Rolling mean

Outliers

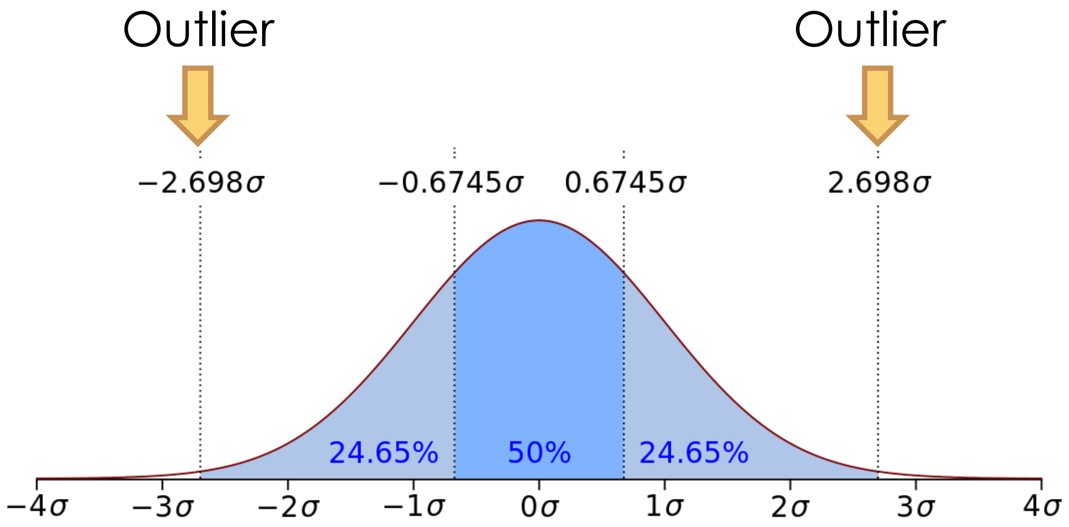
Contents

ROLLING MEAN FOR OUTLIER DETECTION

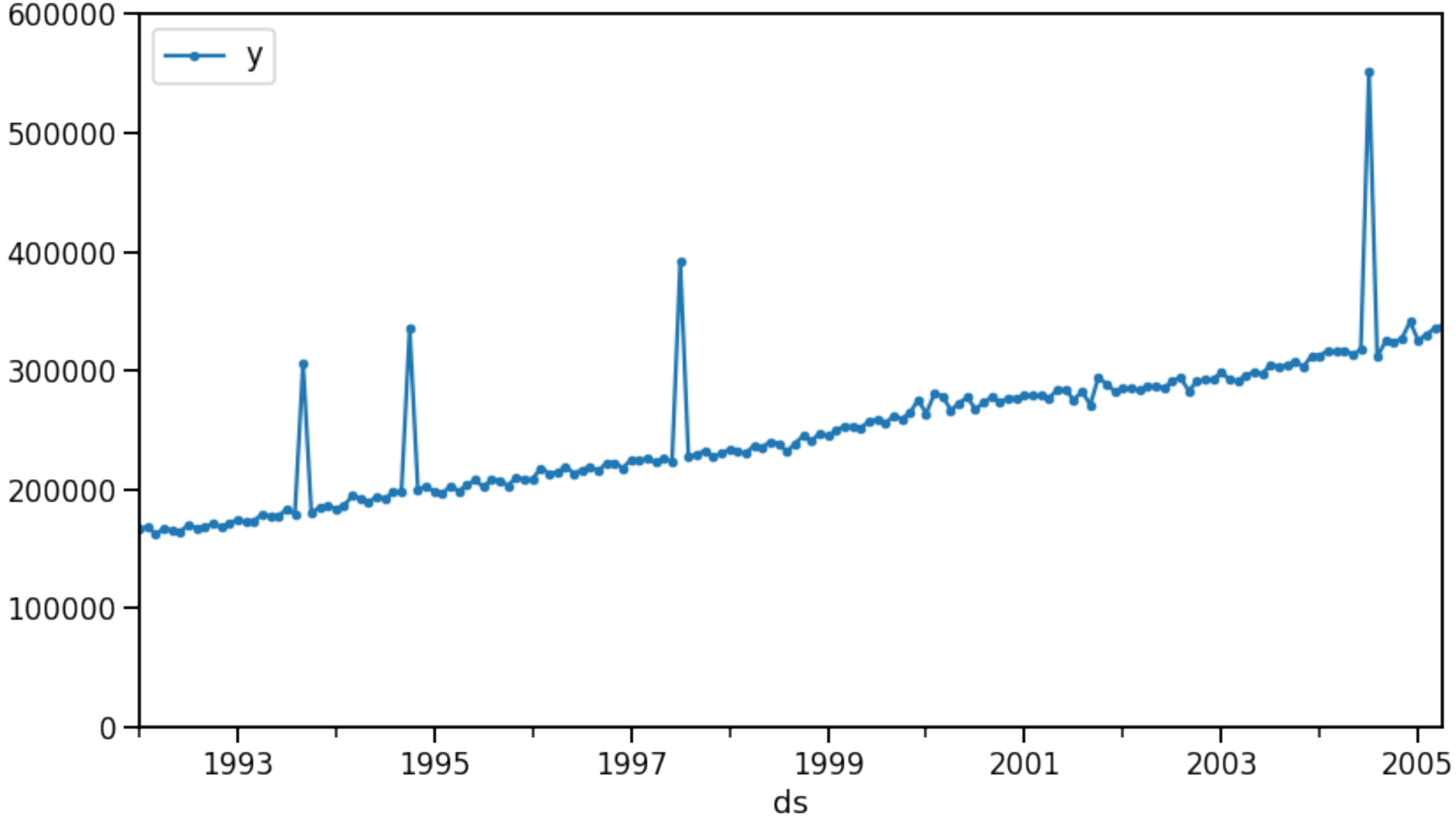
PRACTICAL TIPS

Recap: Estimation method and rolling mean

- $|y_t - \hat{y}_t| > \delta$
- Expected value, $\hat{y}_t$: Use the rolling mean
- Threshold, δ: Use the rolling standard deviation
- Recap: Values outside mean ± 3 × standard deviations can be considered outliers
- $|y_t - \hat{y}_t| > \alpha\hat{\sigma}_t$ where $\alpha = 3$

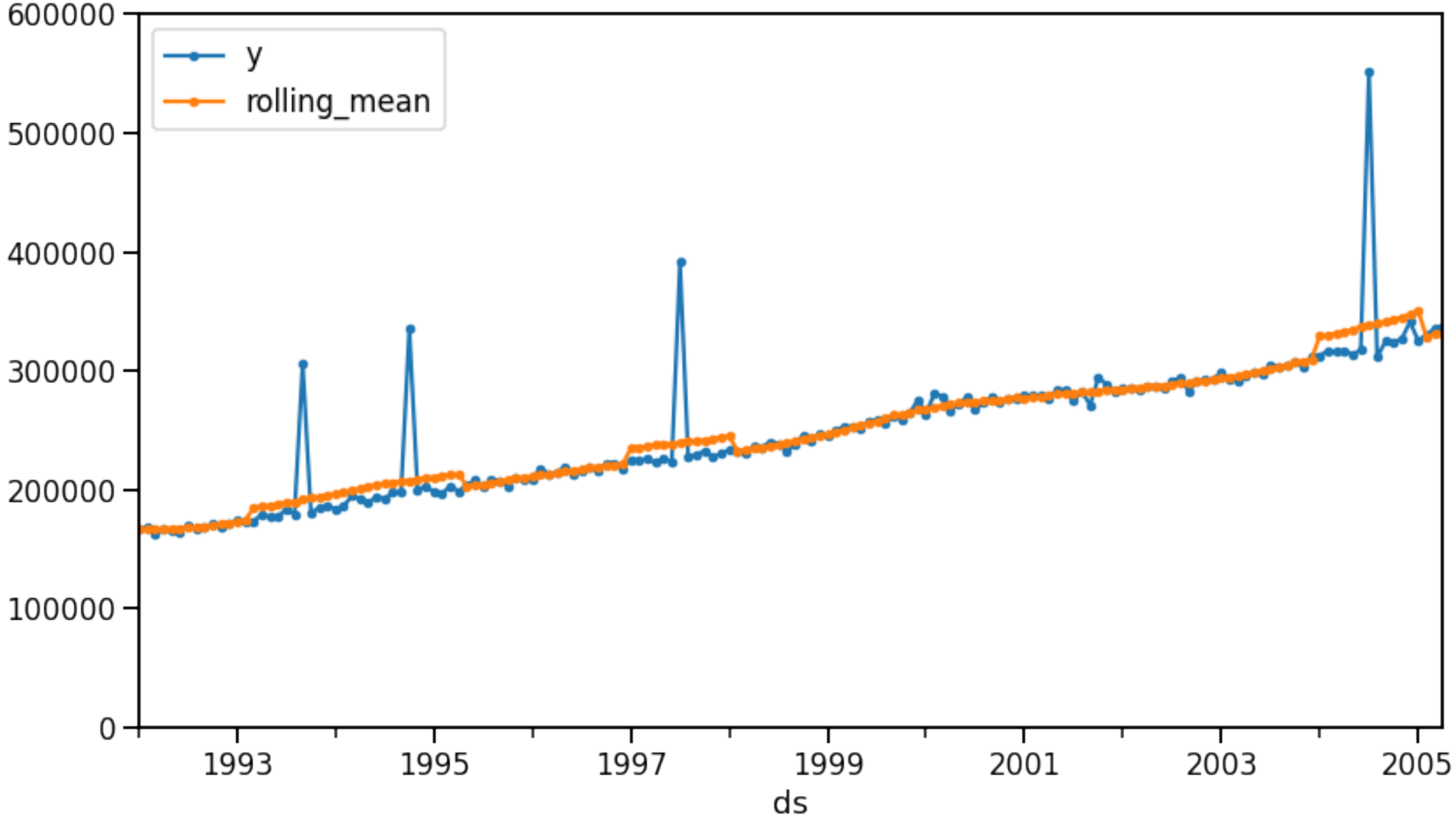


Rolling mean for outlier detection



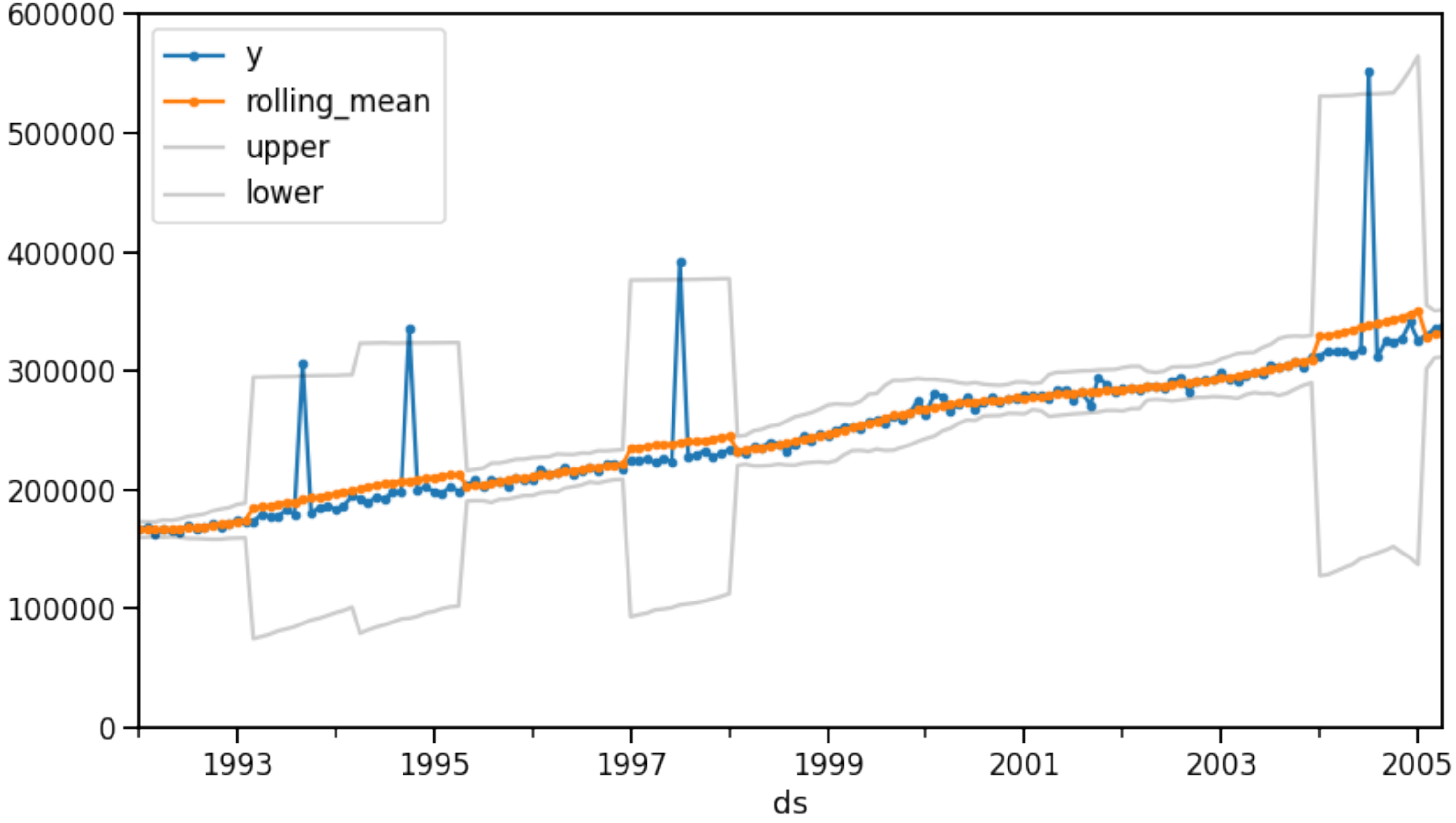
Rolling mean for outlier detection

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



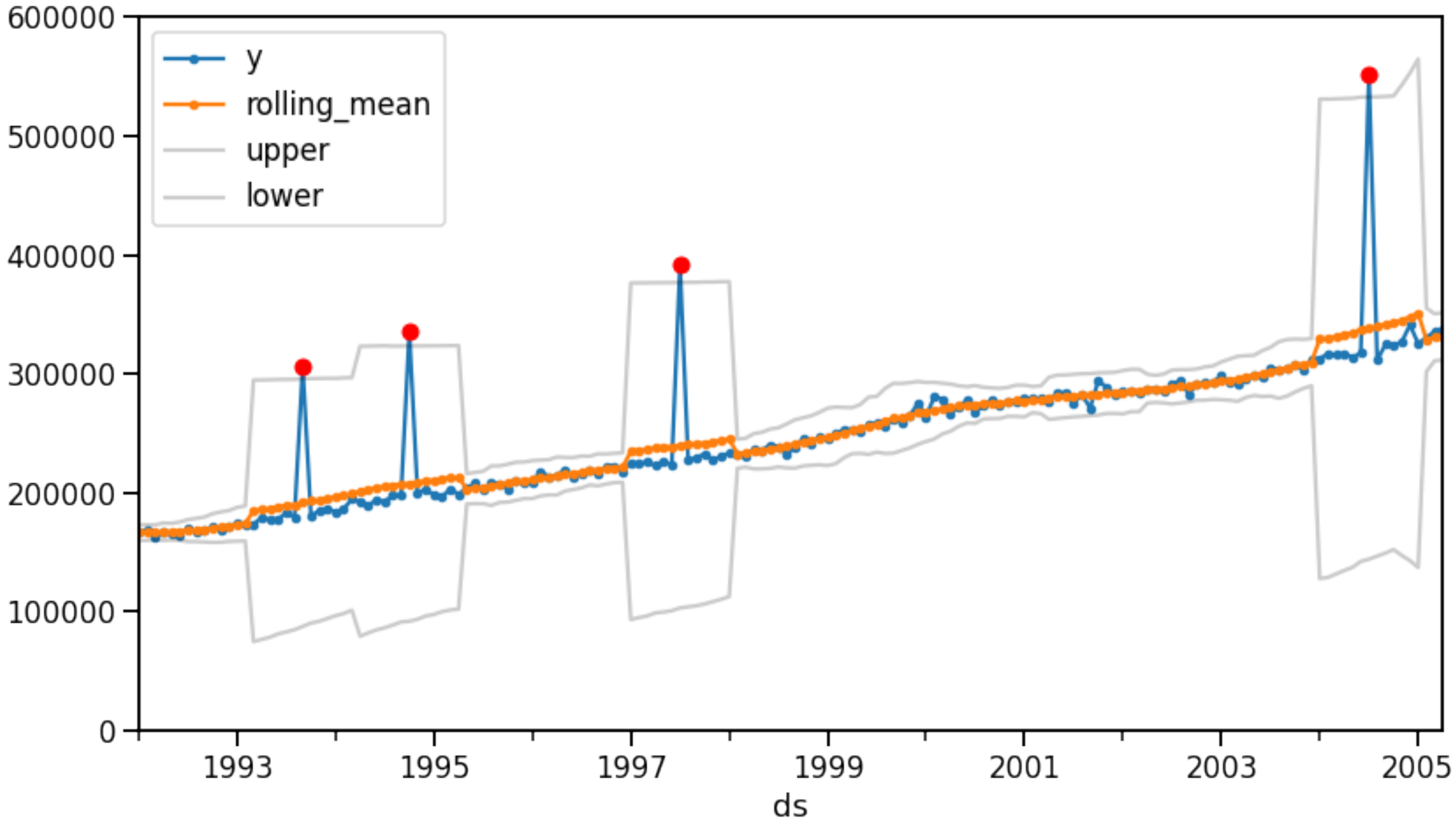
Rolling mean for outlier detection

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



Rolling mean for outlier detection

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



Setting the parameters

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

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

Window size

Want to smooth out short term fluctuations. Trial and error.

The seasonal period is common to smooth out any periodicity.

For example, window size = 7 for daily data to smooth weekly seasonality.

Threshold

$\alpha = 3$ is common but can be tuned to desired sensitivity

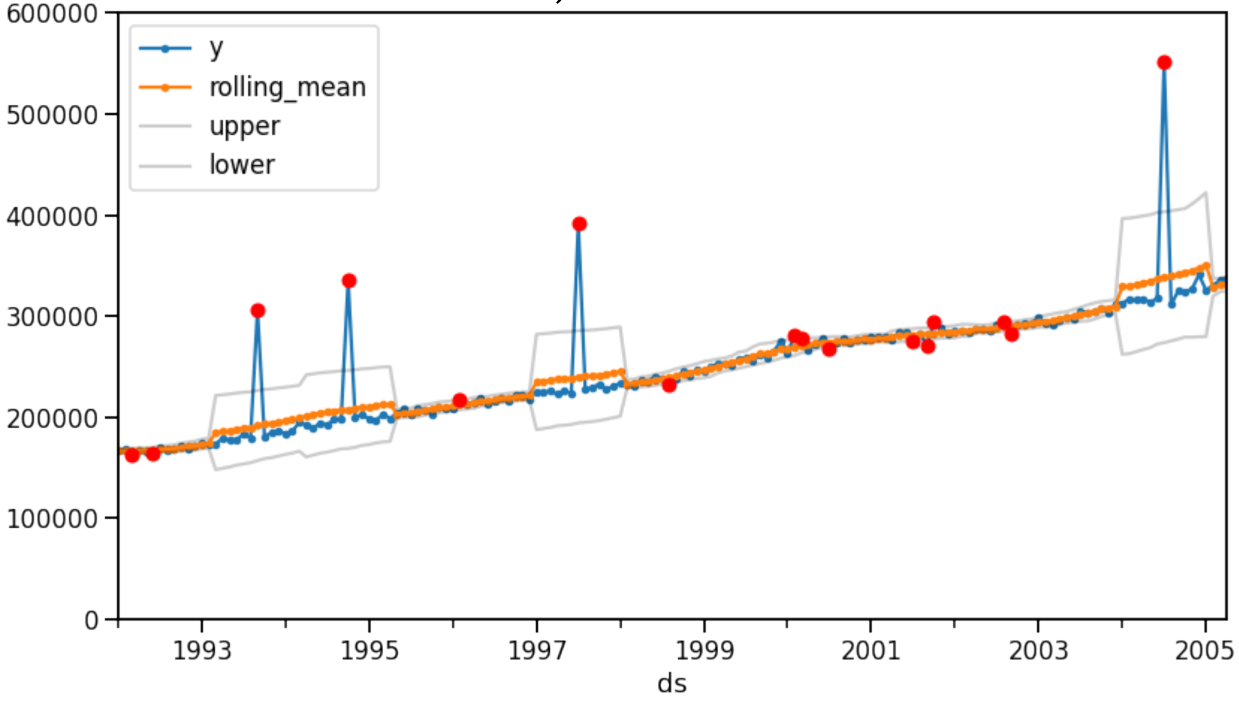
Low α : more sensitive

High α: less sensitive, good if interested only in extreme outliers

Setting the parameters - threshold

- $\delta_t = \alpha \times std(y_{t-T}, \ldots, y_{t-1}, y_t, y_{t+1}, \ldots, y_{t+T})$
- Threshold $\alpha = 3$ is common but can be tuned to desired sensitivity

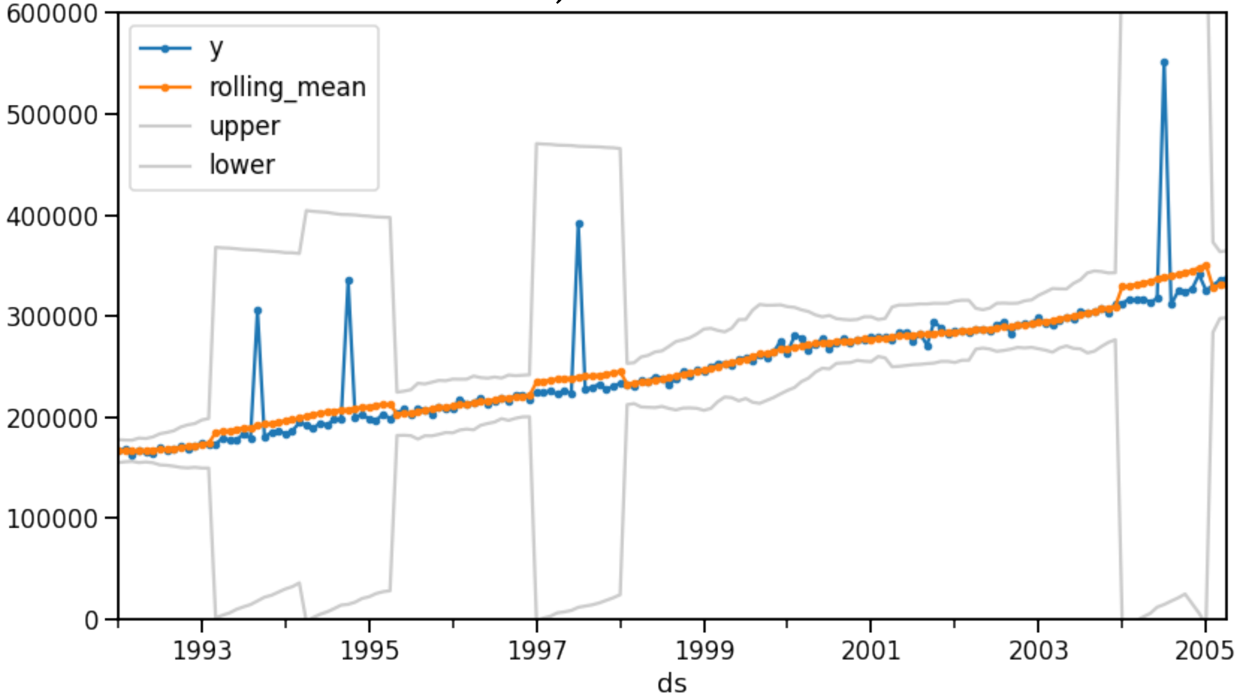
$\alpha = 1$; Over sensitive



Setting the parameters - threshold

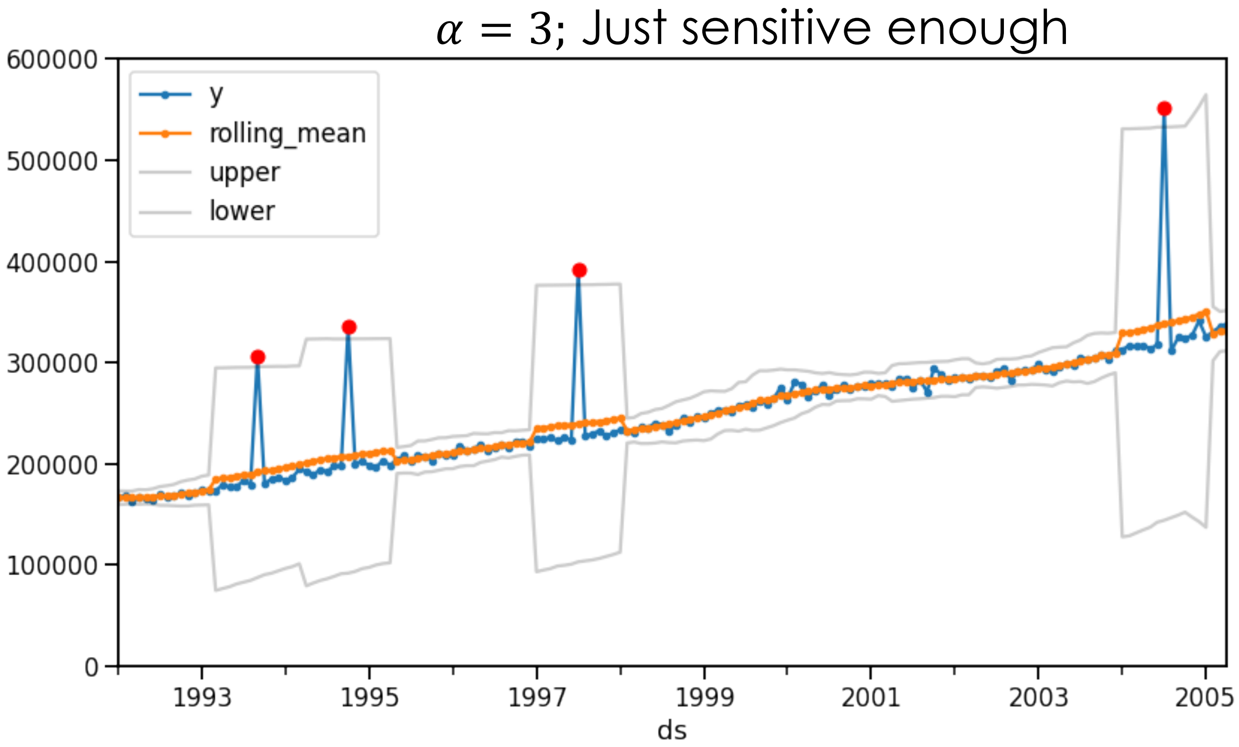
- $\delta_t = \alpha \times std(y_{t-T}, \dots, y_{t-1}, y_t, y_{t+1}, \dots, y_{t+T})$
- Threshold $\alpha = 3$ is common but can be tuned to desired sensitivity

$\alpha = 5$; Under sensitive



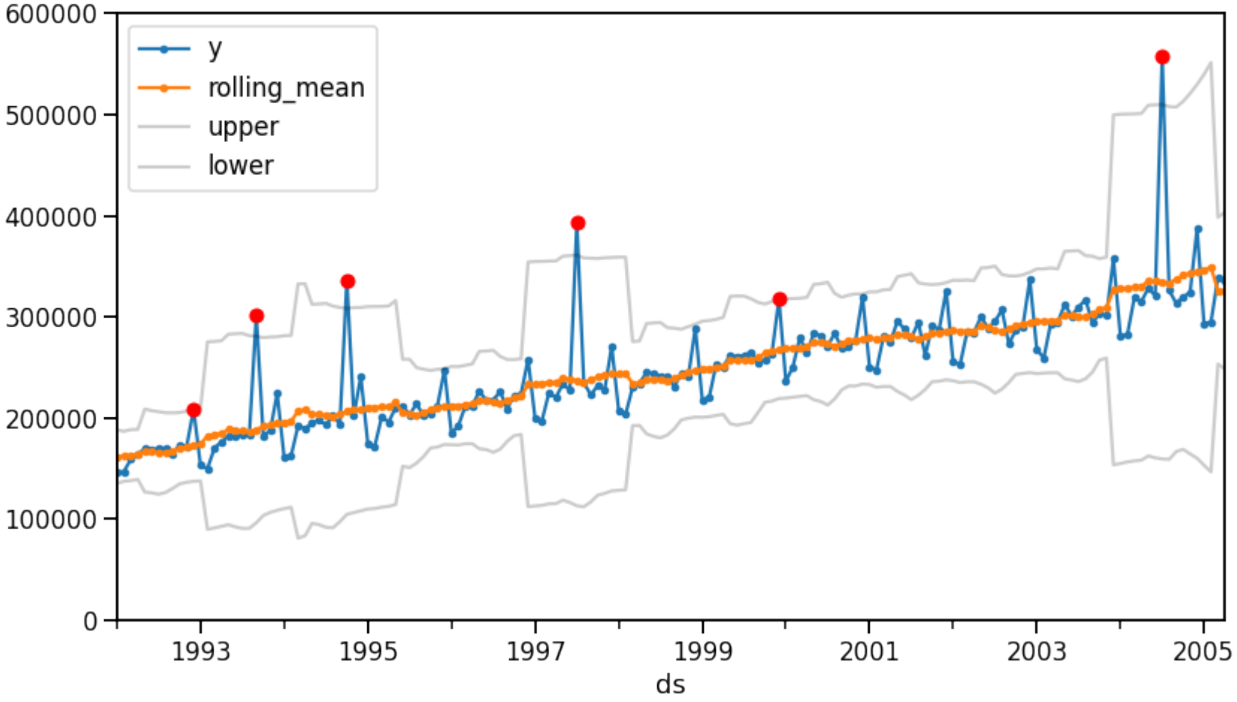
Setting the parameters - threshold

- $\delta_t = \alpha \times std(y_{t-T}, \dots, y_{t-1}, y_t, y_{t+1}, \dots, y_{t+T})$
- Threshold $\alpha = 3$ is common but can be tuned to desired sensitivity



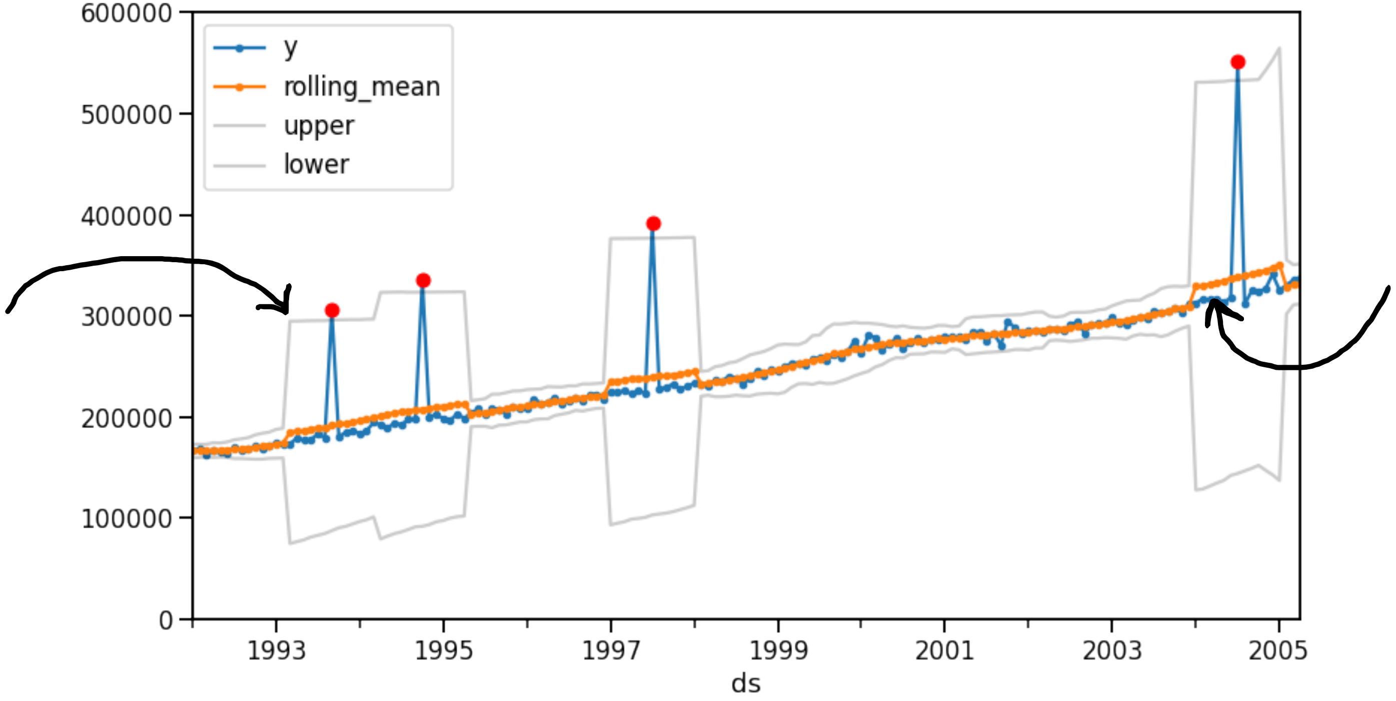
Seasonality complicates matters

- Seasonal spikes can be mistaken as outliers. They also inflate the standard deviation making outlier detection less sensitive.
- Solution: De-seasonalise the data prior to outlier detection



Mean and standard deviation are not robust to outliers

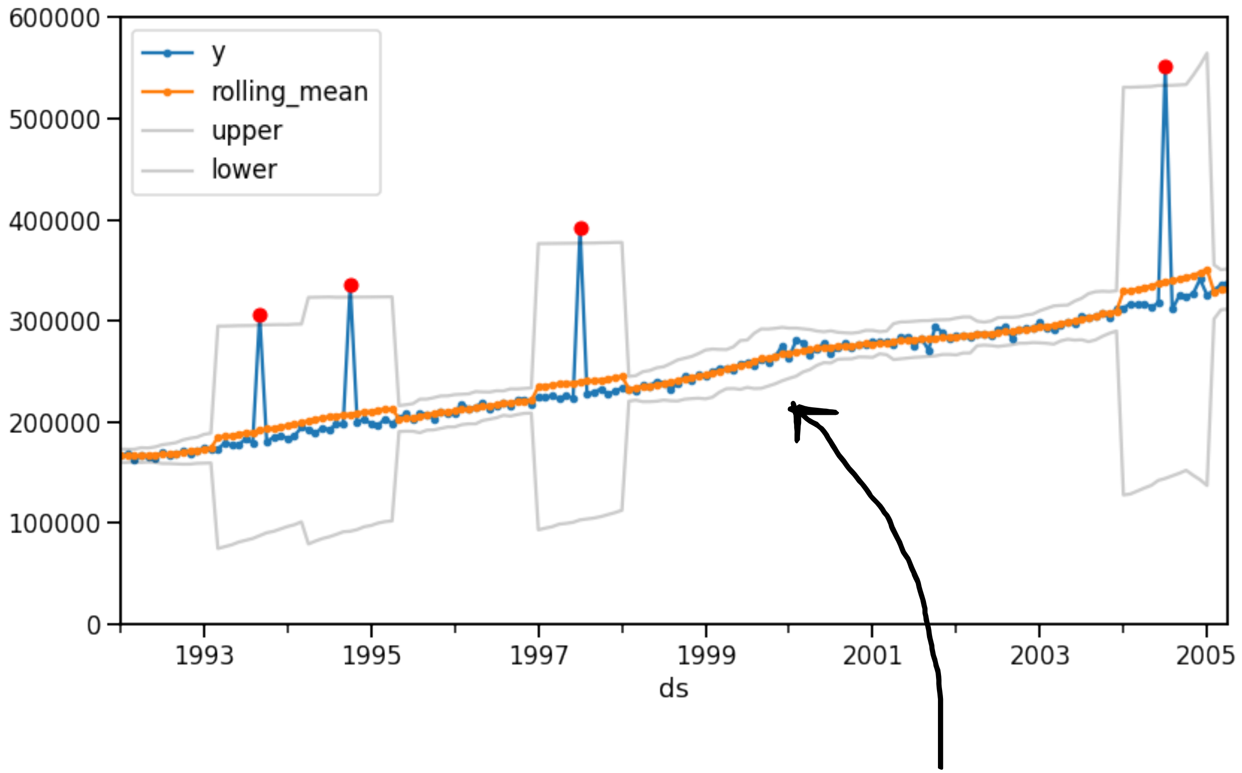
The thresholds change abruptly and the variance is over-estimated whilst outlier in window. Hence less sensitive here.



The expected value changes abruptly and over-predicts whilst outlier in window.

The detection of outliers here is very sensitive to the choice of the window size, threshold, and the data itself.

Sensitivity depends on gradient of trend



The standard deviation increases in regions of steeper trend.
This makes outlier detection less sensitive in these areas.

Rolling mean - summary

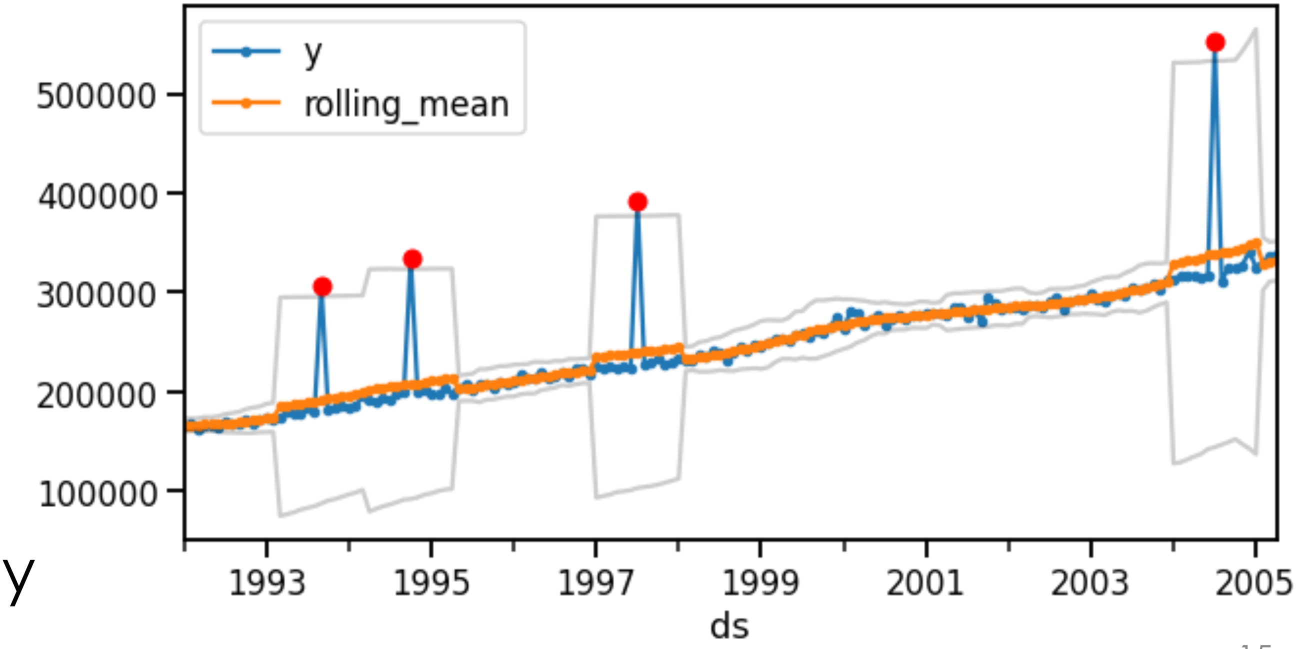
- Parameters:
  - Window size
  - Threshold

- Pros:
  - Simple
  - Adaptive threshold

- Cons:
  - Not robust to outliers
  - Edge effects
  - Sensitivity related to trend
  - Need to remove seasonality

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

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



Summary

Estimation methods compare the actual to the expected value

Rolling mean and std can be used to compute expected value and threshold

Parameters are window size and threshold. Heuristics exist to set these, but ultimately depends on the data