Anomaly Detection
Data adaptation: time series

Paul Irofti
Cristian Rusu
Andrei Pătraşcu

Computer Science Department
University of Bucharest

Class structure

Topics for today:

- discuss characteristics of time series
- basic statistics for anomaly detection
- change-point model for time series
- average and linear models for time series
- anomaly detection

Time series characteristics

What is a time series?
For us, in this class: time series = vector of real values + time stamp

They appear everywhere where a phenomenon is monitored:

- finance (performance indicator measurements)
- healthcare (vital sign measurements)
- industry (sensor measuments)
- ...

Two questions that will interest us today:

- do time series suffer significant changes over time?
- is there something anomalous in the time series?

Time series characteristics

Two questions that will interest us today:

- do time series suffer significant changes over time?
- is there something anomalous in the time series?

We are interested in two situations:

- **a sudden large change in the time series**
  - short event
  - big impact
  - time series returns quickly to “normal”
- **a change in the time series statistics**
  - slow drift
  - effect not noticable immediately
  - time series changes fundamentally

Time series characteristics

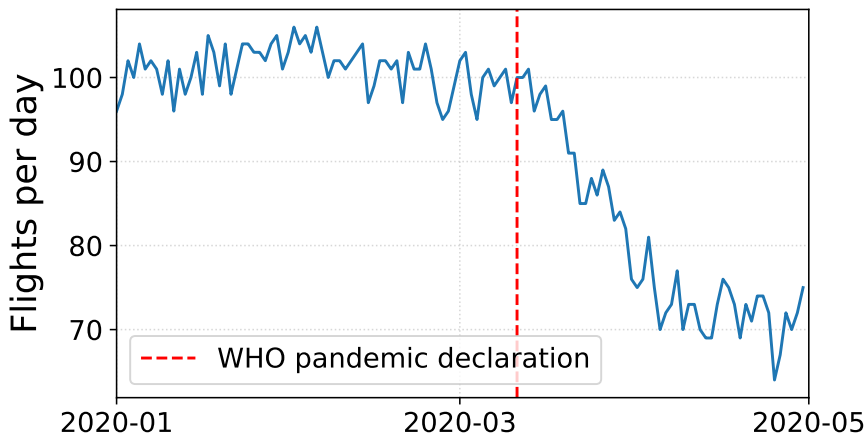
What is the first thing that comes to mind when trying to work with time series?

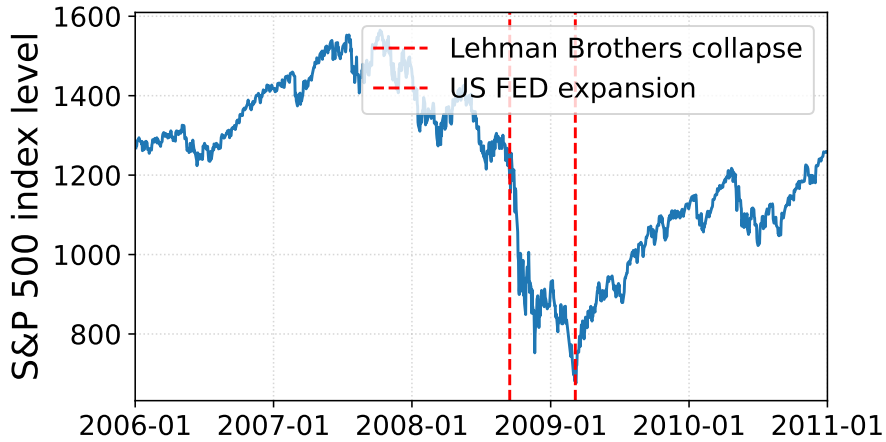
Time series characteristics

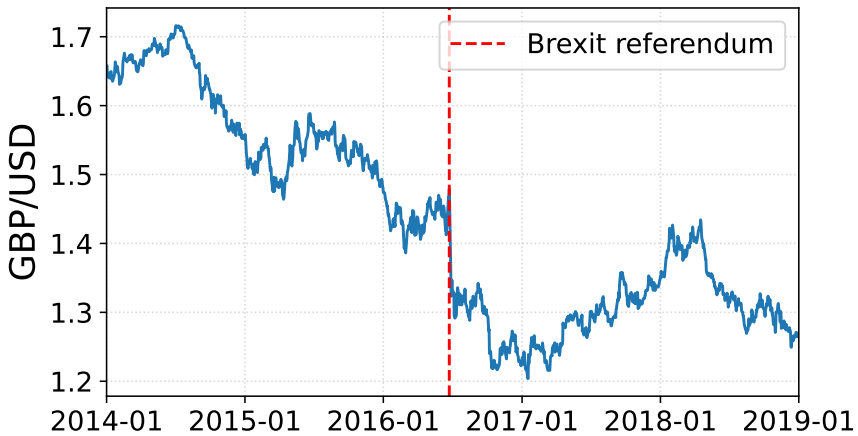
What is the first thing that comes to mind when trying to work with time series?

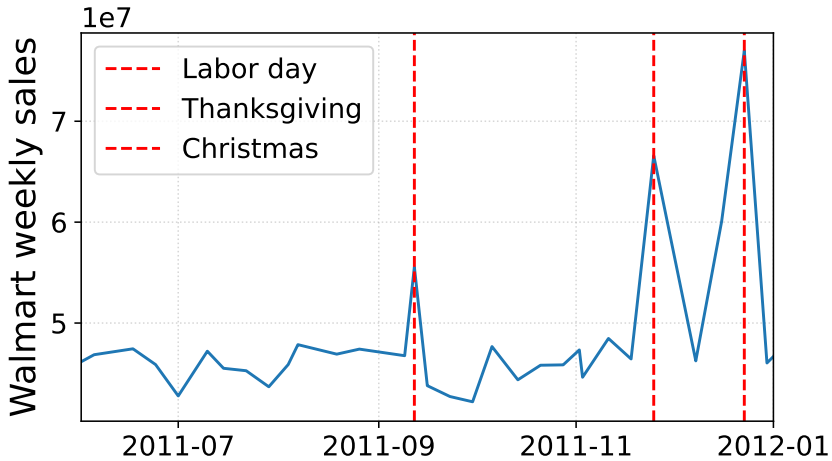
- seems to be a 1D regression problem (if we ignore the time information)
  - how can we turn in into a regression problem and keep the time information?
- the regression problem can be extended into multiple dimensions (feature engineering)
- average a couple of values from the past to try to predict new values (in the style of K-NN)
- try to find seasonal components in the data (peridicity analysis - Fourier)

Typical anomalies for the real world



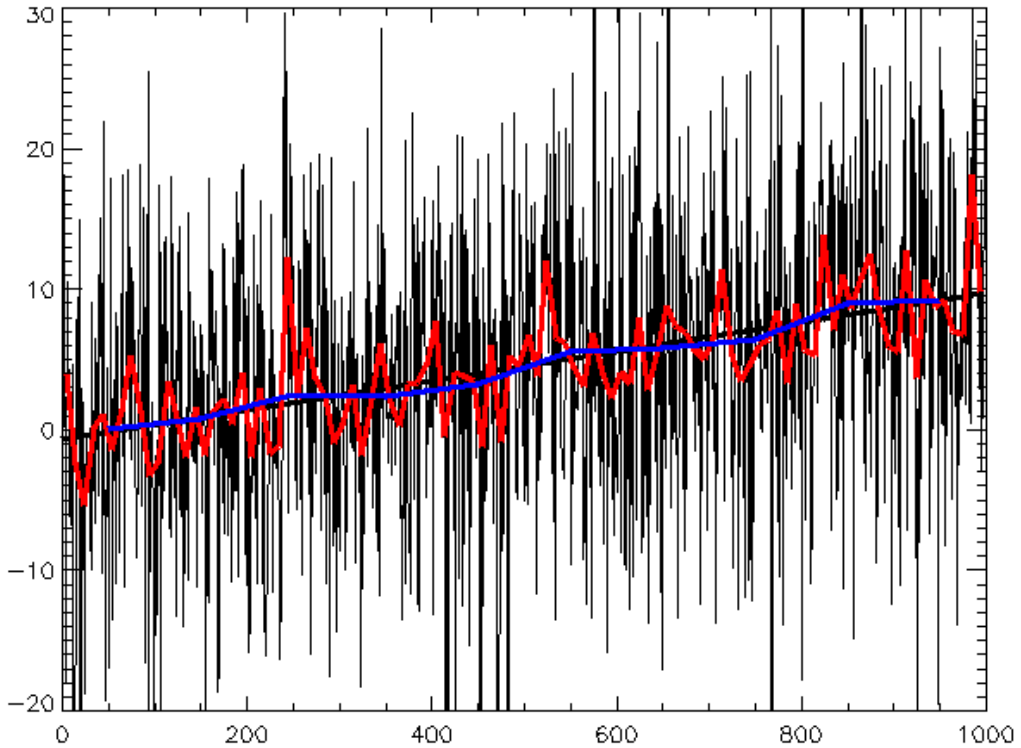




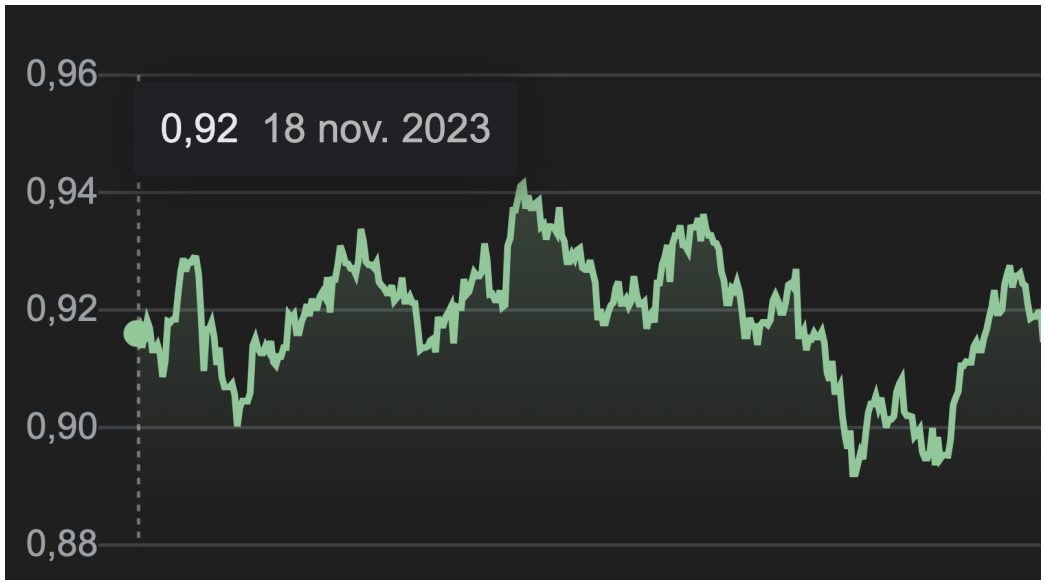


Time series characteristics

A typical time series composed of random data plus a linear trend (source: Wikipedia)



Another typical time series Dollar vs. Euro exchange (source: Google)

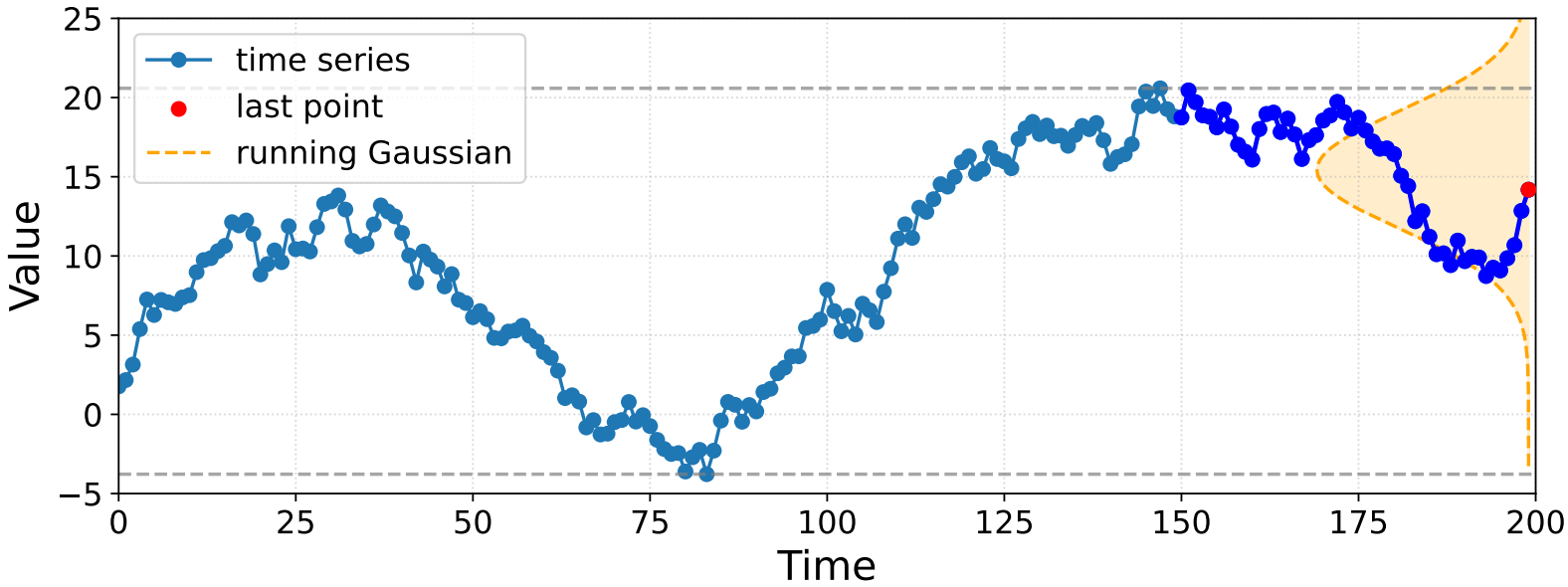


Another typical time series Dollar vs. Euro exchange, with an orange change point (source: Google)

time series like this are not stationary

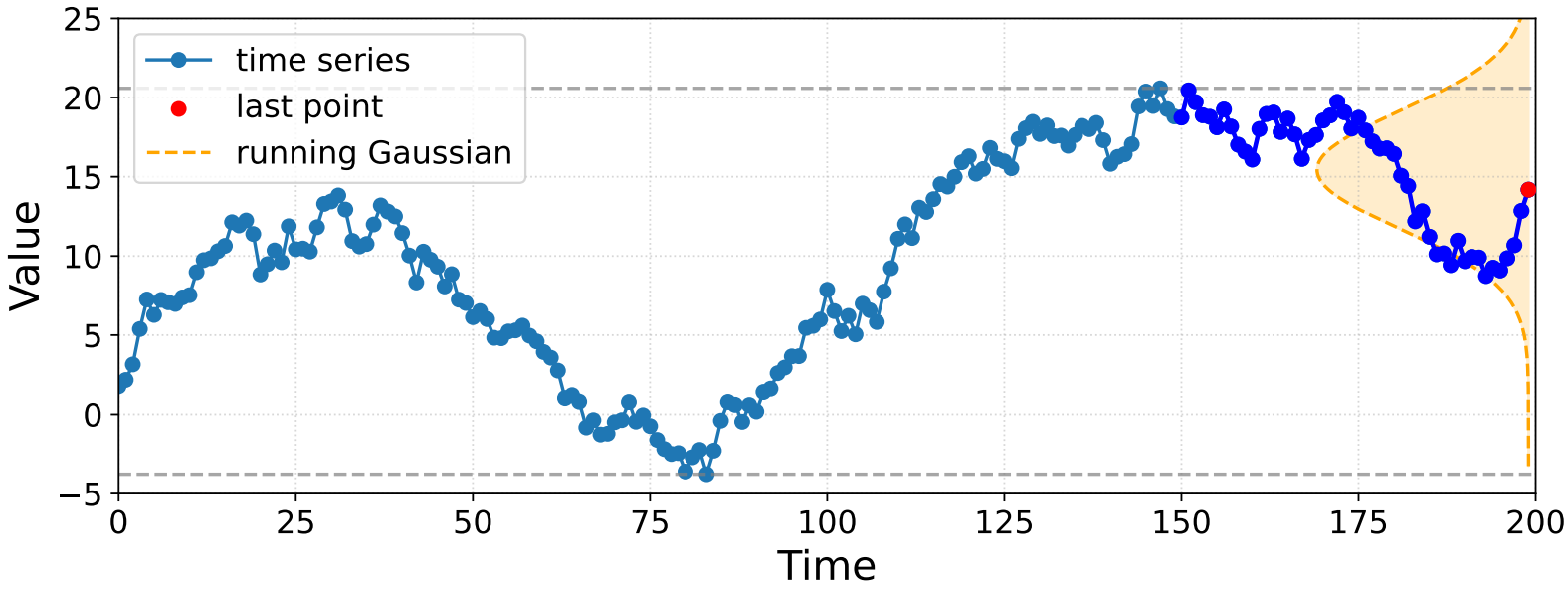
what would linear regression look like on this data?

Model latest points as being sampled form a Gaussian



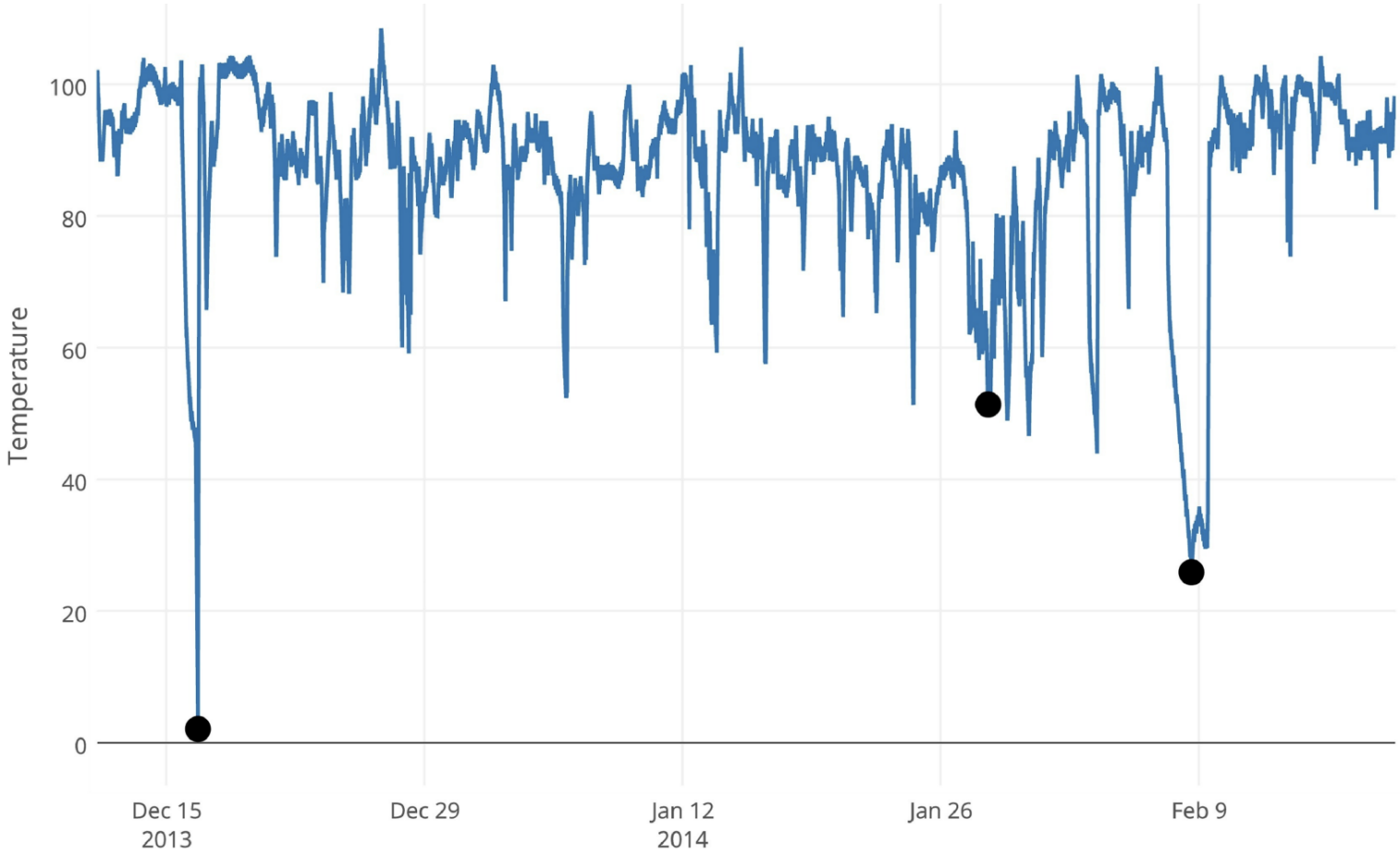
How do we decide if something is an anomaly?

Model latest points as being sampled form a Gaussian

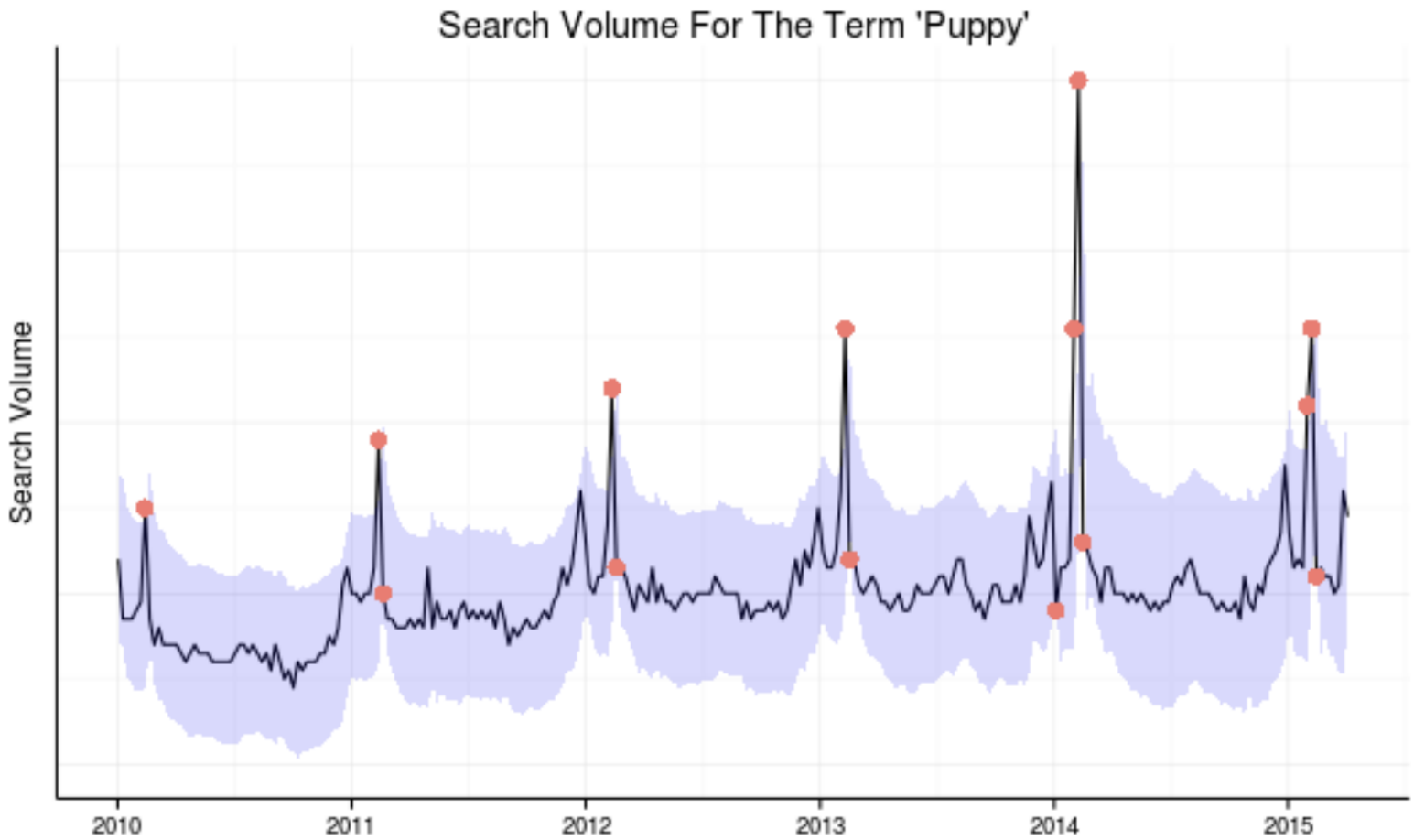


How do we decide if something is an anomaly?
z-scores, Mahalanobis distances

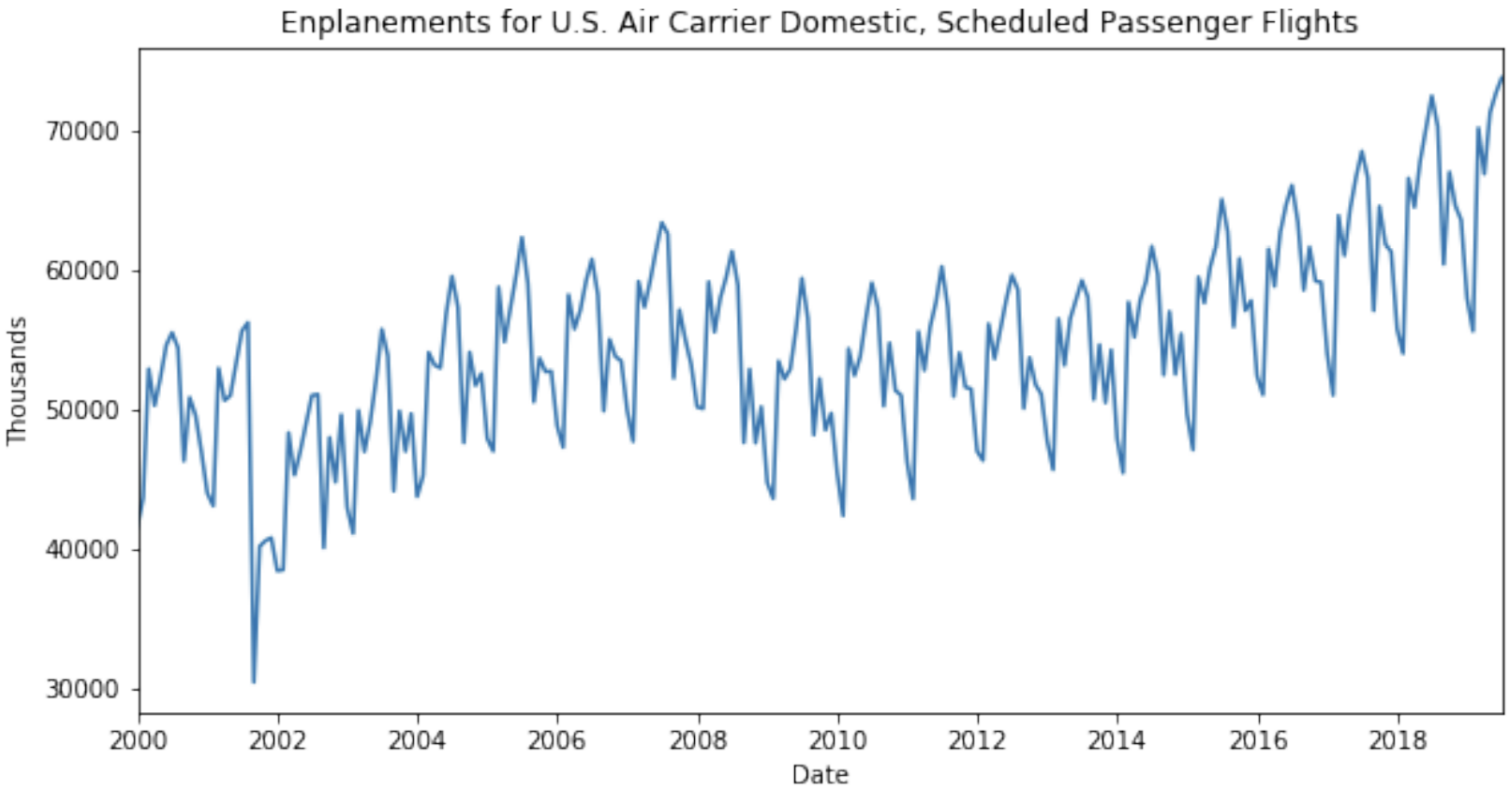
Easy example



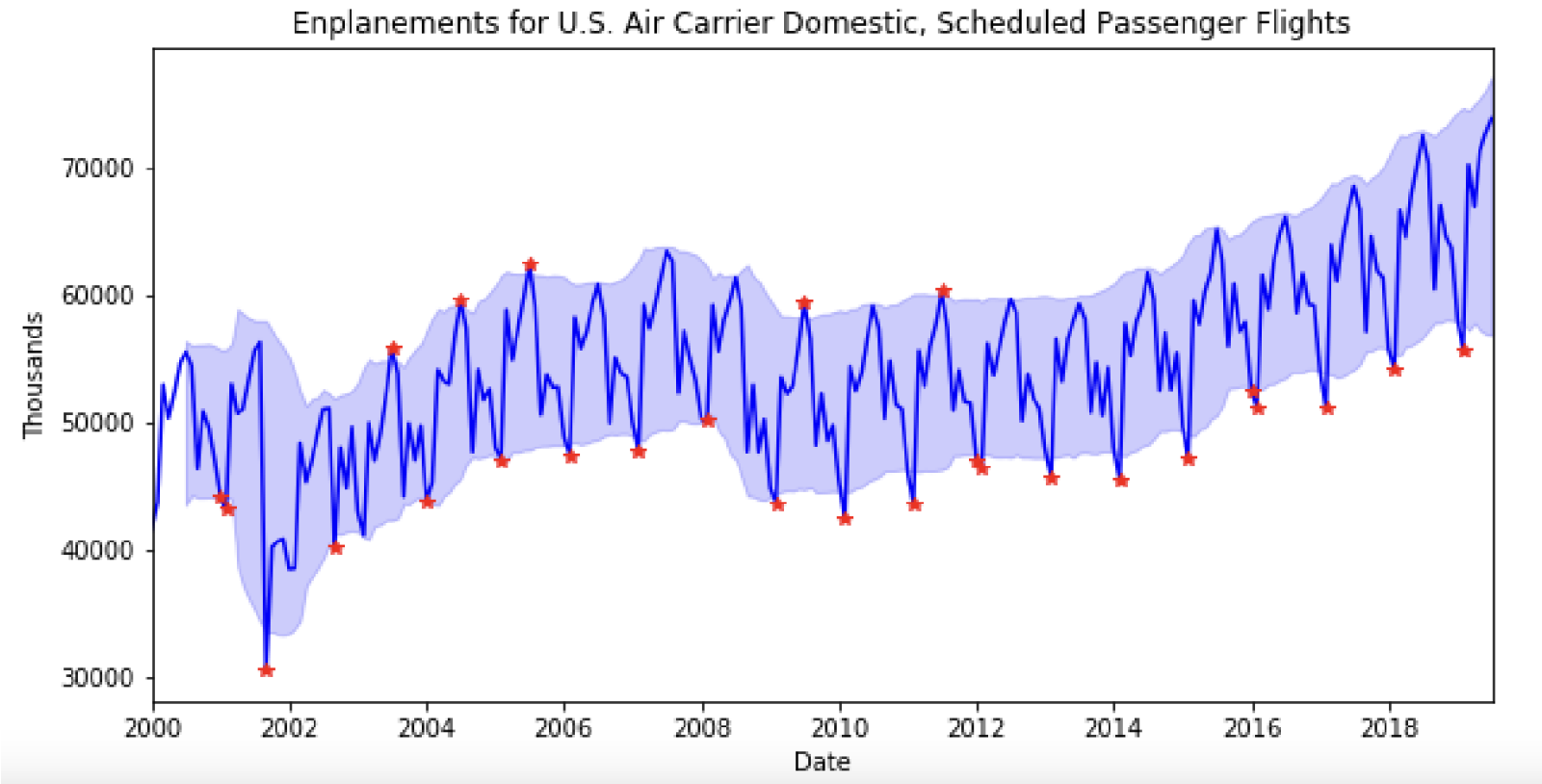
Easy example



Easy example



Mu-sigma model: $|x[n] - \mu| > \lambda\sigma$ ($\lambda = 1.5$ window of size 12)



Change point detection basic statistics (CUSUM)

- We observe a sequence of random variables

$$x[1], x[2], x[3], \ldots$$

we suspect that at some unknown time τ the distribution of the data shifts from a pre-change distribution f_0 to a post-change distribution f_1

- Formally:

$$x[t] \sim \begin{cases} f_0(x), & t \leq \tau, \\ f_1(x), & t > \tau, \end{cases}$$

where τ is unknown and to be detected as quickly as possible

- How would you solve this problem?

Change point detection basic statistics (CUSUM)

- Formally:

$$x[t] \sim \begin{cases} f_0(x), & t \leq \tau, \\ f_1(x), & t > \tau, \end{cases}$$

  where τ is unknown and to be detected as quickly as possible

- Under this hypothesis (change at $k+1$), the likelihood is

$$L_k^{(1)}(x[1:t]) = \prod_{i=1}^{k} f_0(x[i]) \prod_{i=k+1}^{t} f_1(x[i])$$

- Under the no-change hypothesis H_0 (no change up to time t), the likelihood is

$$L^{(0)}(x[1:t]) = \prod_{i=1}^{t} f_0(x[i])$$

Change point detection basic statistics (CUSUM)

- The likelihood ratio for the hypothesis "change at $k+1$" versus "no change" is

$$\begin{aligned}\Lambda_k(t) &= \frac{L_k^{(1)}(x[1:t])}{L^{(0)}(x[1:t])} \\ &= \frac{\prod_{i=1}^{k} f_0(x[i]) \prod_{i=k+1}^{t} f_1(x[i])}{\prod_{i=1}^{t} f_0(x[i])} \\ &= \prod_{i=k+1}^{t} \frac{f_1(x[i])}{f_0(x[i])}\end{aligned}$$

- We do not know the true change time k, so we consider the maximum over all possible changepoints:

$$\Lambda(t) = \max_{0 \le k < t} \Lambda_k(t) = \max_{0 \le k < t} \prod_{i=k+1}^{t} \frac{f_1(x[i])}{f_0(x[i])}$$

Change point detection basic statistics (CUSUM)

- Define the log-likelihood ratio increment and the cumulative sum

$$s_i := \log \frac{f_1(x[i])}{f_0(x[i])}, \quad S_t := \sum_{i=1}^{t} s_i, S_0 := 0$$

- Then the log-likelihood ratio for a change at $k+1$ is

$$\log \Lambda_k(t) = \sum_{i=k+1}^{t} s_i = S_t - S_k$$

- Therefore the likelihood ratio is:

$$\begin{aligned} \log \Lambda(t) &= \max_{0 \le k < t} \log \Lambda_k(t) \\ &= \max_{0 \le k < t} (S_t - S_k) \\ &= S_t - \min_{0 \le k < t} S_k \end{aligned}$$

- Define the running minimum and then for $t \ge 1$,

$$m_t := \min_{0 < k < t} S_k, \log \Lambda(t) = S_t - \min_{0 < k < t} S_k = S_t - m_{t-1}$$

- define the CUSUM statistic

$$C_t := S_t - \min_{0 \le k \le t} S_k.$$

One can show that C_t satisfies the recursion

$$C_t = \max\{0,\ C_{t-1} + s_t\}, \quad C_0 = 0,$$

where $s_t = \log \frac{f_1(x[t])}{f_0(x[t])}$.

- Indeed:
  - $S_t = S_{t-1} + s_t$,
  - if S_t drops below the previous minimum, the minimum becomes S_t and C_t is reset to 0,
  - otherwise $C_t = C_{t-1} + s_t$ remains positive.
- In fact, one can prove that

$$C_t = S_t - \min_{0 \le k \le t} S_k = \max_{0 \le k < t} \left(S_t - S_k\right) = \log \Lambda(t)$$

- CUSUM declares that a change has occurred as soon as C_t exceeds a chosen threshold $h > 0$:

$$T := \inf\{t \geq 1 : C_t \geq h\}$$

Here T is the stopping time (alarm time). The choice of h is used to control the false alarm rate: larger h yields fewer false alarms but later detection, and vice versa.

- CUSUM for the Gaussian case:

$$H_0: \quad x[t] \sim \mathcal{N}(\mu_0, \sigma^2), \quad \text{(in-control)}$$

$$H_1: \quad x[t] \sim \mathcal{N}(\mu_1, \sigma^2), \quad \text{(out-of-control)}$$

  with $\mu_1 > \mu_0$ (upward shift)

- Both μ_0 and μ_1 and the variance σ^2 are assumed known (or estimated).
- The densities are

$$f_0(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2}\right)$$

$$f_1(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)$$

Change point detection basic statistics (CUSUM)

- CUSUM for the Gaussian case
- Log-likelihood ratio increment is

$$
\begin{aligned}
s_i &= \log \frac{f_1(x[i])}{f_0(x[i])} \\
&= \frac{(x[i]-\mu_0)^2}{2\sigma^2} - \frac{(x[i]-\mu_1)^2}{2\sigma^2} \\
&= \frac{\mu_1-\mu_0}{\sigma^2}\left(x[i] - \frac{\mu_0+\mu_1}{2}\right)
\end{aligned}
$$

- What do you think? Are the terms in s_i obvious?

- CUSUM for the Gaussian case
- Log-likelihood ratio increment is

$$
\begin{aligned}
s_i &= \log \frac{f_1(x[i])}{f_0(x[i])} \\
&= \frac{(x[i]-\mu_0)^2}{2\sigma^2} - \frac{(x[i]-\mu_1)^2}{2\sigma^2} \\
&= \frac{\mu_1-\mu_0}{\sigma^2}\left(x[i] - \frac{\mu_0+\mu_1}{2}\right)
\end{aligned}
$$

- What do you think? Are the terms in s_i obvious?
  - when $x[i]$ is close to the mean of the means, nothing useful can be said (half-way between the changes)
  - when the difference between the means is small, nothing useful can be said (highly sensitive)

Change point detection basic statistics (CUSUM)

- CUSUM for the Gaussian case
- Log-likelihood ratio increment is

$$\begin{aligned} s_i &= \log \frac{f_1(x[i])}{f_0(x[i])} \\ &= \frac{(x[i]-\mu_0)^2}{2\sigma^2} - \frac{(x[i]-\mu_1)^2}{2\sigma^2} \\ &= \frac{\mu_1-\mu_0}{\sigma^2}\left(x[i] - \frac{\mu_0+\mu_1}{2}\right) \end{aligned}$$

- To detect both upward and downward shifts in the mean, we maintain two CUSUM statistics:

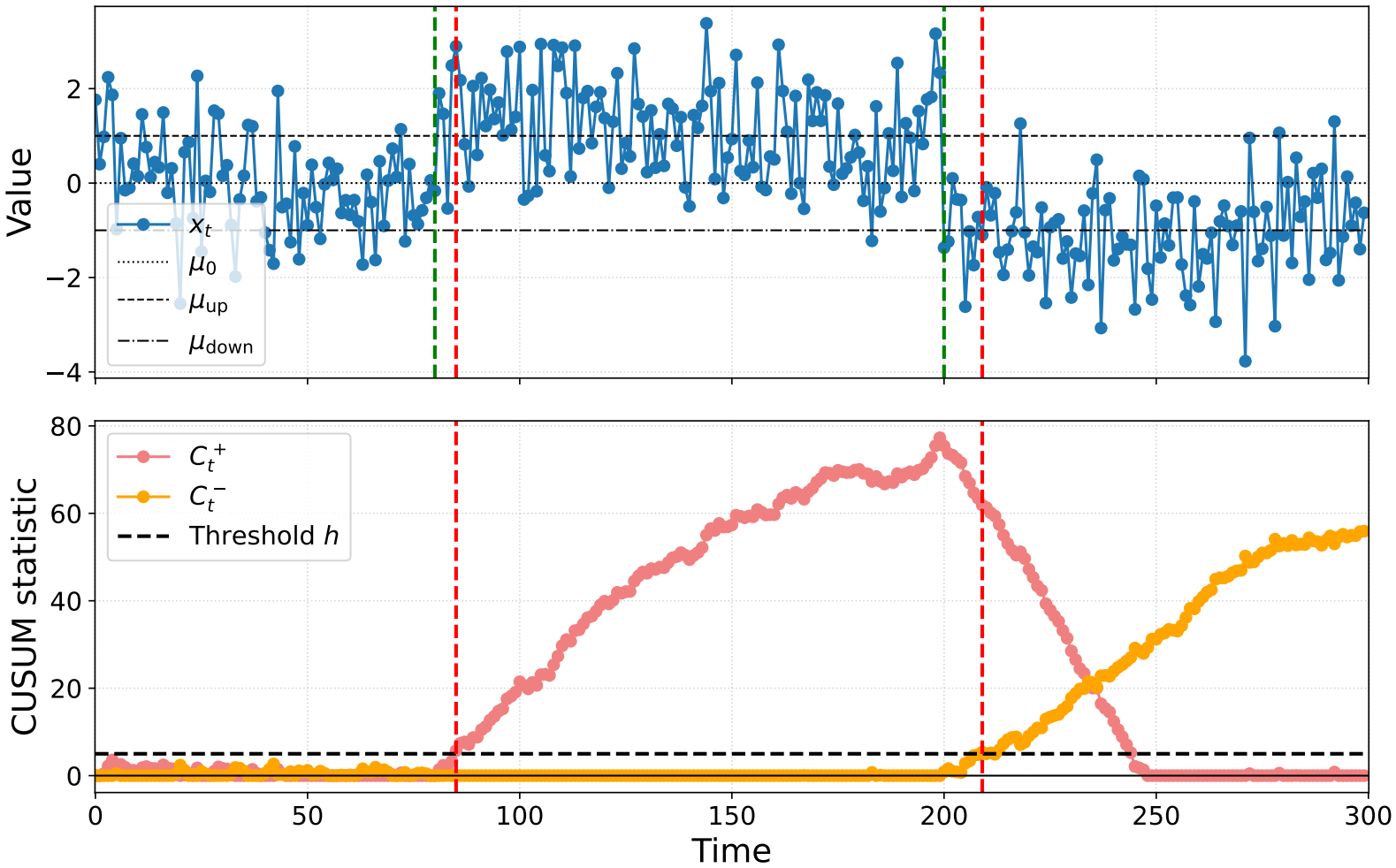
$$C_t^+ = \max\{0,\ C_{t-1}^+ + s_t\},\ C_t^- = \max\{0,\ C_{t-1}^- - s_t\}$$

- The stopping rule is:

Upward alarm if $C_t^+ > h$

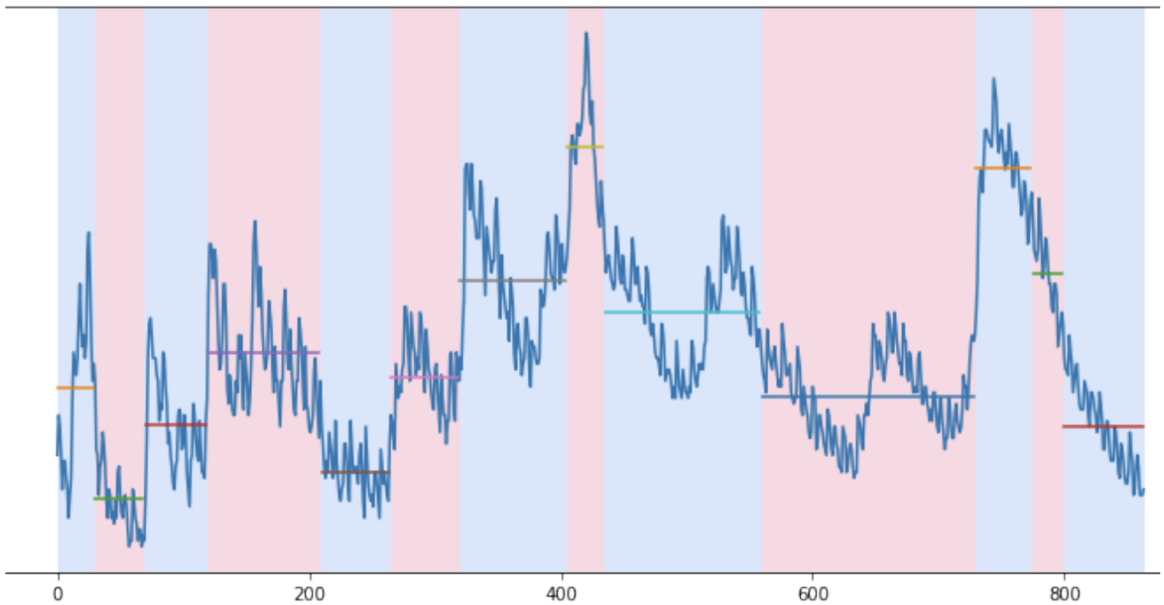
Downward alarm if $C_t^- > h$

CUSUM example



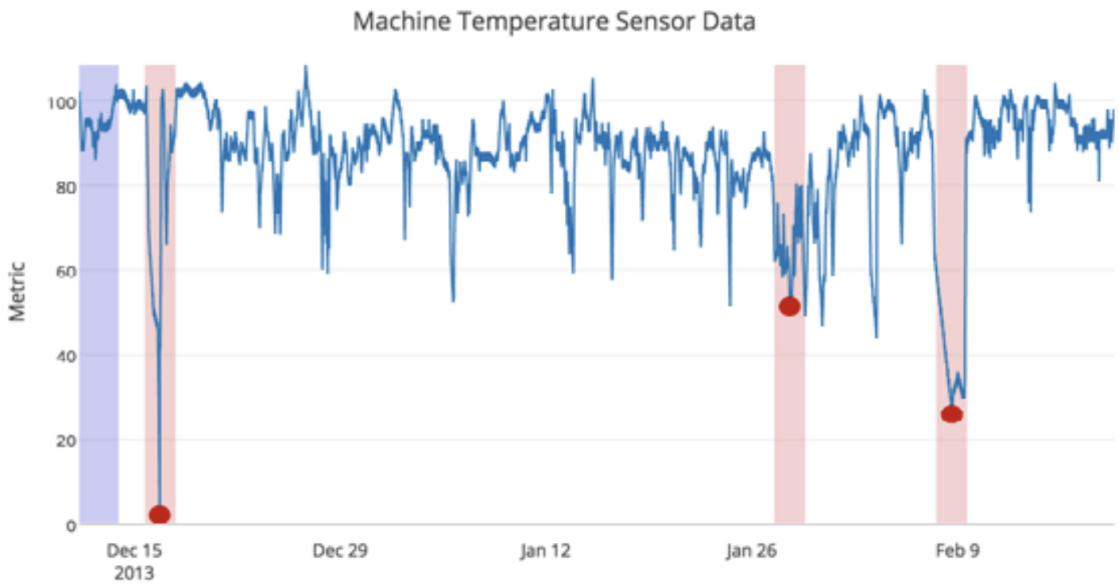
Problem: given a time series, retrieve K points in the time series where a significant change occurs (source: US unemployment data)

- find how many change points there are
- tell us where these change points are



Problem: given a time series, retrieve the points where something unusual happens

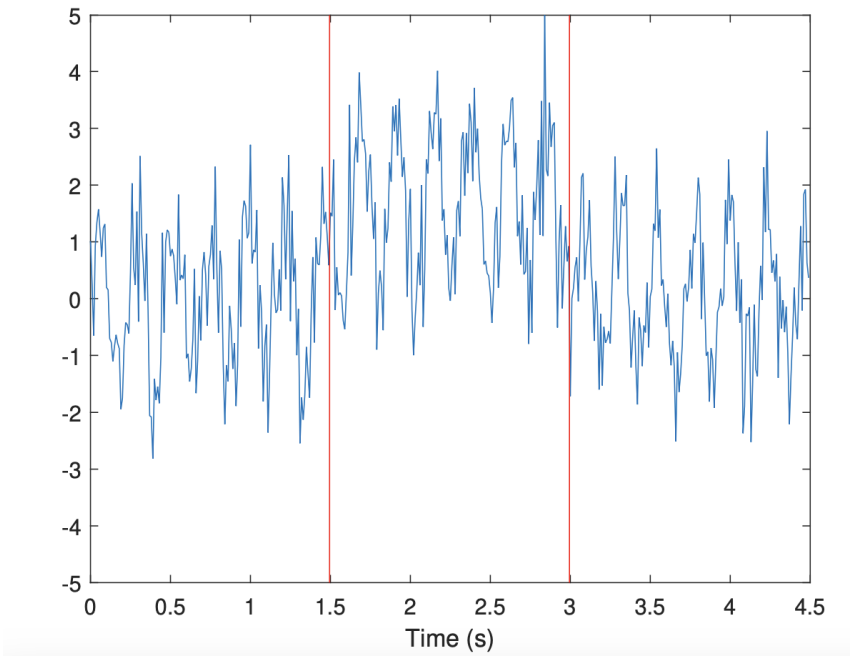
- what is the definition of unusual?
- how many unusual points are you looking for?



Change point detection

Goals:

- find the abrupt changes in the time series $x[n]$
- find the time at which these happen
- find how many there are
- we call the set of times where an abrupt change happens $\mathcal{T}^\star$



Problem statement

$$(\hat{t}_1, \ldots, \hat{t}_K) = \arg\min_{t_1,\ldots,t_K} \sum_{k=1}^{K} c(x[t_k : t_{k+1}]) \tag{1}$$

We have made the following notation:

- $t_k : t_{k+1}$ is Matlab notation for the set $\{t_k, t_k + 1, \ldots, t_{k+1} - 1\}$
- c is a cost function that measures homogeneity
- what are some good picks for the cost function c?

Problem statement

$$(\hat{t}_1, \ldots, \hat{t}_K) = \arg\min_{t_1,\ldots,t_K} \sum_{k=1}^{K} c(x[t_k : t_{k+1}]) \tag{1}$$

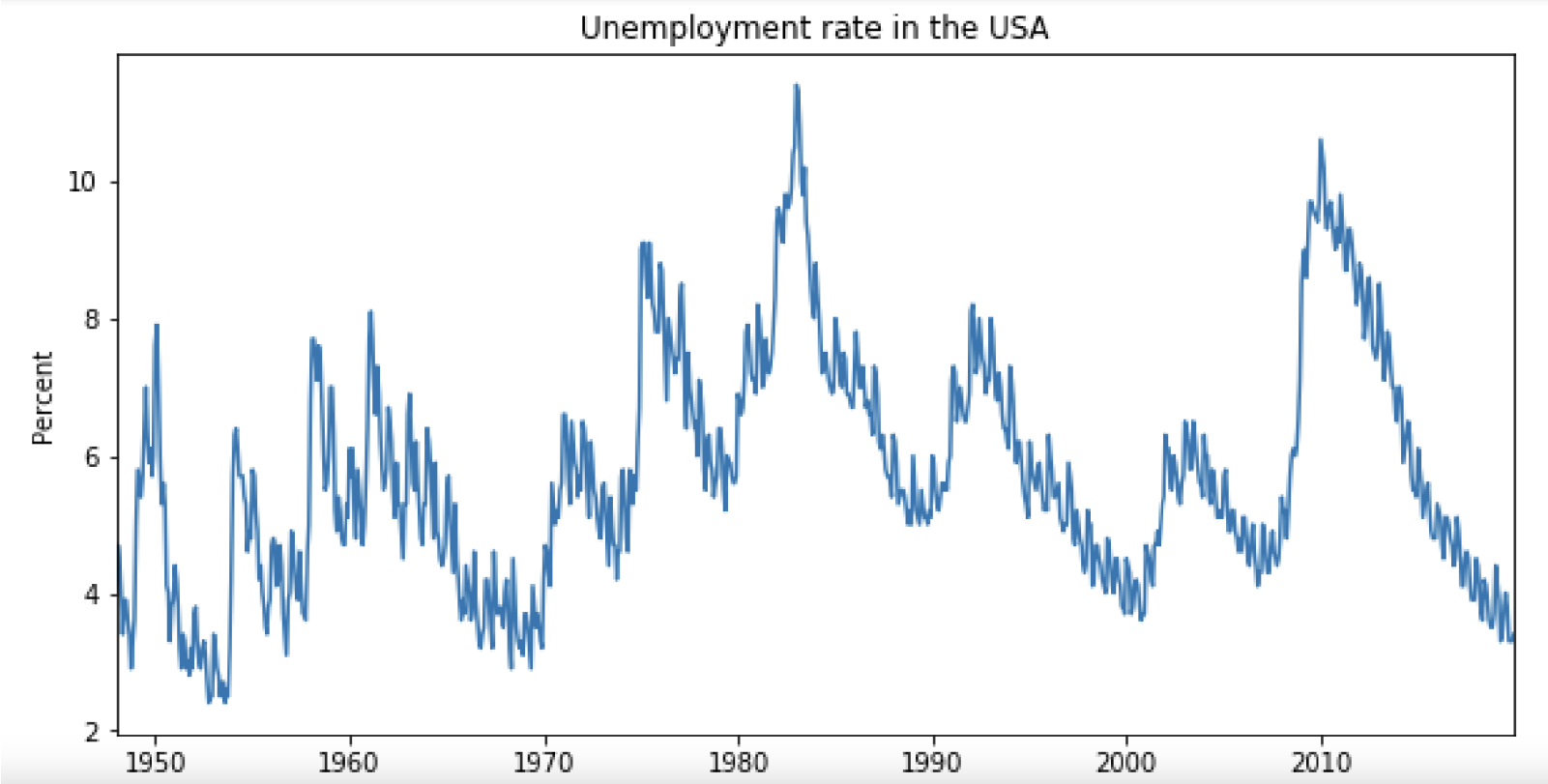
We have made the following notation:

- $t_k : t_{k+1}$ is Matlab notation for the set $\{t_k, t_k + 1, \ldots, t_{k+1} - 1\}$
- c is a cost function that measures homogeneity
- what are some good picks for the cost function c?
  - log likelihood term
  - the mean
  - the median
  - the error (RMSE) of a linear model
  - the error (RMSE) of a more sophisticated linear model

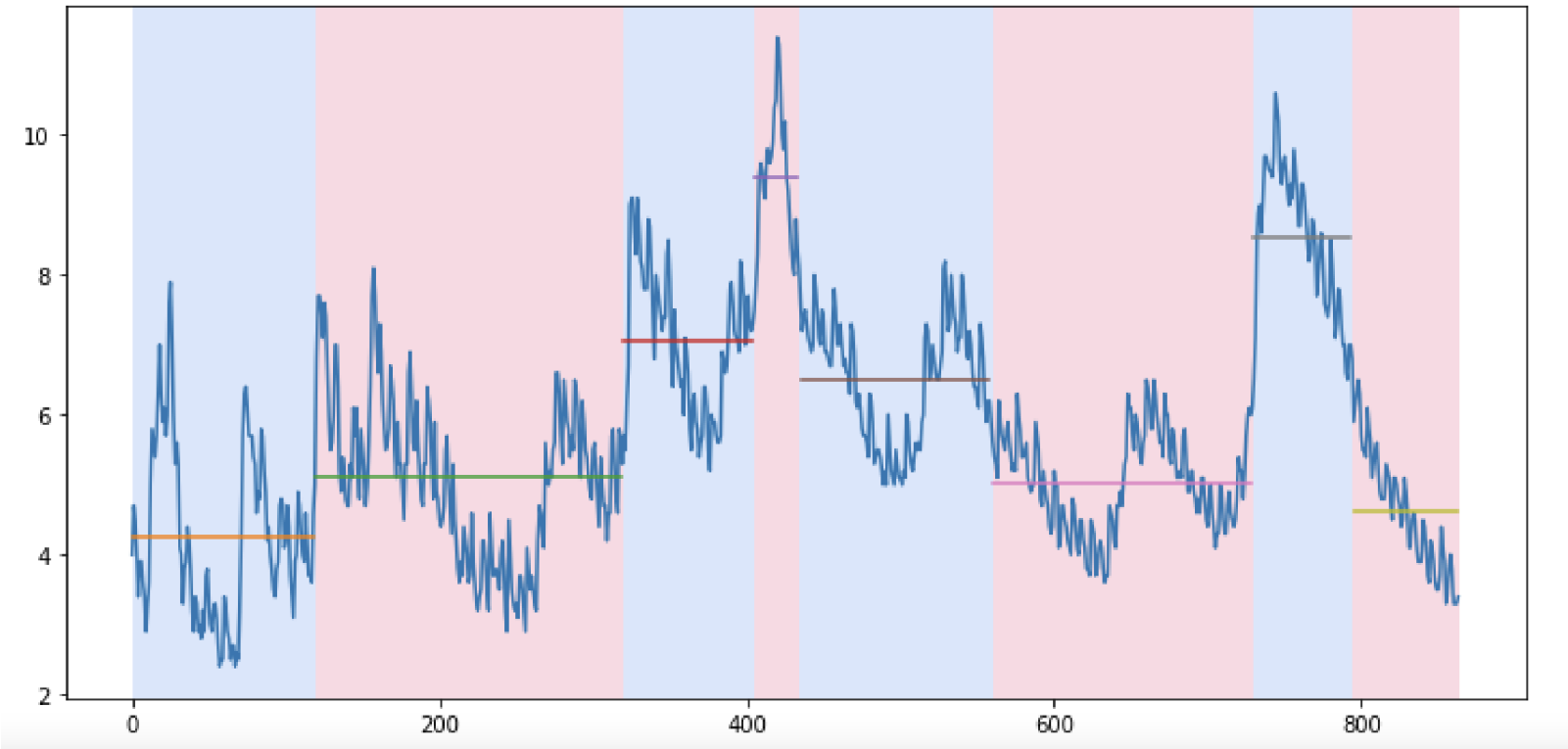
Choices for the cost function:

1. $c_{\text{prob}}(t_k : t_{k+1}) = -\max_{\theta} \sum_{n=t_k}^{t_{k+1}-1} \log p(x[n]|\theta)$
2. $c_{\text{L2}}(t_k : t_{k+1}) = \sum_{n=t_k}^{t_{k+1}-1} \|x[n] - \mu_{t_k:t_{k+1}}\|_2^2$
3. $c_{\Sigma}(t_k : t_{k+1}) = (t_{k+1} - t_k) \log \sigma^2_{t_k:t_{k+1}} + \frac{1}{\sigma^2_{t_k:t_{k+1}}} \sum_{n=t_k}^{t_{k+1}-1} \|x[n] - \mu_{t_k:t_{k+1}}\|_2^2$
4. $c_{\text{lin}}(t_k : t_{k+1}) = \min_{\alpha} \sum_{n=t_k}^{t_{k+1}-1} \|x[n] - \sum_{i=1}^{M} \alpha_i x[n-i]\|_2^2$

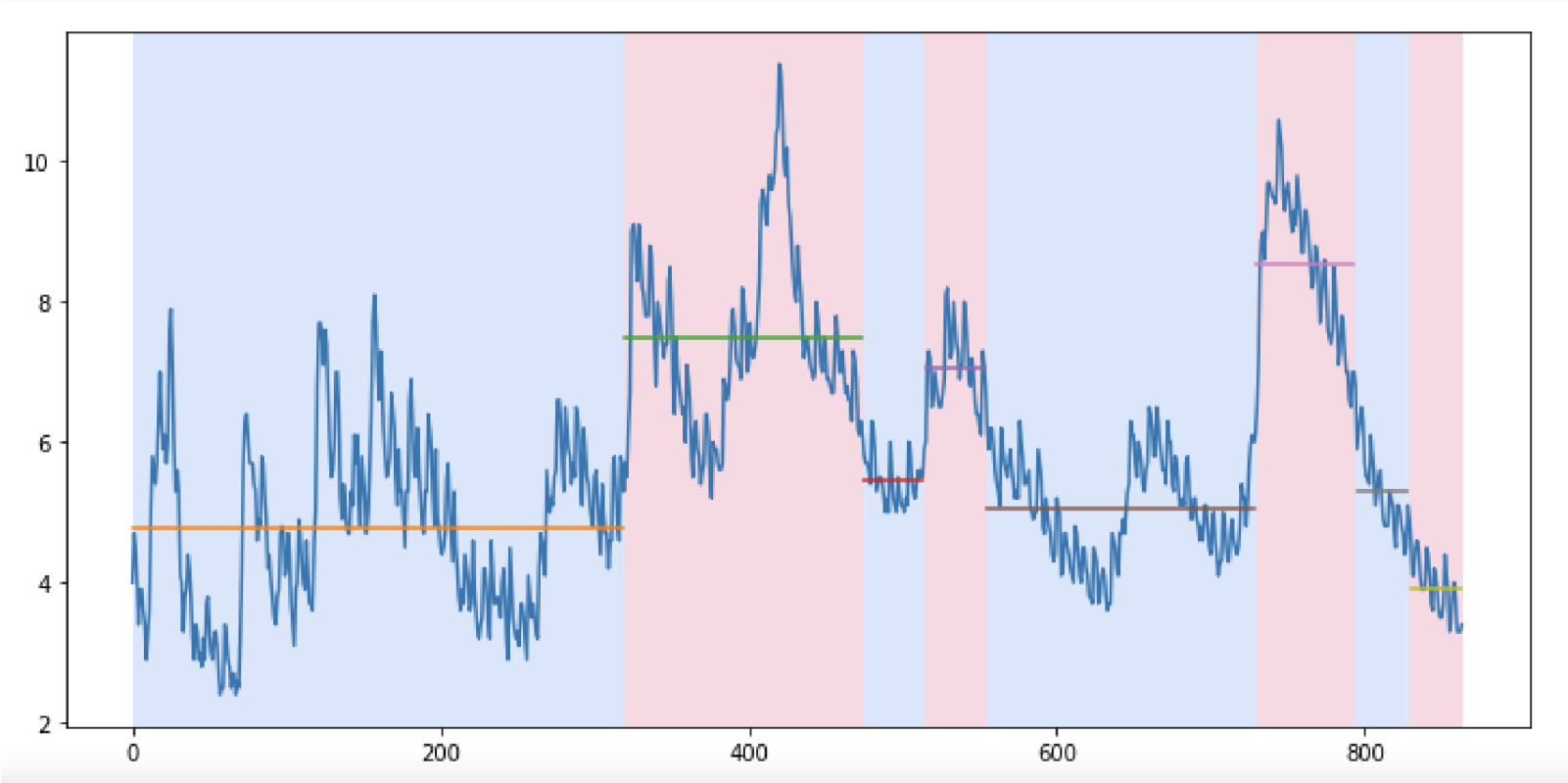
US Unemployment



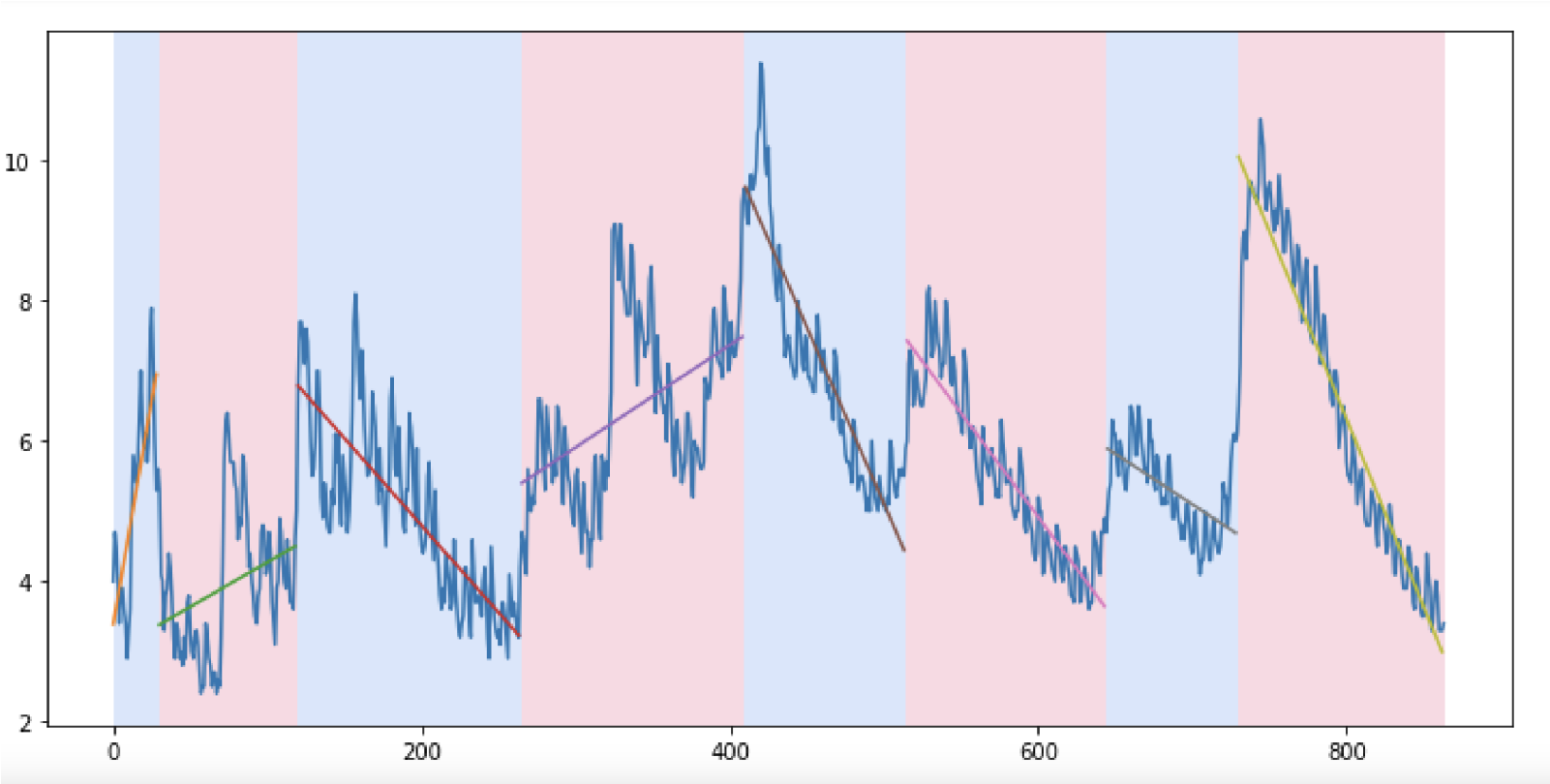
US Unemployment - $c_{L2}, K = 7$



US Unemployment - $c_{\Sigma}, K = 7$



US Unemployment - $c_{\text{lin}}, K = 7$



Solve optimally by dynamic programming

For

$$\mathcal{V}(\mathcal{T}, \mathbf{x}) = \sum_{k=0}^{K} c(x[t_k : t_{k+1}])$$

we have that

$$\begin{aligned}
\min_{|\mathcal{T}|=K} V(\mathcal{T}, \mathbf{x}) &= \min_{0=t_0<t_1<\cdots<t_K<t_{K+1}=N} \sum_{k=0}^{K} c(x[t_k : t_{k+1}]) \\
&= \min_{t \leq N-K} \left[c(x[0:t]) + \min_{t_0=t<t_1<\cdots<t_{K-1}<t_K=N} \sum_{k=0}^{K-1} c(x[t_k : t_{k+1}]) \right] \\
&= \min_{t \leq N-K} \left[c(x[0:t]) + \min_{|\mathcal{T}|=K-1} V(\mathcal{T}, x[t:N]) \right]
\end{aligned}$$

Complexity?

Change point detection: the algorithm

Solve optimally by dynamic programming

For

$$\mathcal{V}(\mathcal{T}, \mathbf{x}) = \sum_{k=0}^{K} c(x[t_k : t_{k+1}])$$

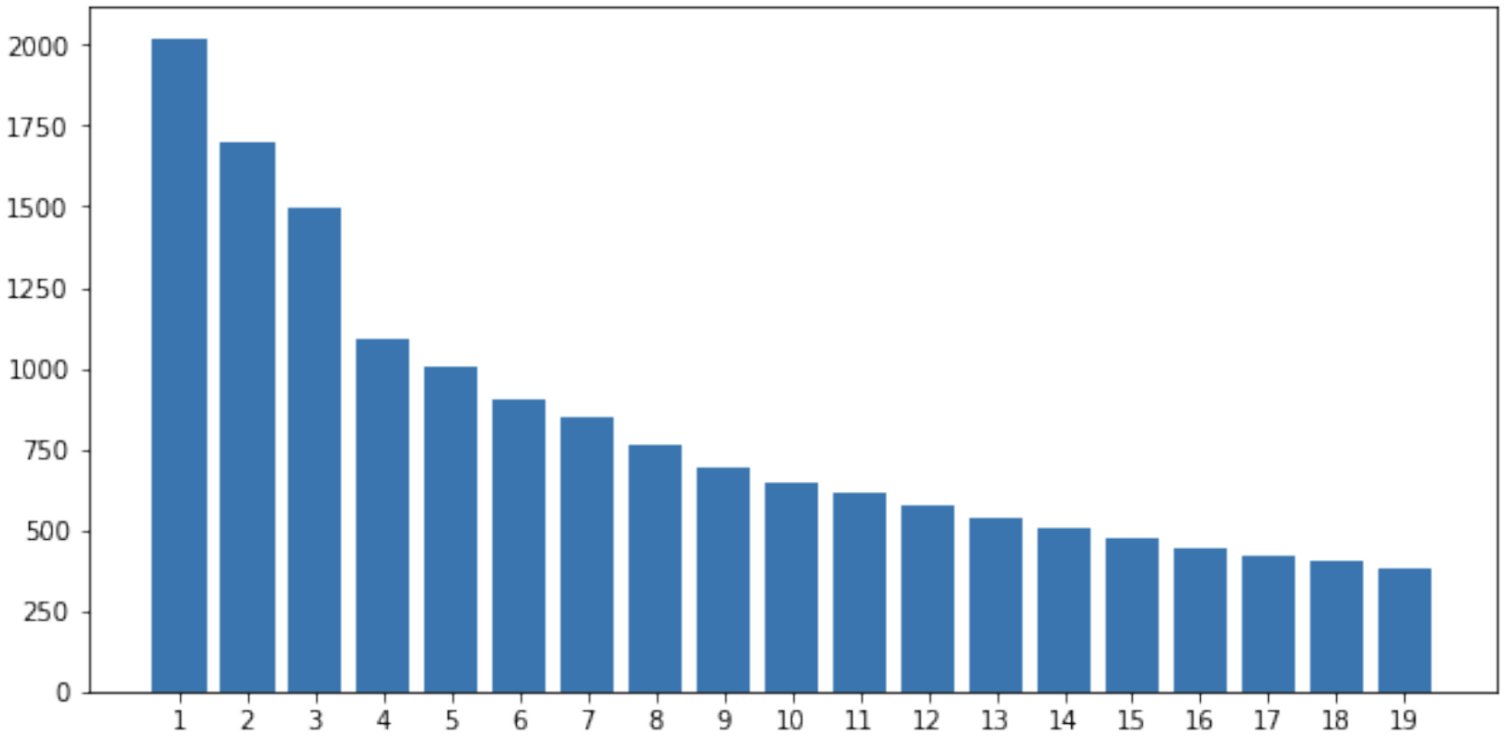
we have that

$$\begin{aligned}
\min_{|\mathcal{T}|=K} V(\mathcal{T}, \mathbf{x}) &= \min_{0=t_0<t_1<\cdots<t_K<t_{K+1}=N} \sum_{k=0}^{K} c(x[t_k : t_{k+1}]) \\
&= \min_{t \le N-K} \left[c(x[0:t]) + \min_{t_0=t<t_1<\cdots<t_{K-1}<t_K=N} \sum_{k=0}^{K-1} c(x[t_k : t_{k+1}]) \right] \\
&= \min_{t \le N-K} \left[c(x[0:t]) + \min_{|\mathcal{T}|=K-1} V(\mathcal{T}, x[t:N]) \right]
\end{aligned}$$

Complexity? $O(KN^2)$

Change point detection: finding optimum K

- run for all values K from 1 to $K_{\max}$
- find an “elbow” in the resulting curve



how would you integrate the K into the problem itself?

New, regularized problem statement (penalized change point detection)

$$(\hat{t}_1, \ldots, \hat{t}_K) = \underset{t_1,\ldots,t_K}{\arg\min} \sum_{k=1}^{K} c(x[t_k : t_{k+1}]) + \lambda K \tag{2}$$

- this type of regularization is typical in machine learning
- the size of the solution set $\mathcal{T}$ is taken into account at each step of the algorithm
- many algorithms have been proposed for this task
- new problem: find $\lambda \in \mathbb{R}_+$ (in general there is no clear formula between λ and K)

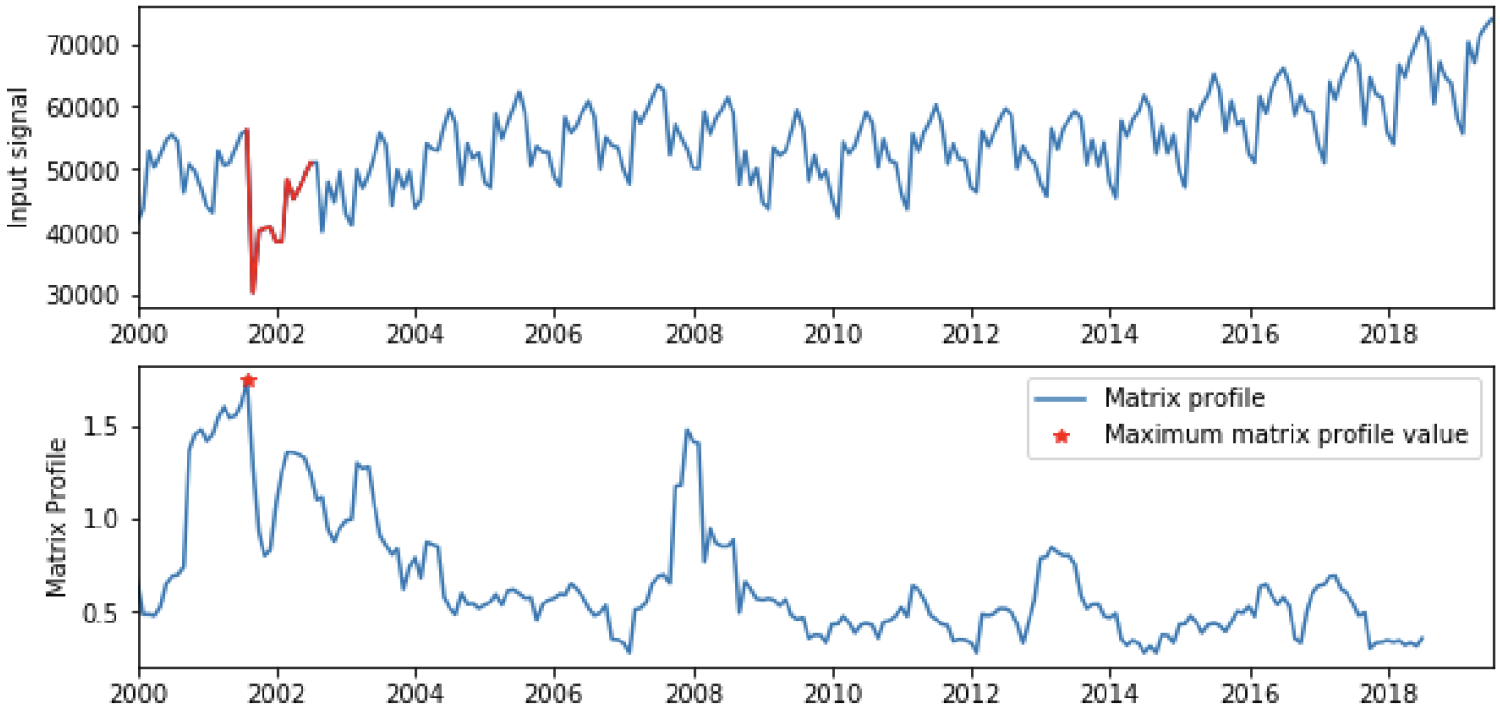
Anomaly detection

After we have the change point splitting done we can do several things:

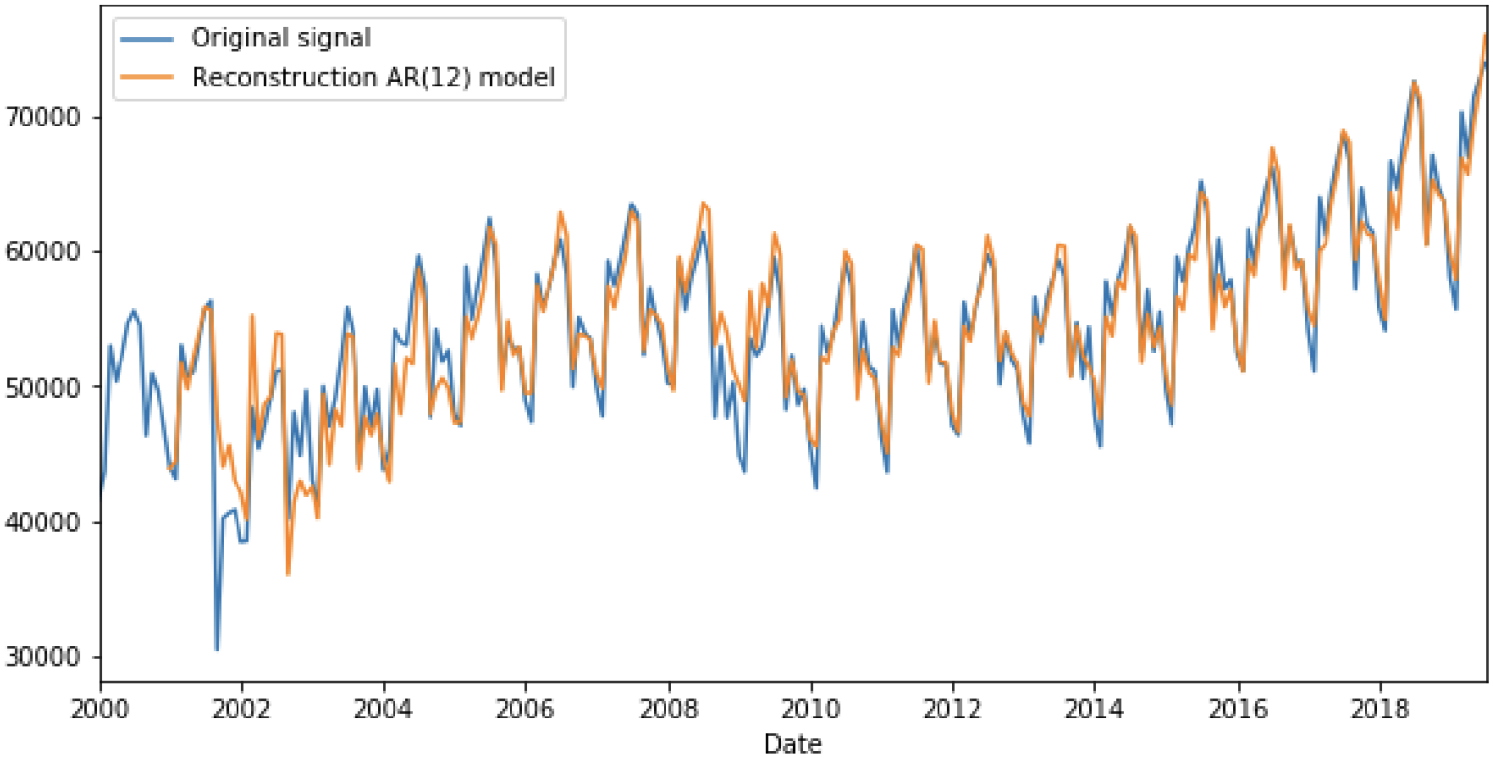
- for the mean and statistical cost functions c_{L2} and c_{Σ} we can use methods developed in Lecture 5 Statistical algorithms: truncation, LODA
- for the regression (linear) statistical cost function c_{lin} we can use the leverage scores developed in Lecture 2 Leverage scores for linear regression
- the third option is to use an adaptive model that changes with the time series

Given a chunk of size L try to find in the time series a similar pattern

$$\text{pattern}[n] = \min d(x[n : n+L-1], x[i : i+L-1]) \text{ for } i < n-L \text{ and } i > n+L \quad (3)$$

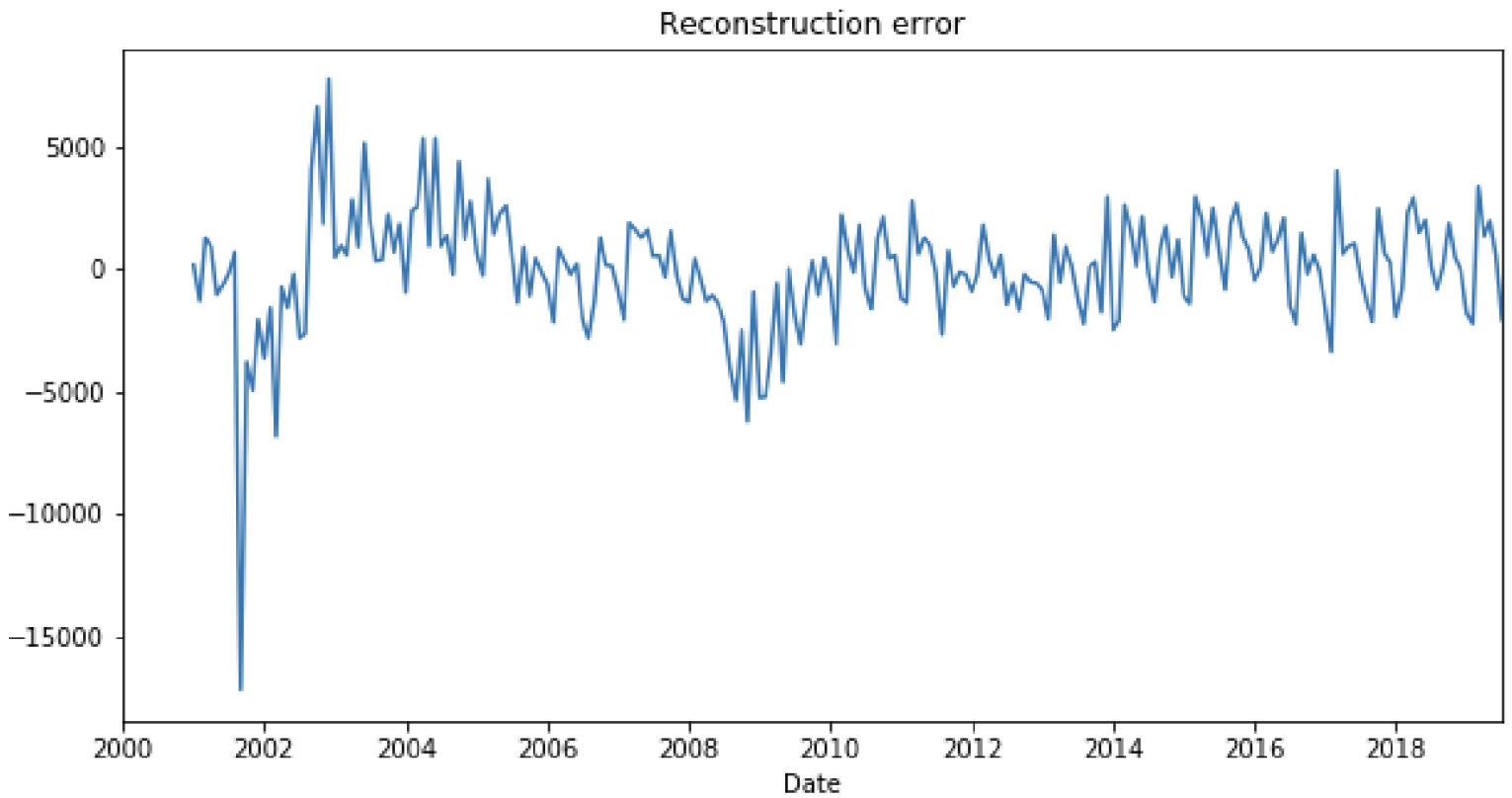


Model based: AR model reconstruction



Anomaly detection: reconstruction

Model based: AR model error



Anomaly detection by reconstruction error (forecasting error)

- The AR model (and more generally ARMA model):

$$\hat{x}[n] = \sum_{i=1}^{W} w[i]x[n-i] \tag{4}$$

- RNN/LSTM/Transformer

$$\hat{x}[n] = f_\theta(x[n-1:n-W-1]), \sum_{n=1} \|x[n] - f_\theta(x[n-W-1:n-1])\|_2^2 \tag{5}$$

- PCA/Autoencoder

$$\hat{x}[n] = g_\phi(f_\theta(x[n])), \sum_{n=1} \|x[n] - g_\phi(f_\theta(x[n]))\|_2^2 \tag{6}$$

The end.