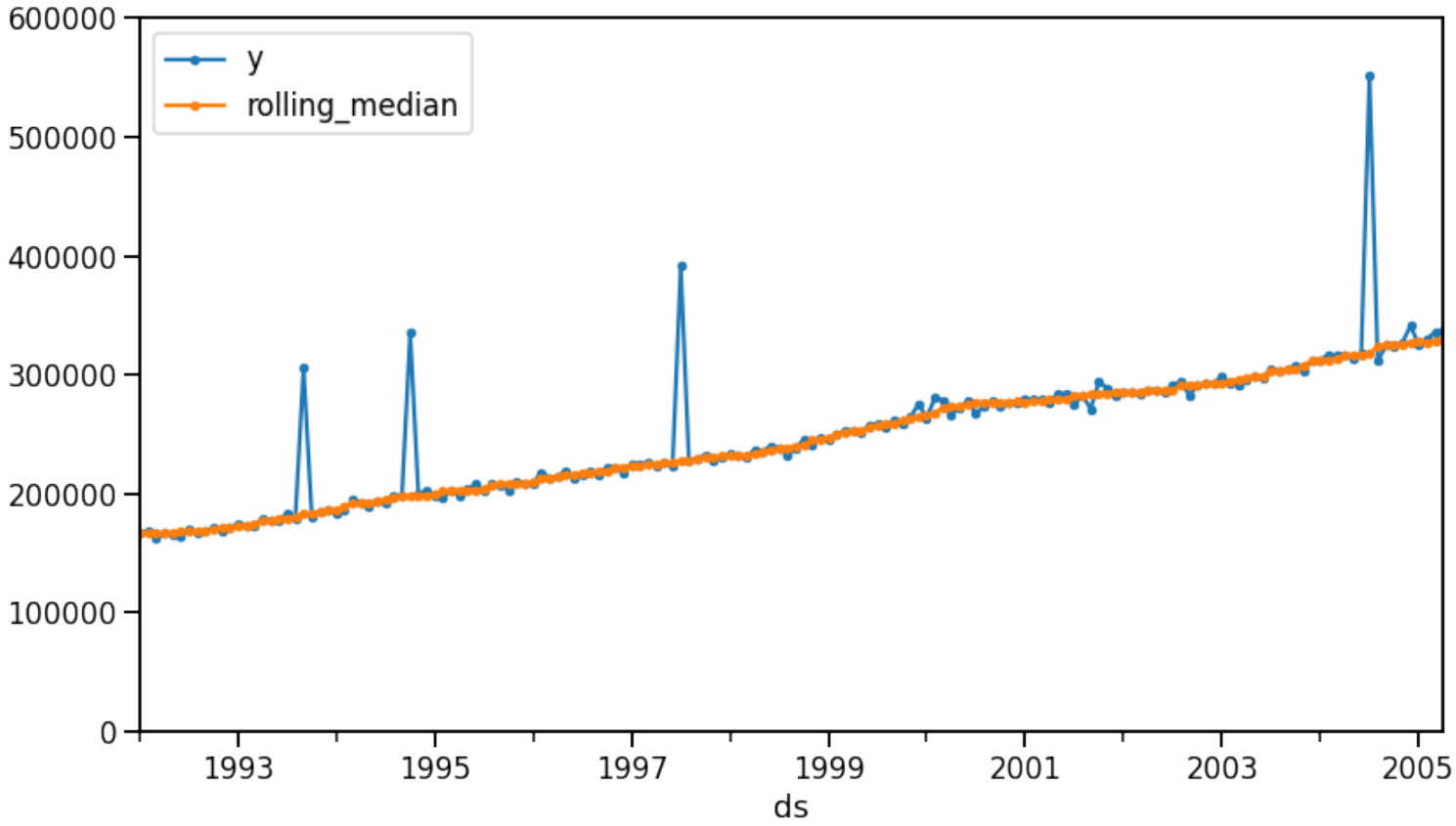
[1] Boris Iglewicz and David Hoaglin (1993), "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor.

Rolling median for outlier detection



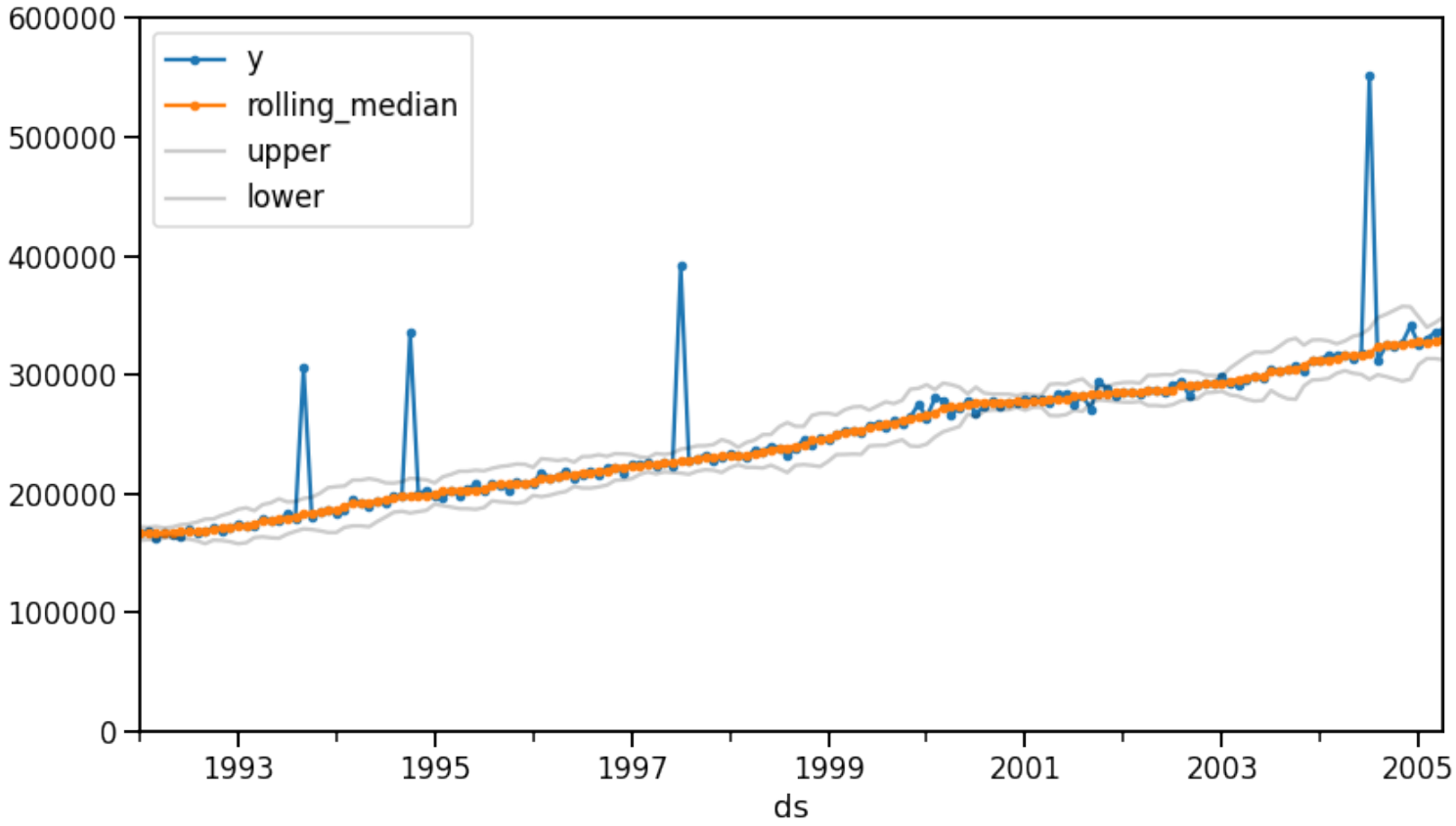
Rolling median for outlier detection

$\hat{y}_t = median(y_{t-T}, \ldots, y_{t-1}, y_t, y_{t+1}, \ldots, y_{t+T})$; Window size = 2T + 1



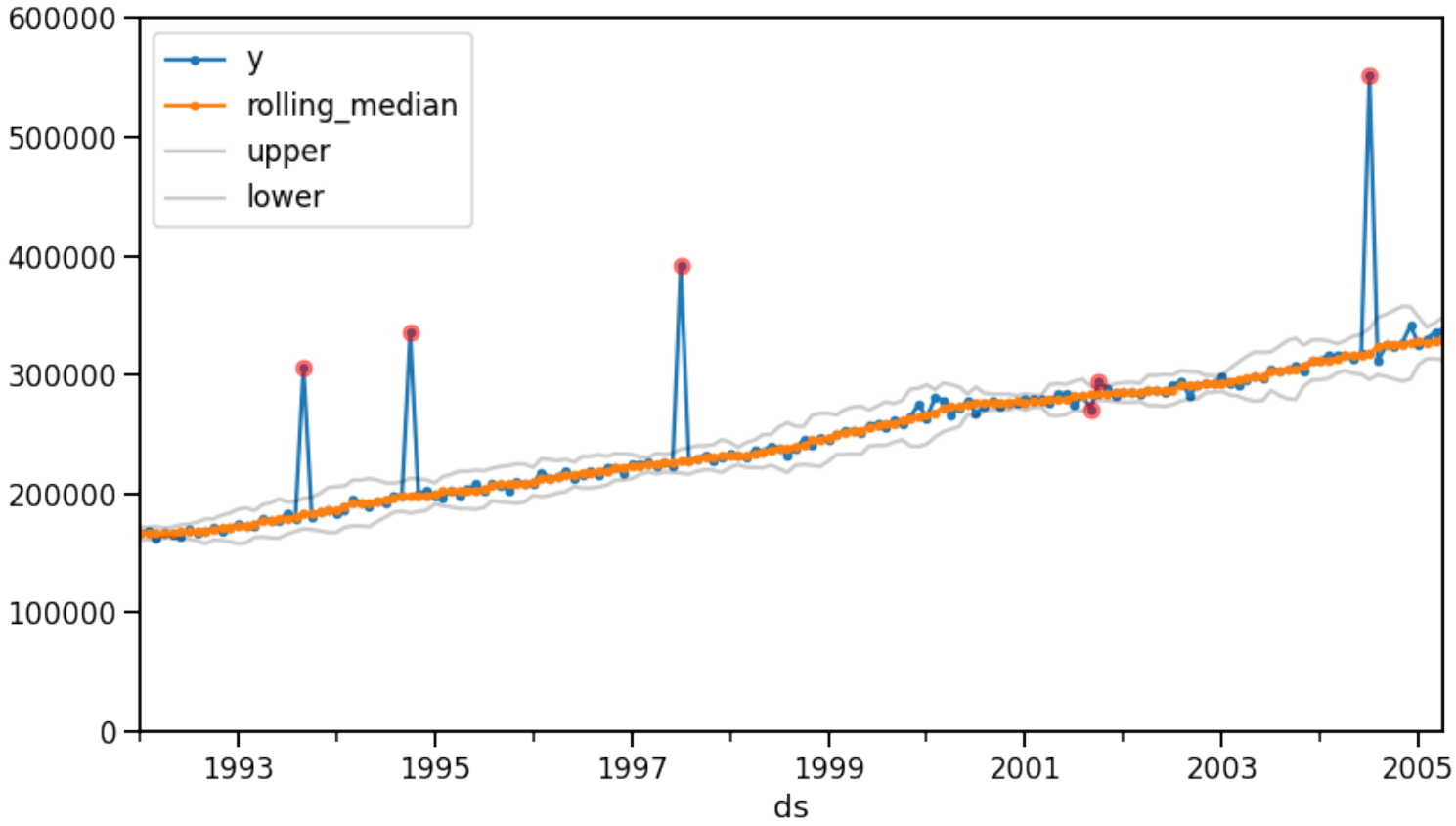
Rolling median for outlier detection

$\delta_t = \alpha \times MAD(y_{t-T}, \ldots, y_{t-1}, y_t, y_{t+1}, \ldots, y_{t+T})$; Window size = 2T + 1



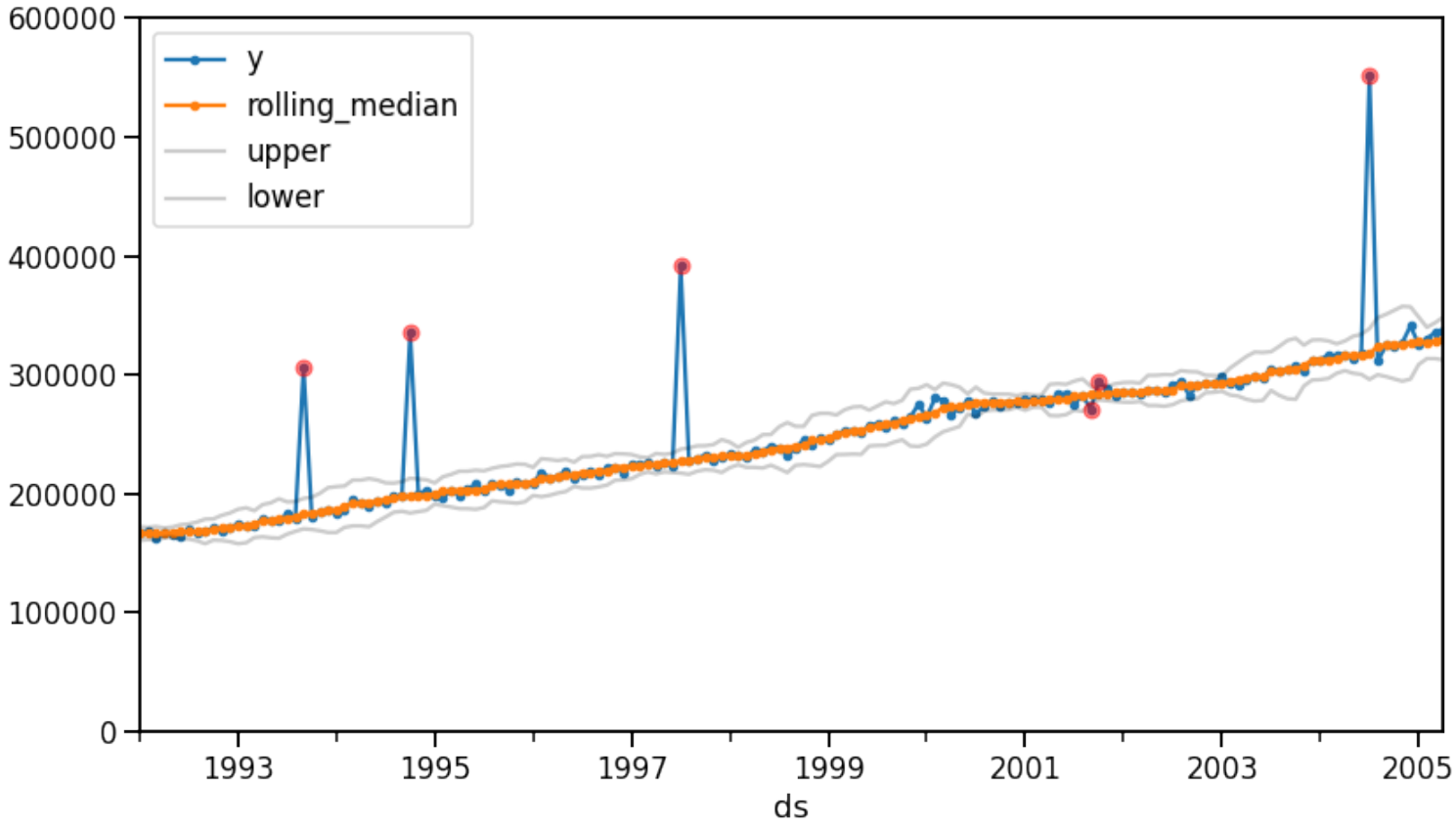
Rolling median for outlier detection

$\delta_t = \alpha \times MAD(y_{t-T}, \ldots, y_{t-1}, y_t, y_{t+1}, \ldots, y_{t+T})$; Window size = 2T + 1



Rolling median for outlier detection

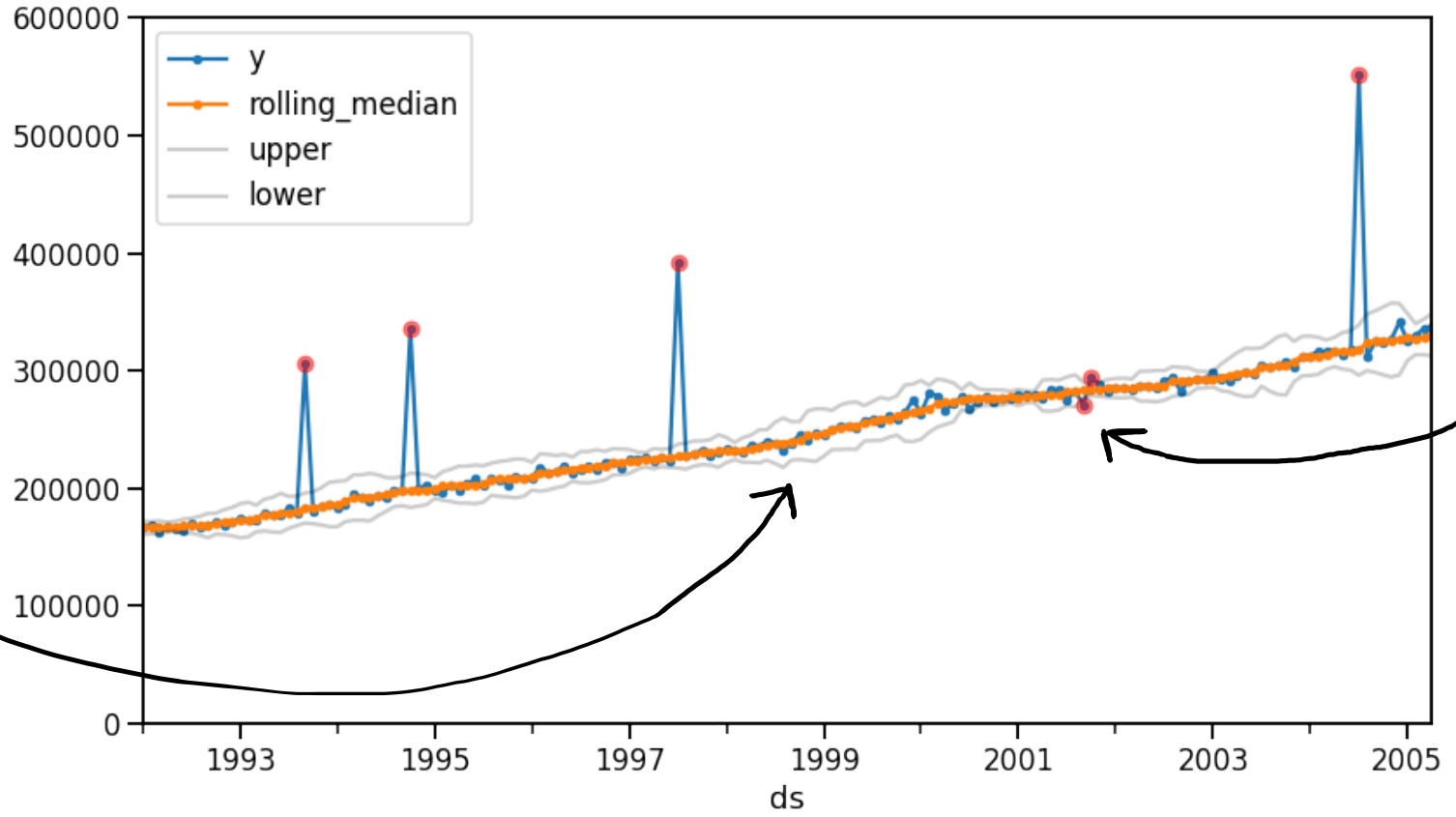
Threshold no longer jumps abruptly



Expected value no longer jumps abruptly

Rolling median for outlier detection

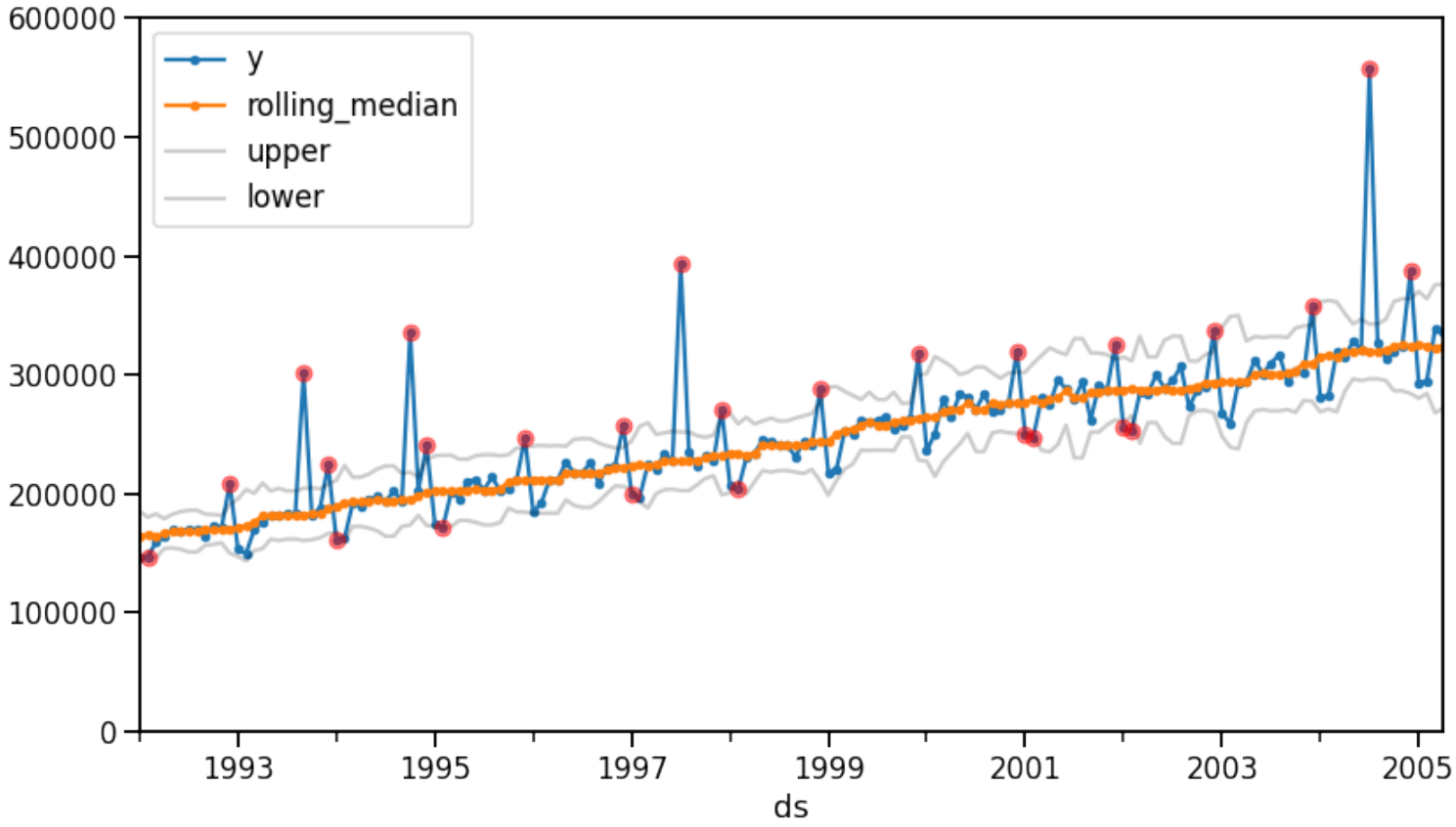
The MAD also increases when the trend is steeper. So the sensitivity is still related to the trend.



Smaller deviations from local behaviour are picked up. Can resolve by adjusting threshold.

Seasonality still complicates matters

- Seasonal spikes can be mistaken for outliers and inflate the threshold. De-seasonalise prior to outlier detection.



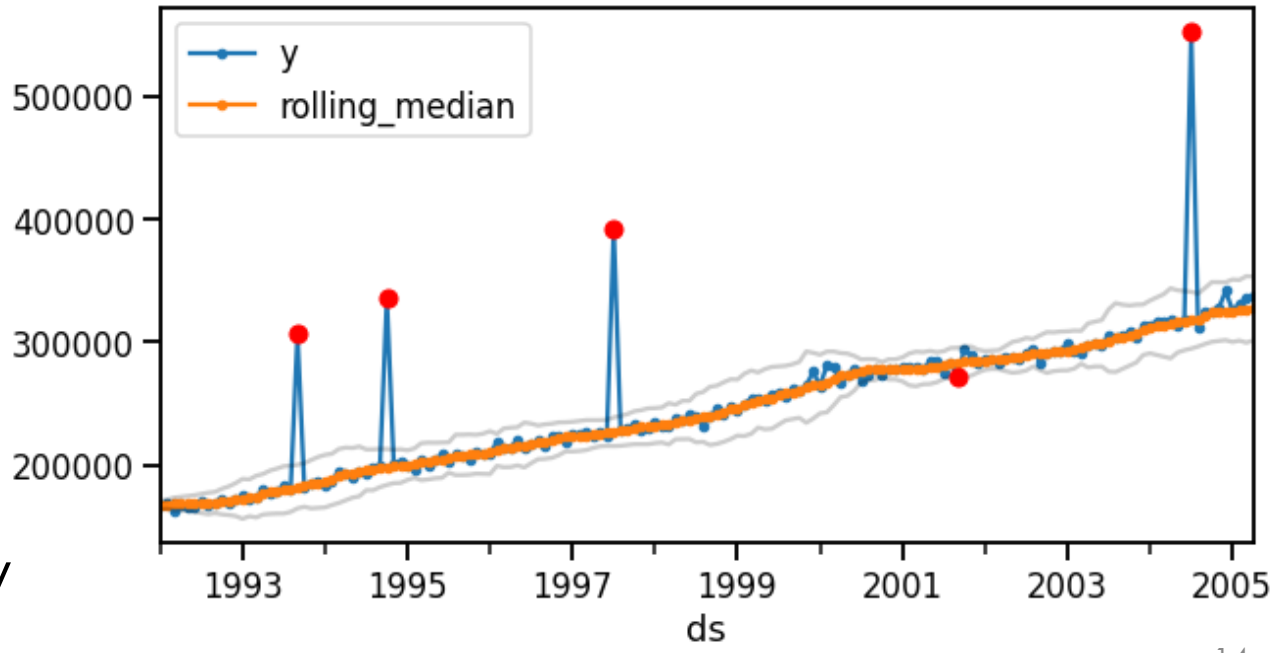
Rolling Median - summary

- Parameters:
  - Window size
  - Threshold
- Pros:
  - Simple
  - Adaptive threshold
  - Robust to outliers
- Cons:
  - Edge effects
  - Sensitivity related to trend
  - Need to remove seasonality

$$\hat{y}_t = median(y_{t-T}, \ldots, y_{t-1}, y_t, y_{t+1}, \ldots, y_{t+T})$$

$$\delta_t = \alpha \times MAD(y_{t-T}, \ldots, y_{t-1}, y_t, y_{t+1}, \ldots, y_{t+T})$$

$$MAD = median(|y - median(y)|)$$



Summary

Mean and standard deviation are not robust to outliers

Median and median absolute deviation are robust to outliers and can be used instead

Some drawbacks remain such as the sensitivity to trend and need for de-seasonalizing the data