

“Space Data Processing: Making Sense of Experimental Data”

Topic 2

"Quasi-optimal approximation under uncertainty"

Tatiana Podladchikova Rupert Gerzer

Term 4, March 28 – May 27, 2016

t.podladchikova@skoltech.ru

The basis of statistical analysis

```
graph TD; A[The basis of statistical analysis] --> B[Least-square method and linear regression]; B --> C[The LSM method leads to divergence and loses its practical value when a model is inadequate or unknown.]; B --> D[Linear regression doesn't provide reliable long-term forecasting];
```

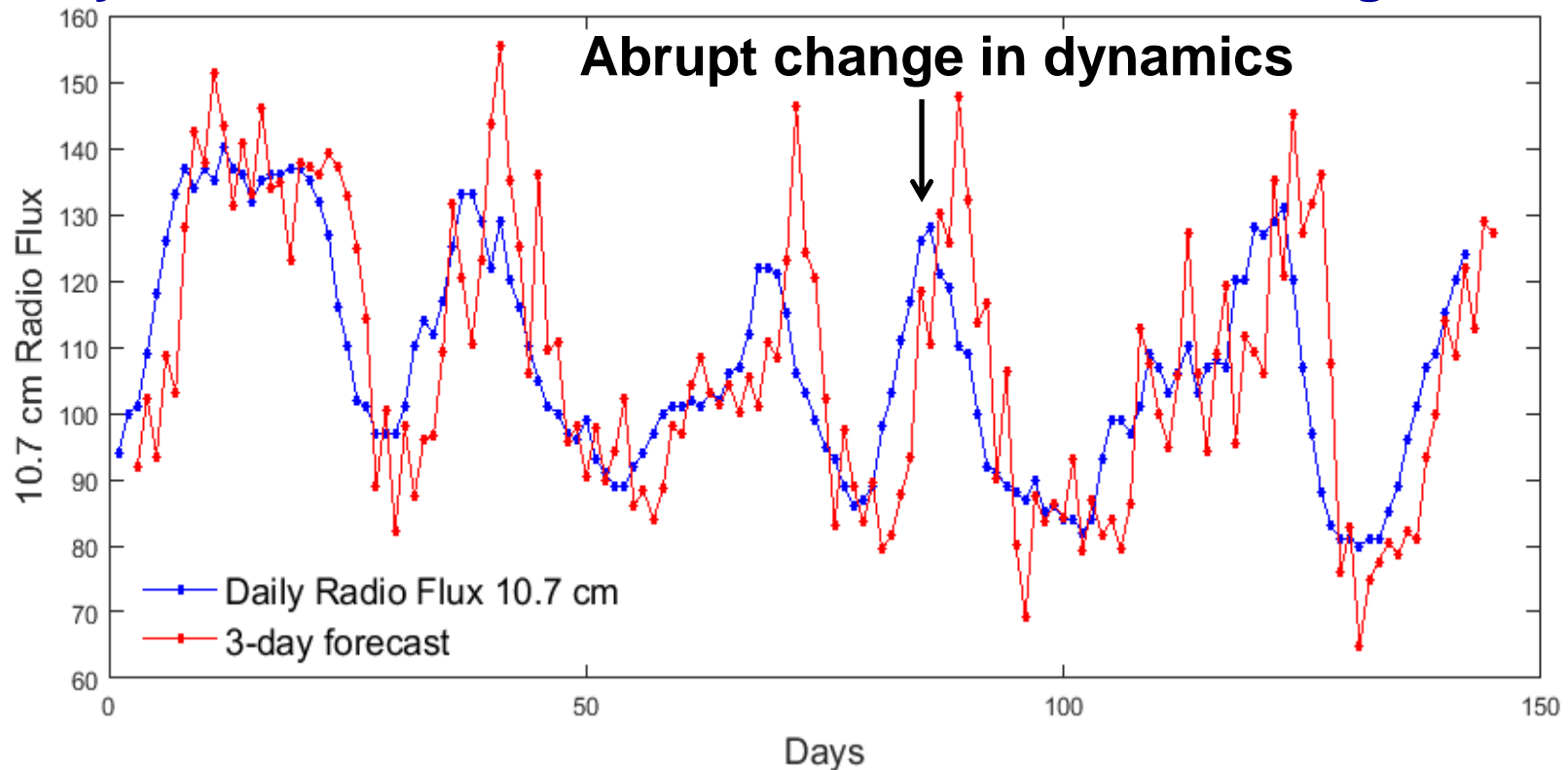
**Least-square method
and linear regression**

**The LSM method leads to
divergence and loses its
practical value when a model
is inadequate or unknown.**

**Linear regression doesn't
provide reliable long-term
forecasting**

Linear regression doesn't provide long-term forecasting

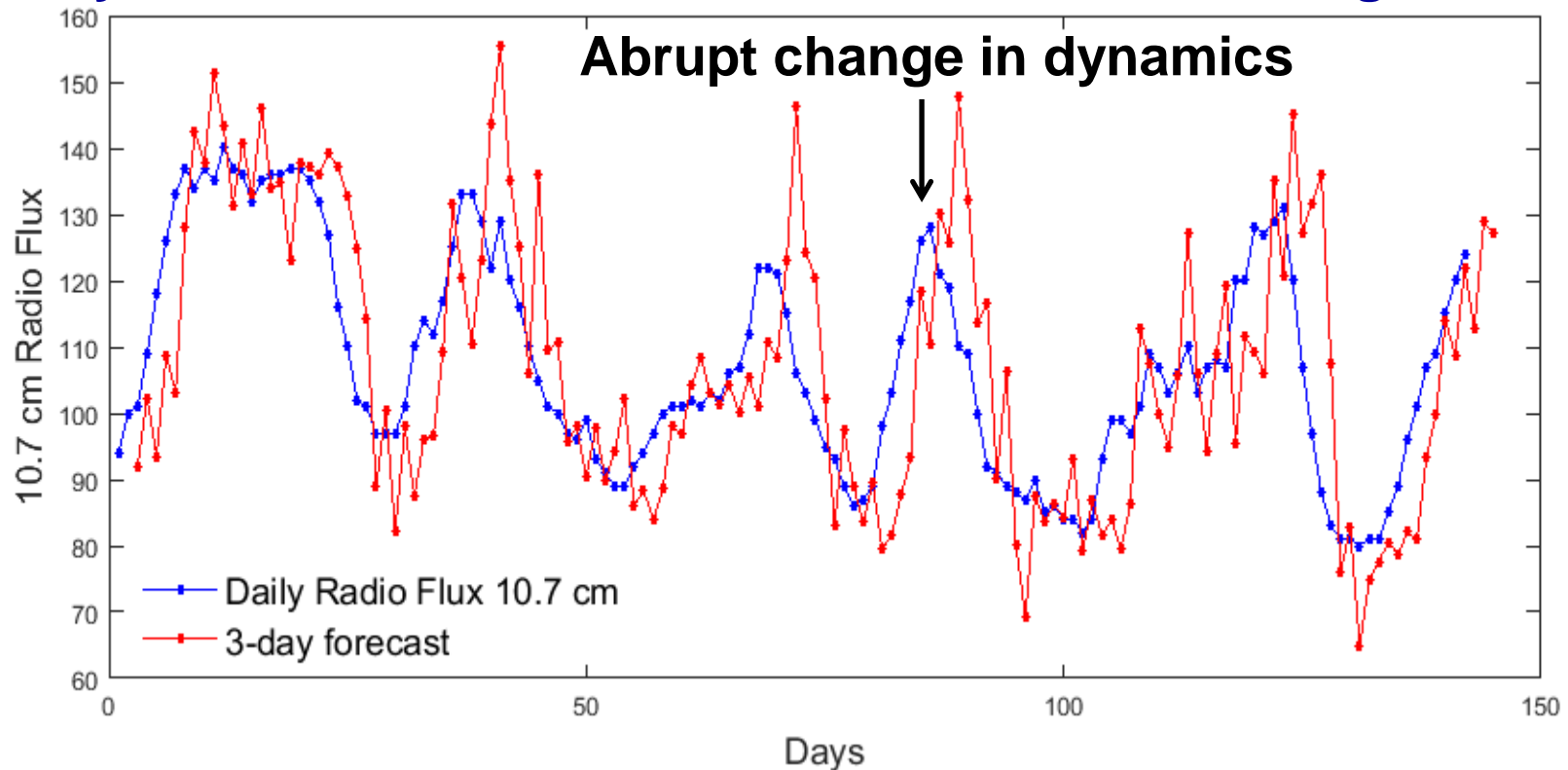
3-day 10.7 cm radio flux forecast based on a linear regression



**Changes in dynamics
of a process leads to great
increase of forecasting errors**

Linear regression doesn't provide long-term forecasting

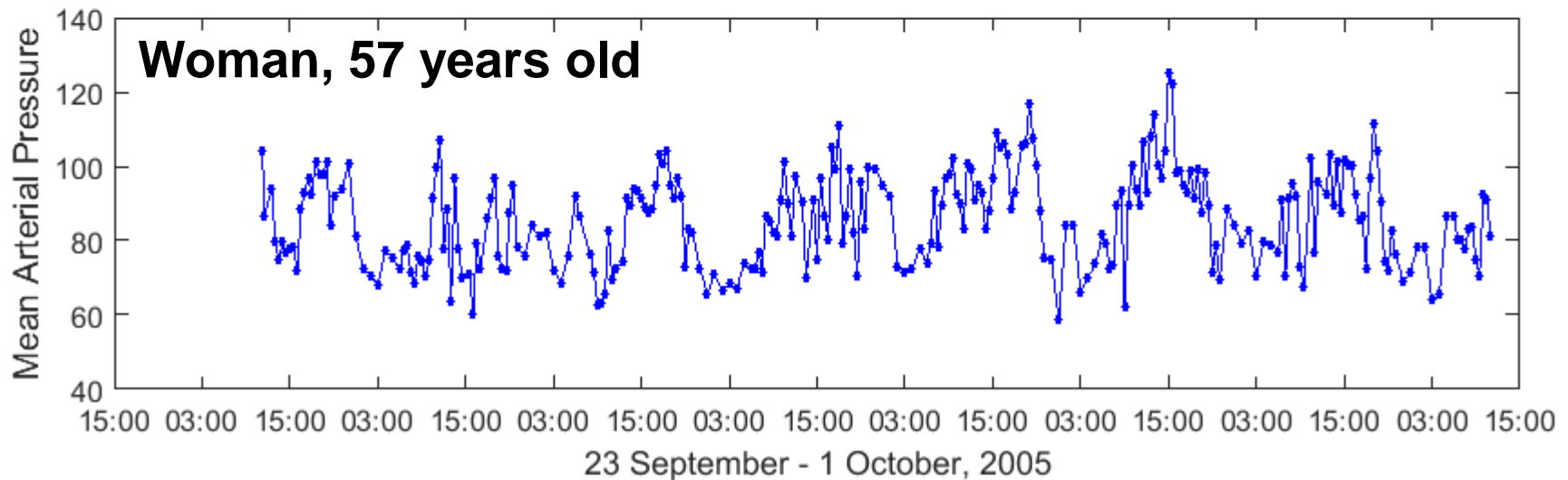
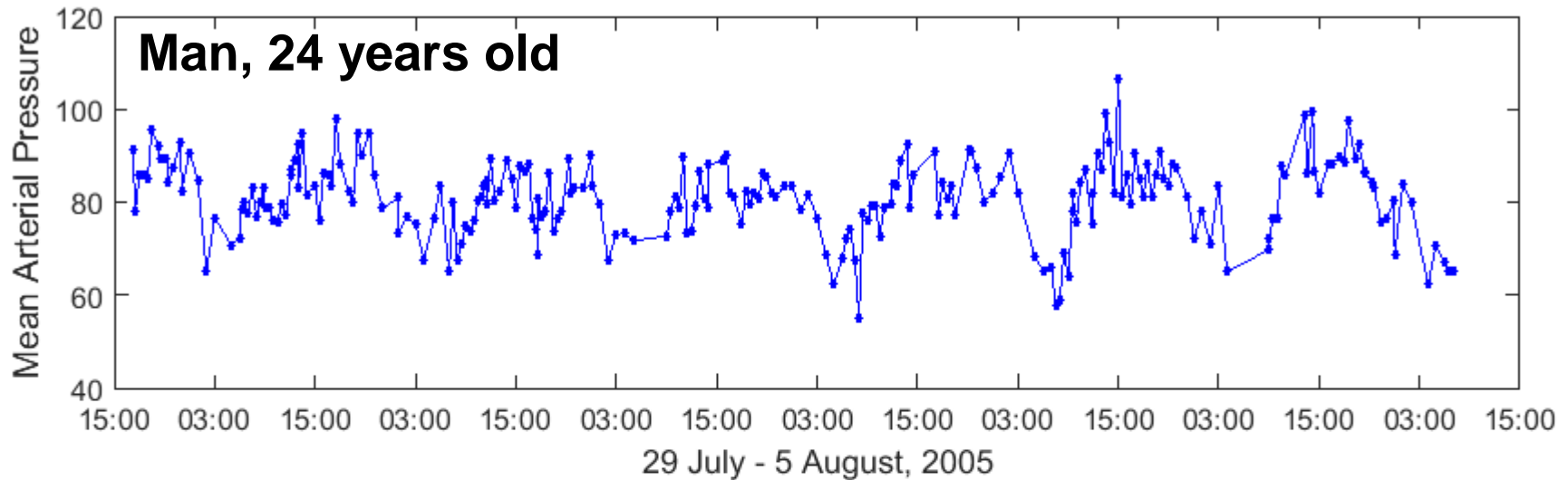
3-day 10.7 cm radio flux forecast based on a linear regression



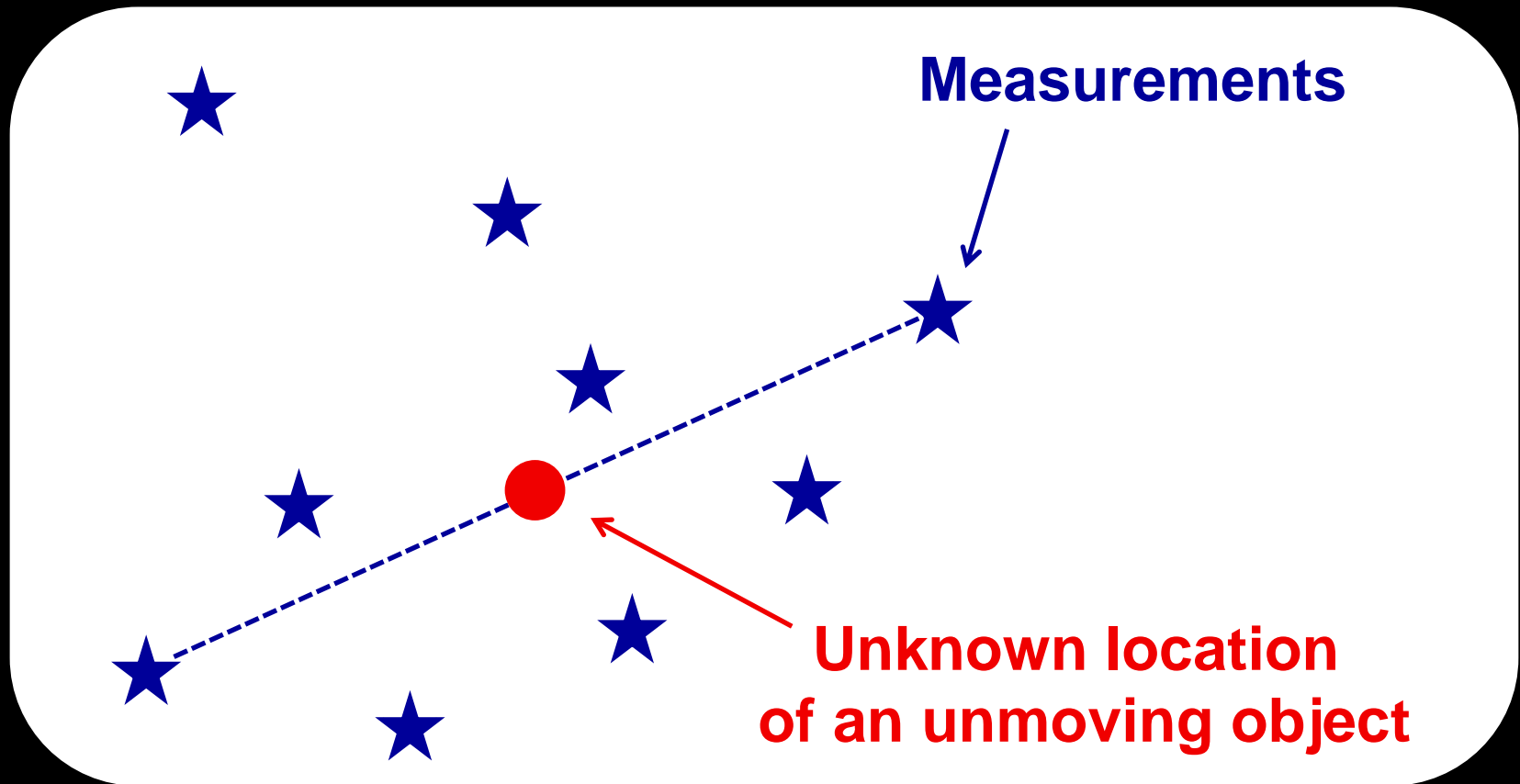
**Changes in dynamics
of a process leads to great
increase of forecasting errors**

**To extract regularities that will
allow long-term forecasting
we need to smooth data**

Which regularities can you extract from measurements of mean arterial pressure?



Estimate the location of an unmoving object



Smoothing is weighted averaging of noisy data.
Fluctuation components are self compensated.

The most popular methods of quasi-optimal estimation

1

**Running
mean**

2

**Exponential
smoothing**



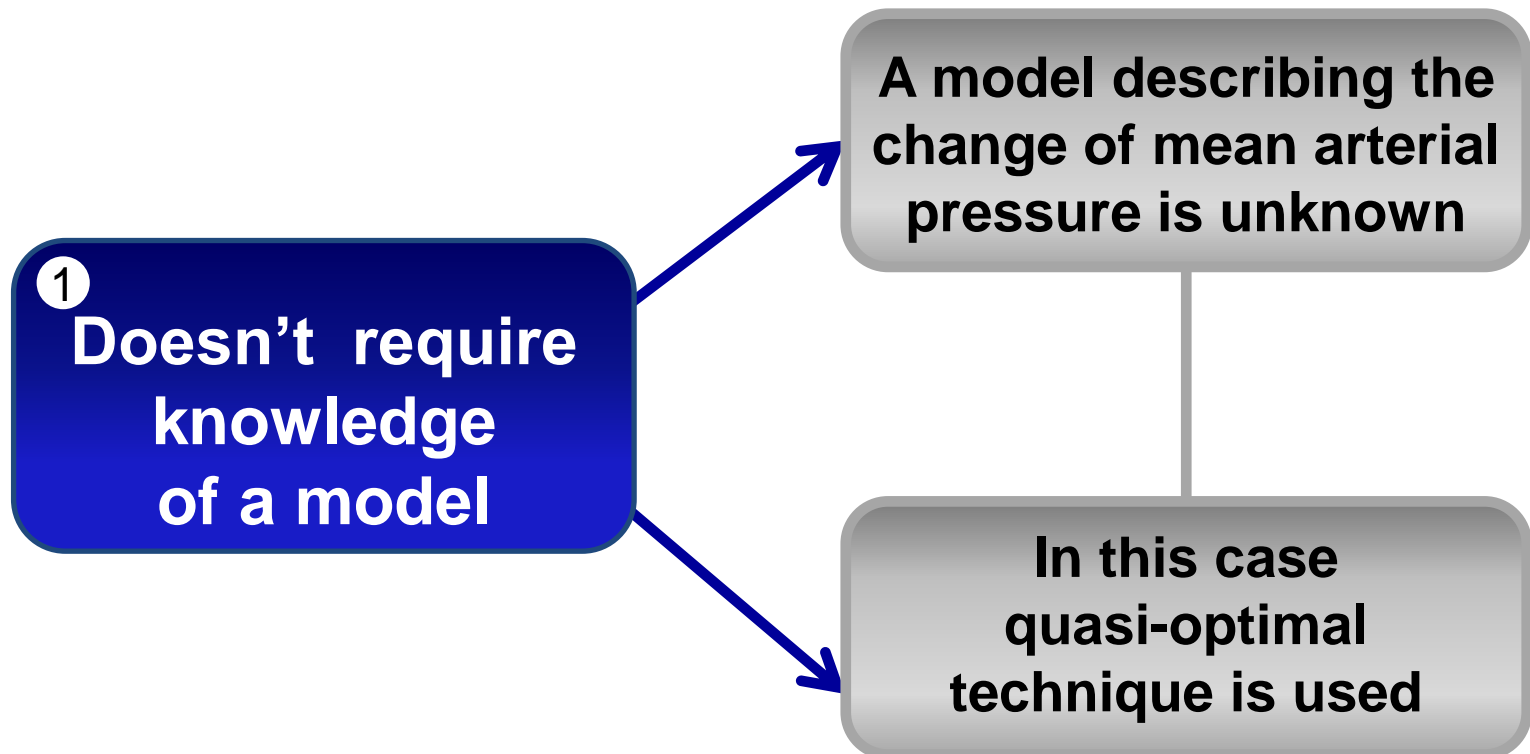
Advantages



Limitations



Advantages of quasi-optimal estimation methods





Advantages of quasi-optimal estimation methods

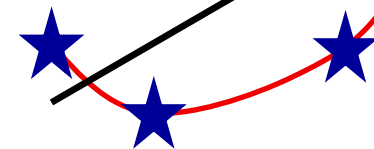
2

Robustness



No risk of divergence

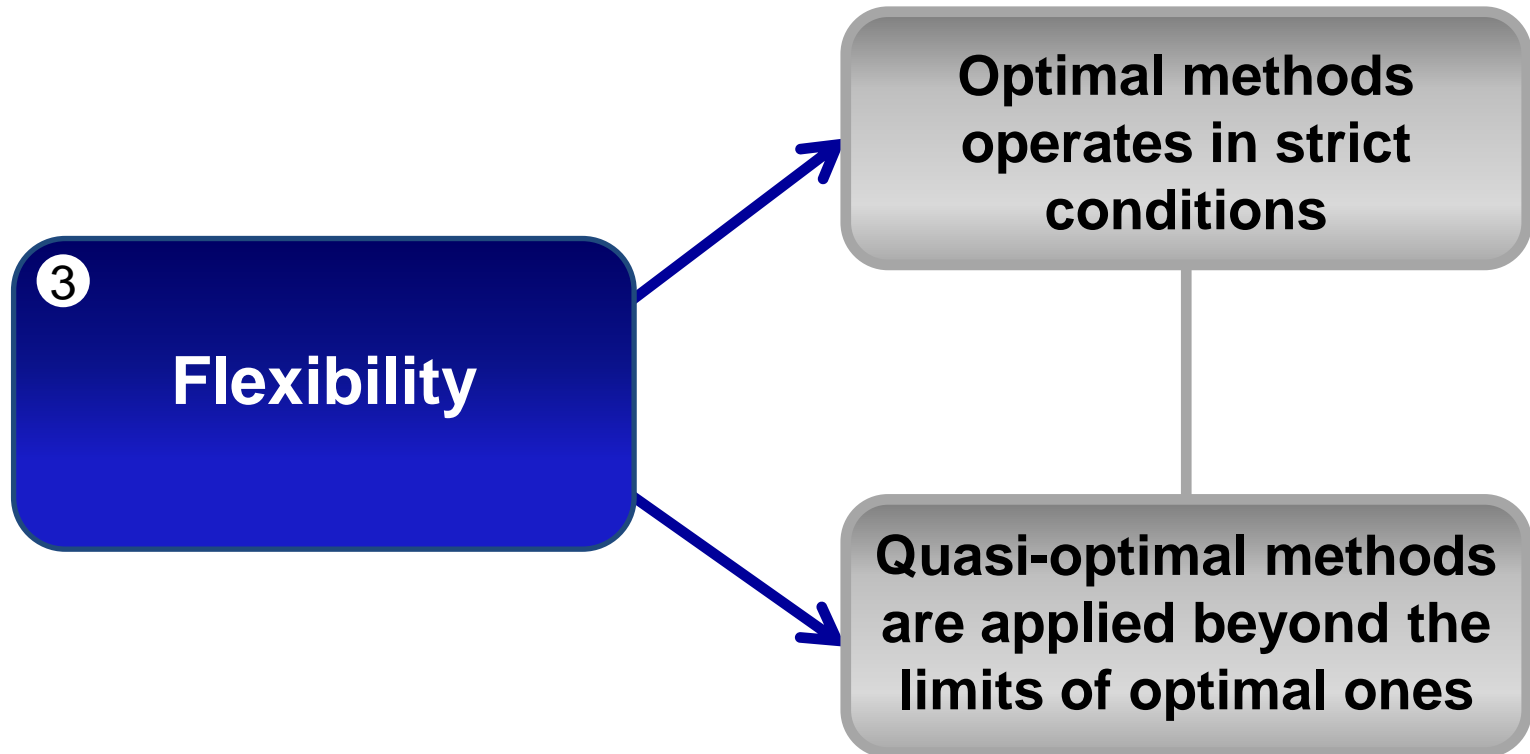
**Optimal estimation
in conditions
of inadequate
model**



**Divergence. Errors
monotonously
increase**



Advantages of quasi-optimal estimation methods





Disadvantages of quasi-optimal estimation methods

1

Unknown estimation errors



Risk of false conclusions



It is needed to search for ways of accuracy estimation

Quasi-optimal approximation under uncertainty

```
graph TD; A[Quasi-optimal approximation under uncertainty] --> B[Learning goals]; B --> C[Analyze conditions for which methods provide effective solution and conditions under which they break down.]; B --> D[Choose the most effective method in conditions of uncertainty];
```

Learning goals

**Analyze conditions for which
methods provide effective
solution and conditions under
which they break down.**

**Choose the most effective
method in conditions
of uncertainty**

① Running mean

$$z_i = X_i + \eta_i$$

z_j
Measurements

X_i
True process
to be
estimated

η_i
Uncorrelated
unbiased noise
with variance
 σ_η^2

① Running mean

$$z_i = X_i + \eta_i$$

z_j
Measurements

X_i
True process
to be
estimated

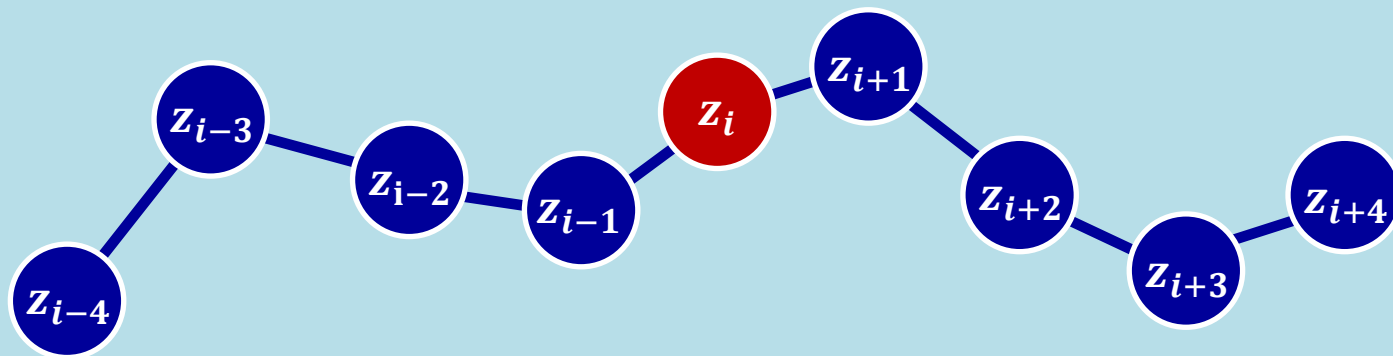
η_i
Uncorrelated
unbiased noise
with variance
 σ_η^2

Our goal

To reconstruct the dynamics of process X_i
using available measurements z_j when
the dynamical model is unknown

1 Running mean

Window size $M = 9$



Last 9 measurements z_i

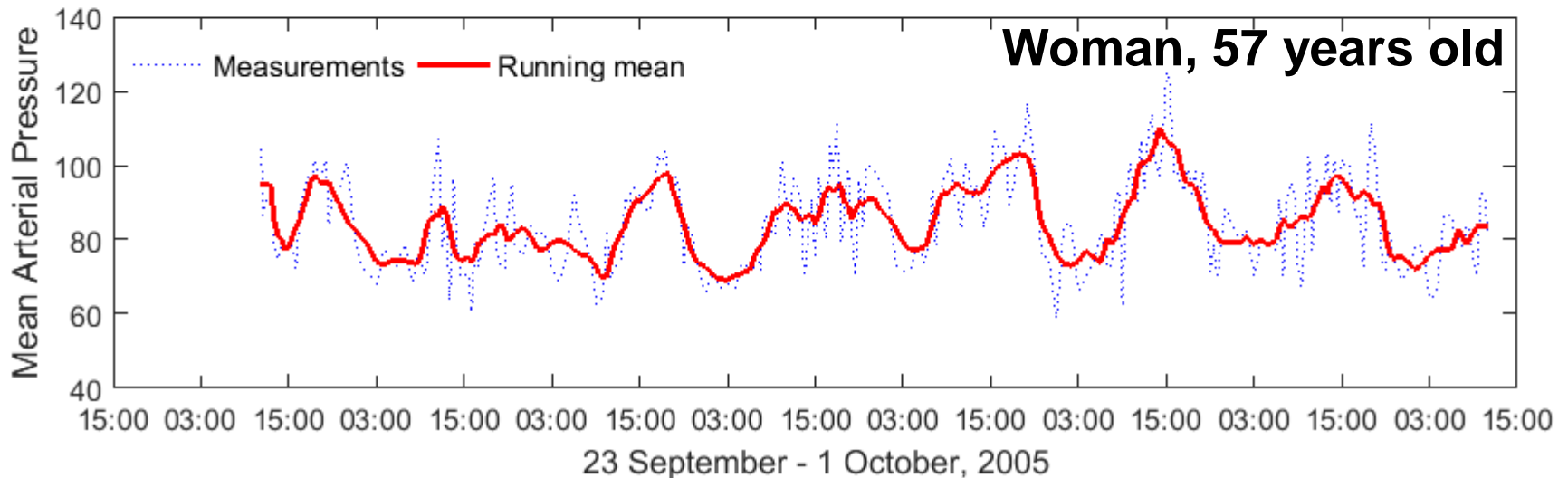
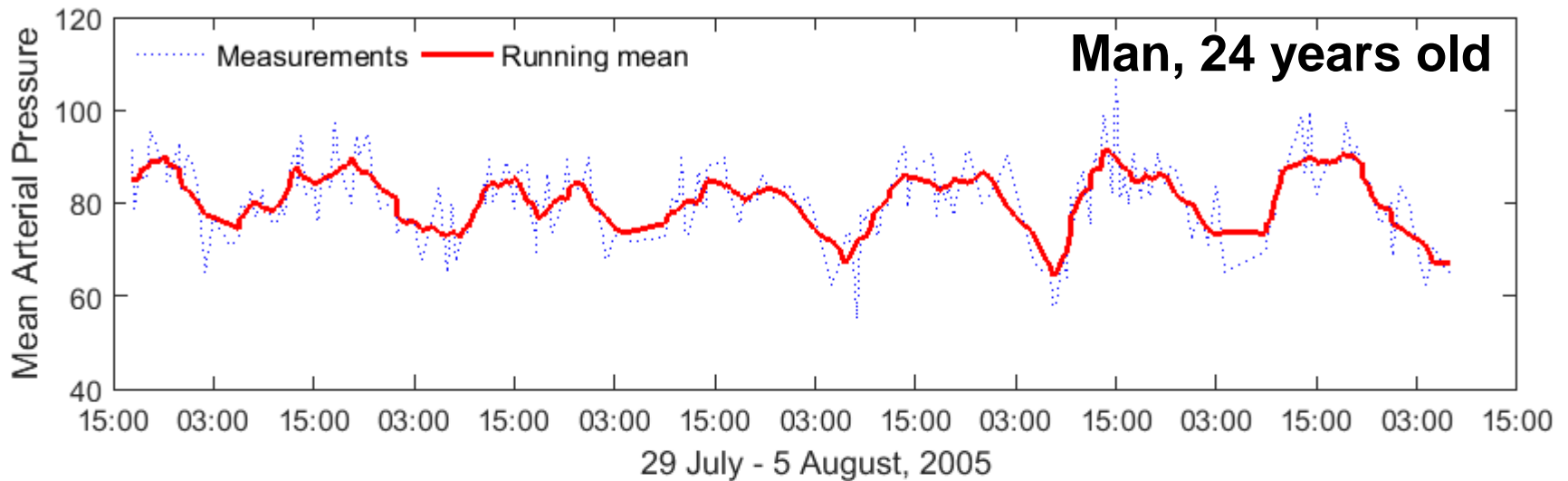
Estimation \hat{X}_i

$$\hat{X}_i = \frac{1}{9} \sum_{k=i-4}^{i+4} z_i$$

$$\hat{X}_i = \frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i+\frac{M-1}{2}} z_i$$

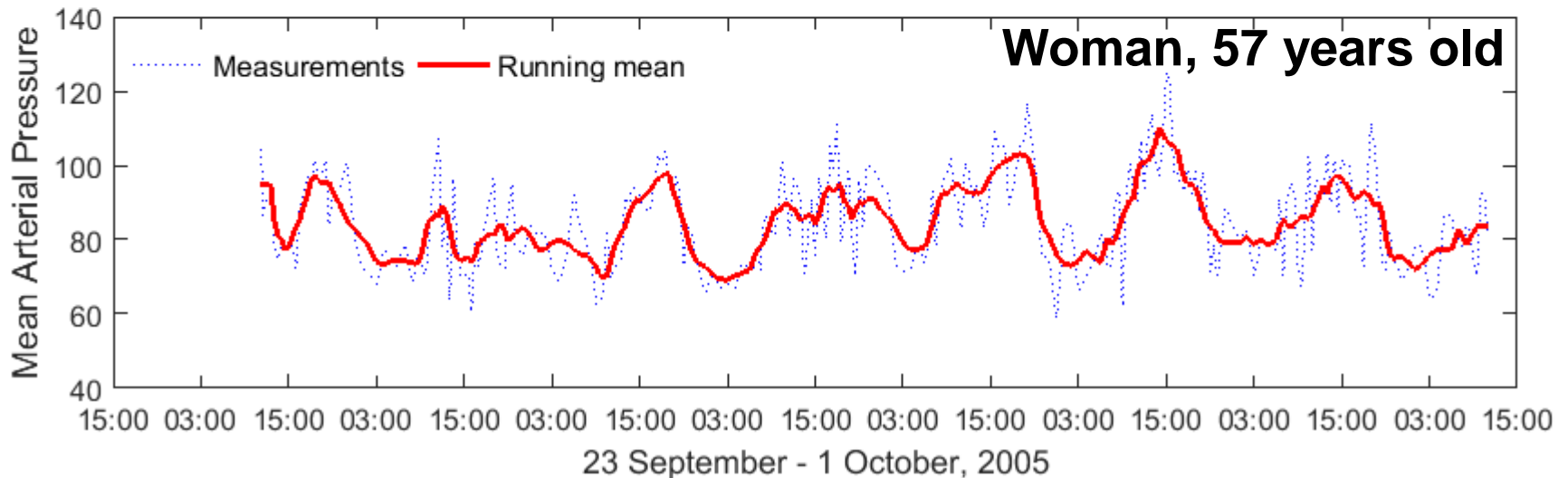
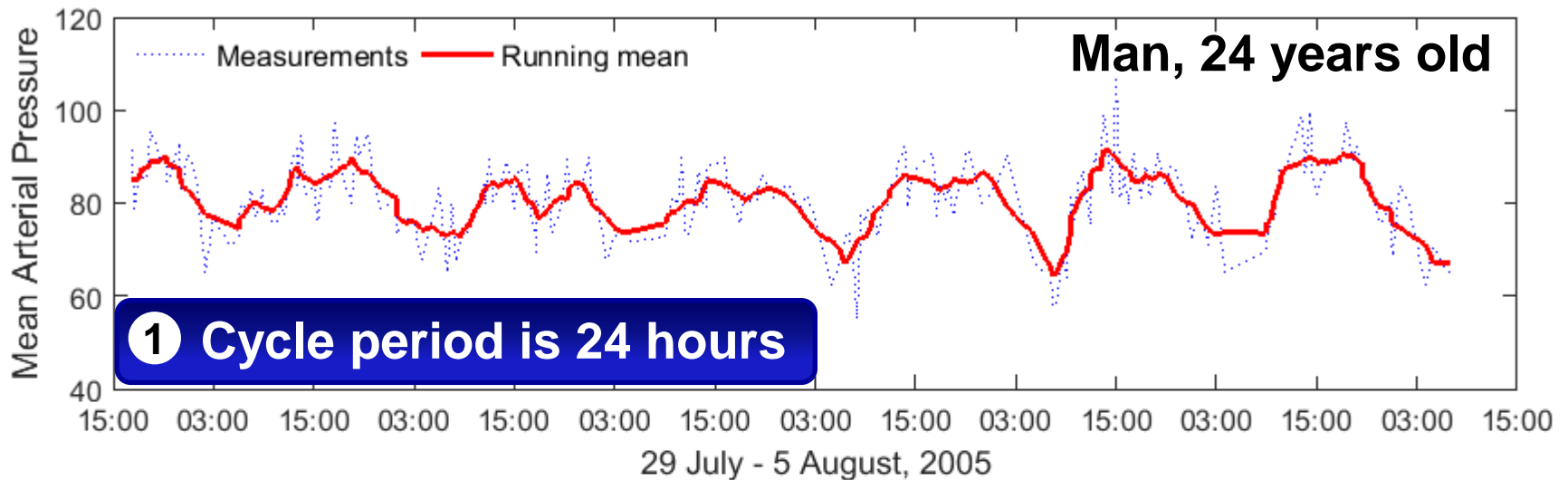
M - window size

Which regularities can you extract from smoothed measurements of mean arterial pressure?



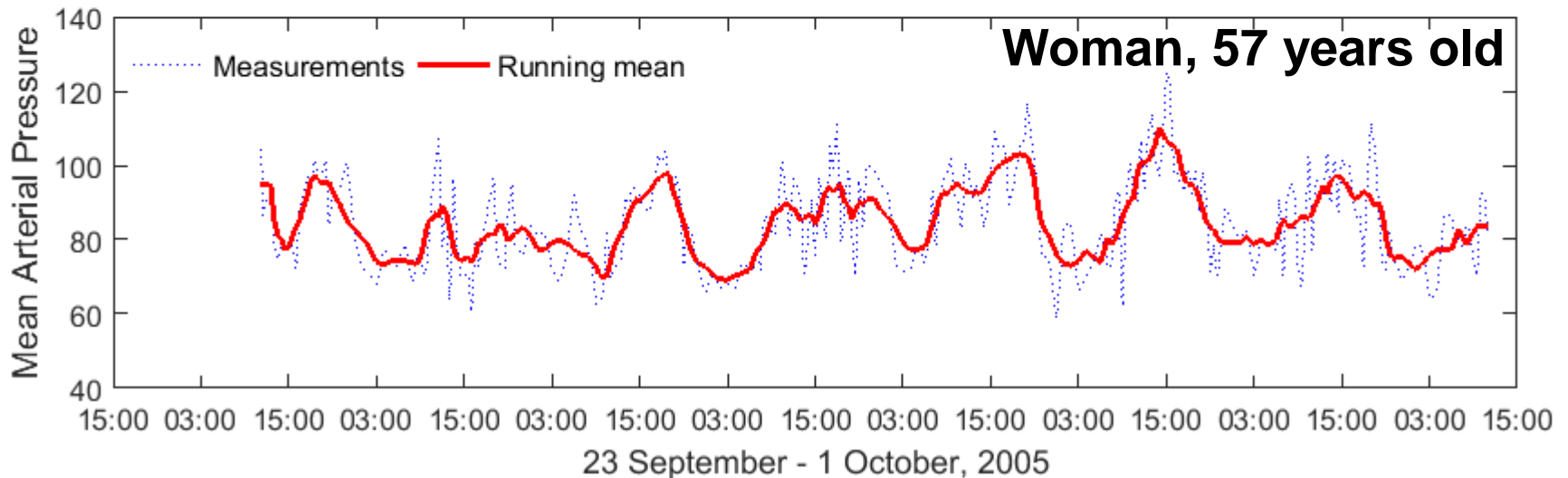
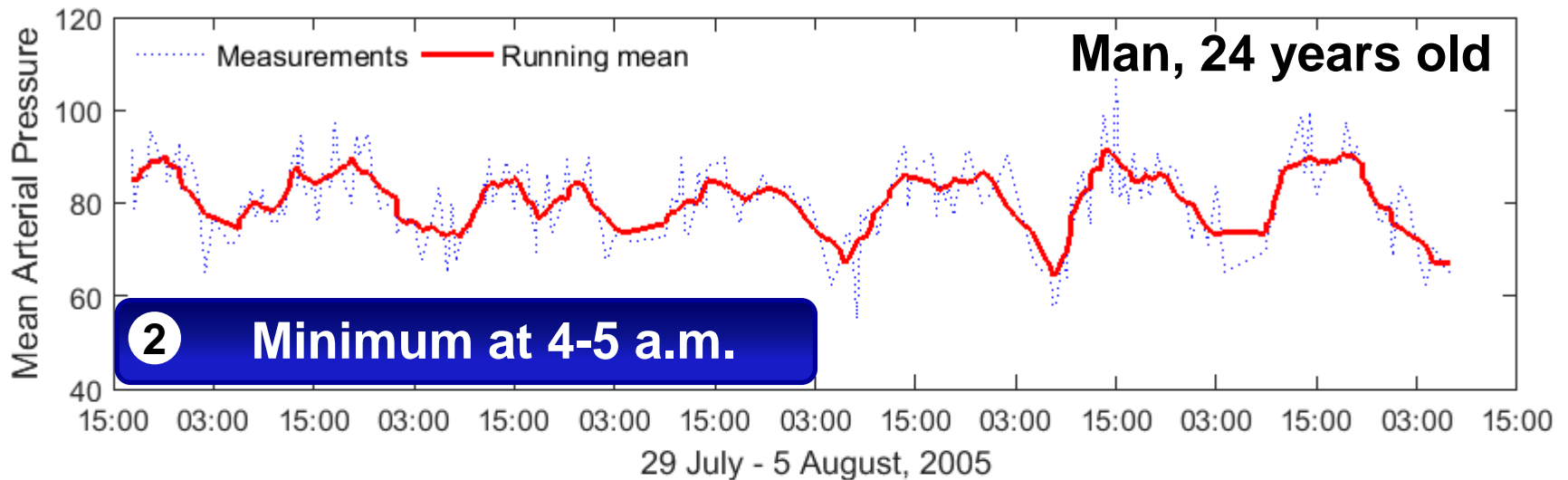
Running mean with window $M = 7$

Which regularities can you extract from smoothed measurements of mean arterial pressure?



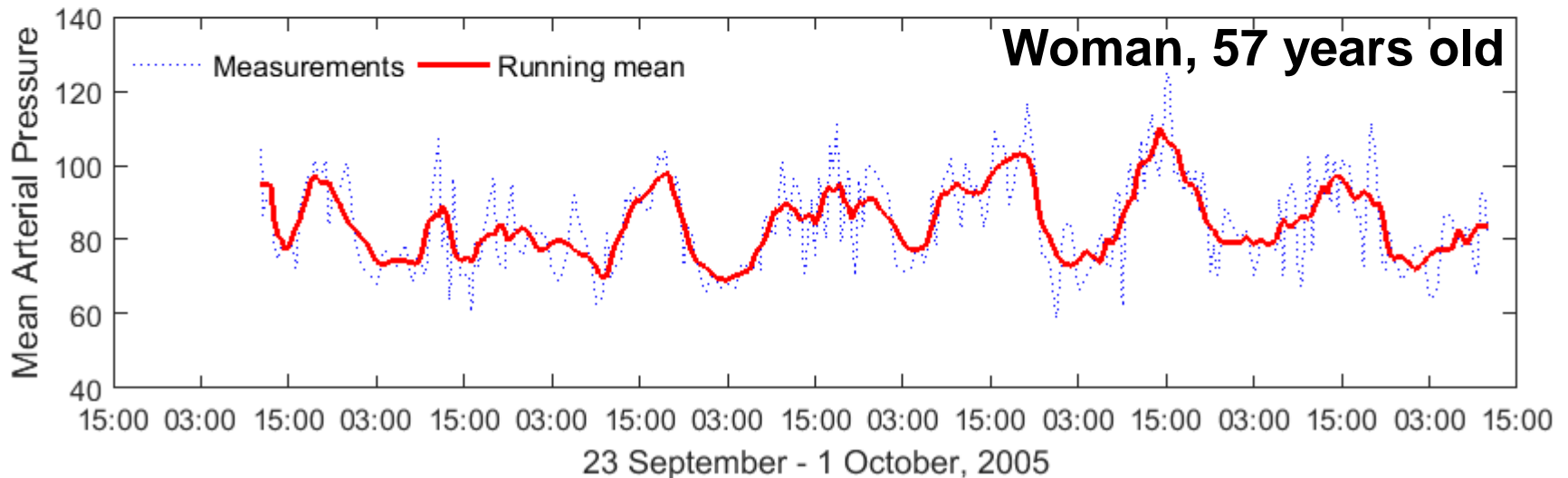
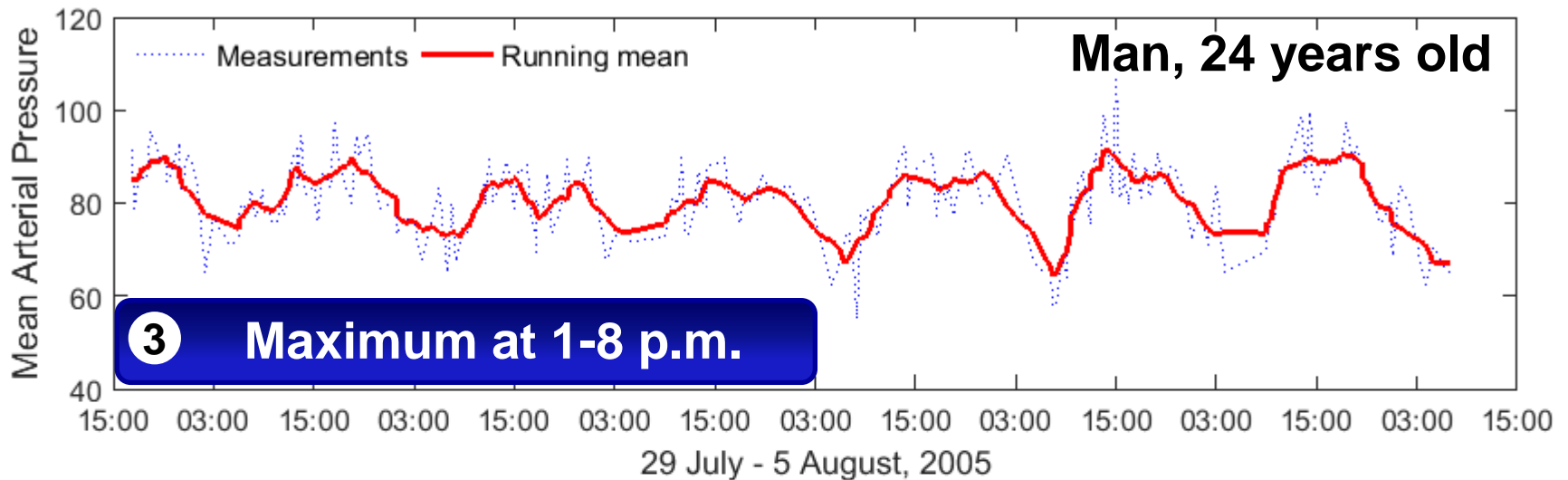
Running mean with window $M = 7$

Which regularities can you extract from smoothed measurements of mean arterial pressure?



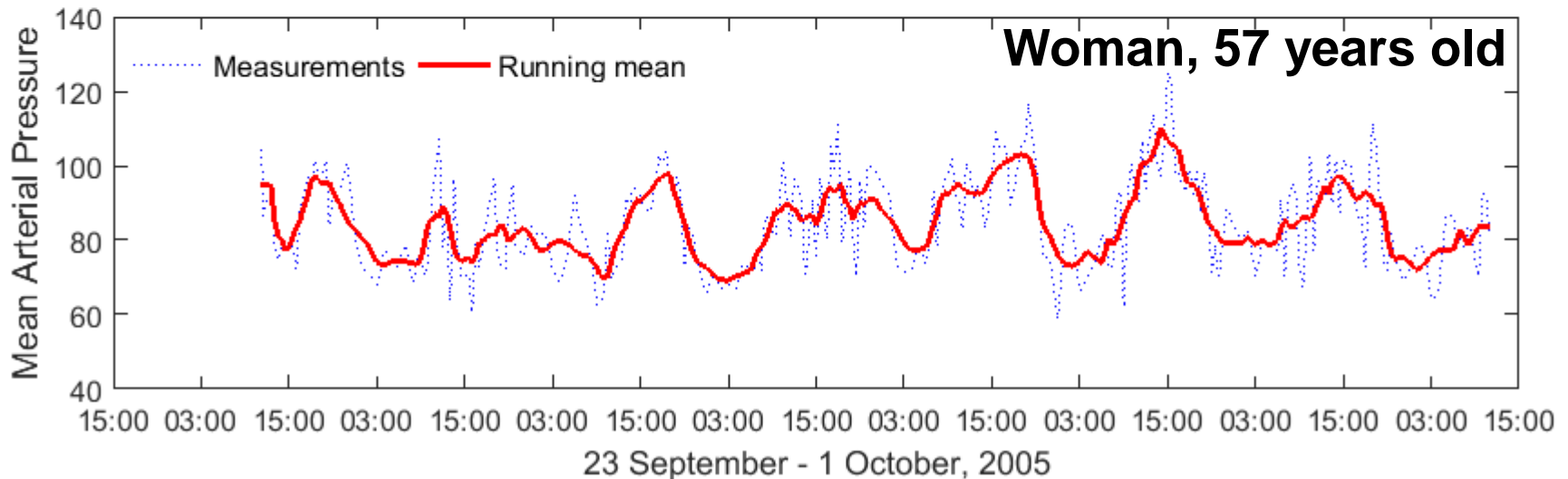
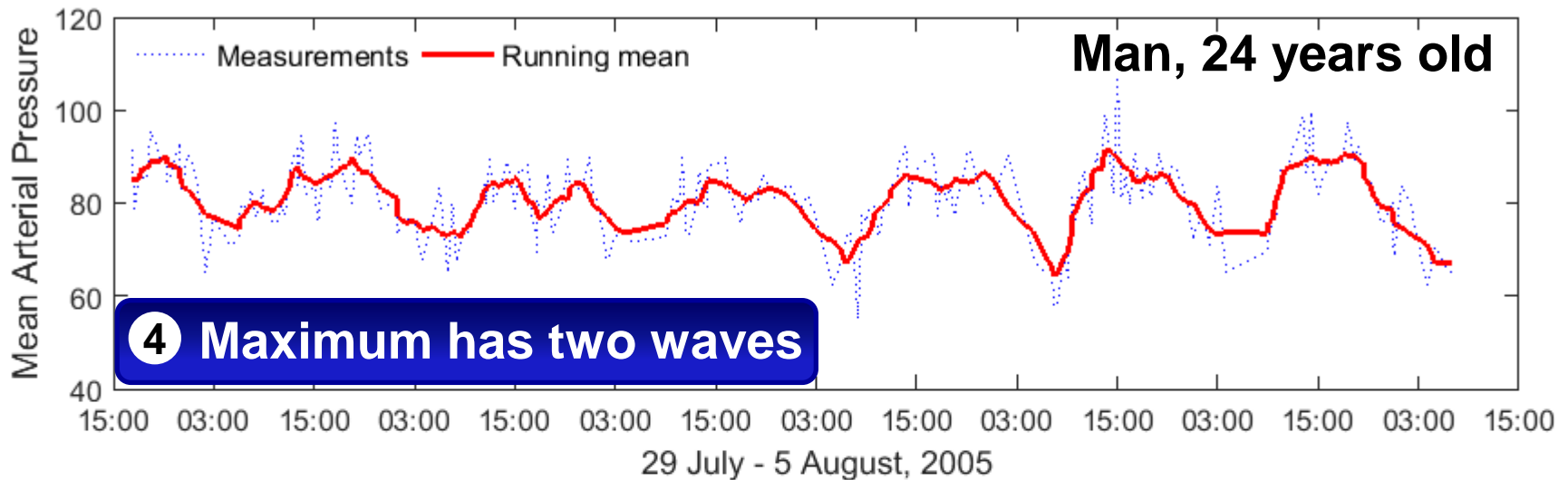
Running mean with window $M = 7$

Which regularities can you extract from smoothed measurements of mean arterial pressure?



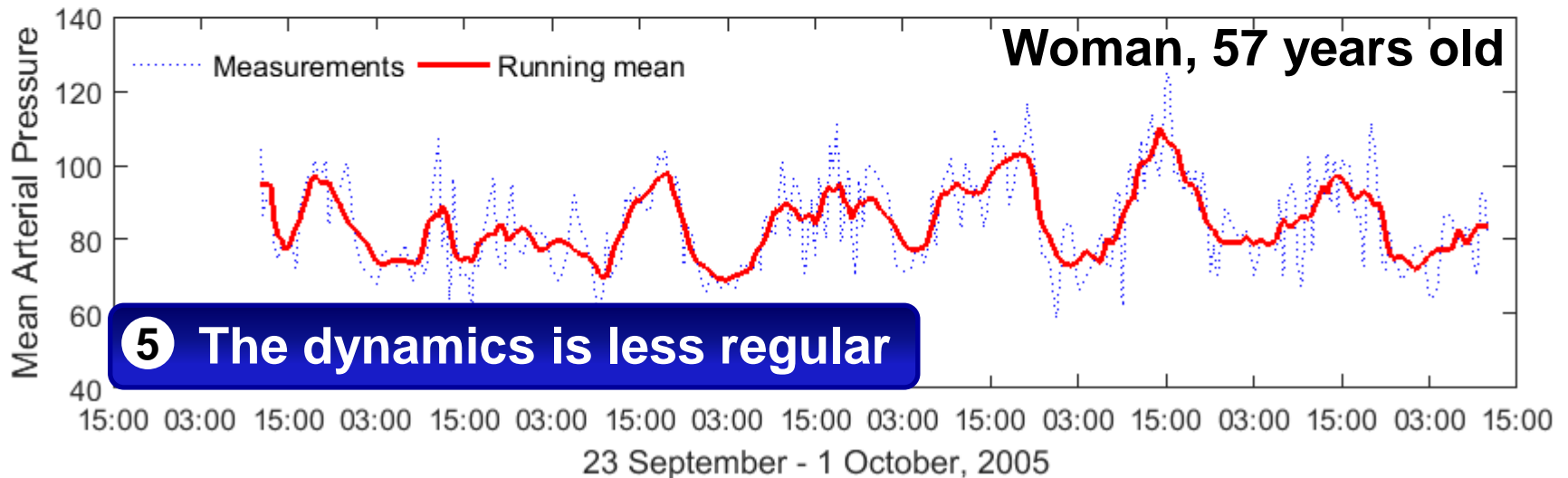
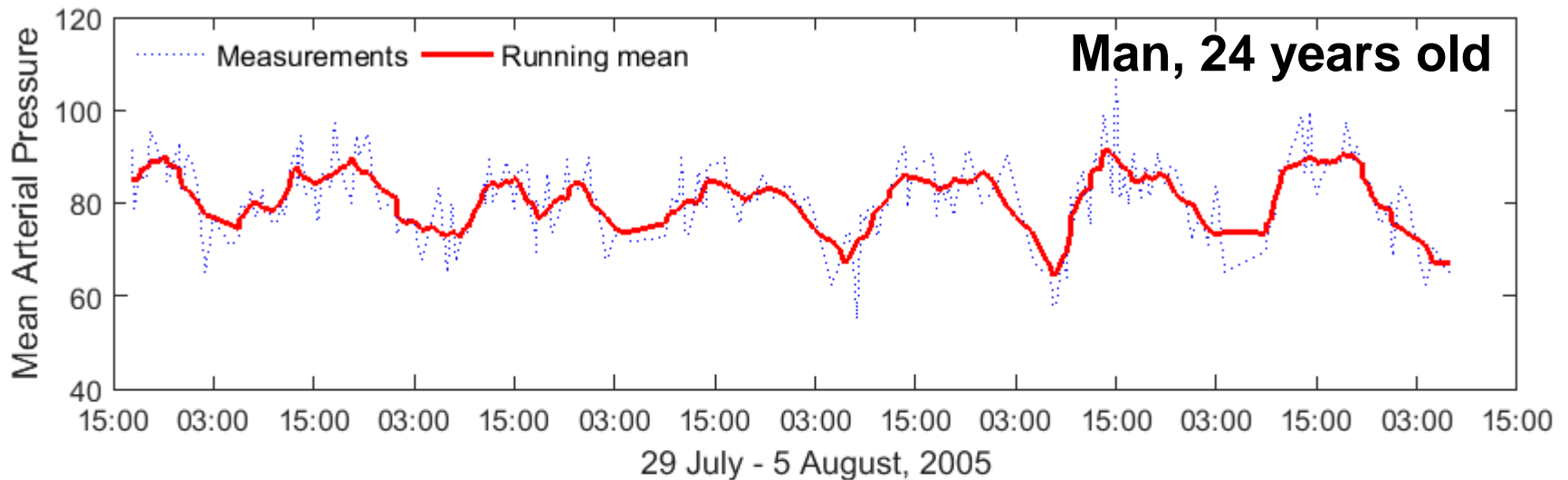
Running mean with window $M = 7$

Which regularities can you extract from smoothed measurements of mean arterial pressure?



Running mean with window $M = 7$

Which regularities can you extract from smoothed measurements of mean arterial pressure?



Running mean with window $M = 7$

① Running mean

```
graph TD; A[① Running mean] --> B((Dilemma of setting goal)); B --> C[Greater size of running window]; B --> D[Smaller size of running window]; C --> E[?]; D --> F[?];
```

The diagram is a flowchart on a black background. At the top is a white rounded rectangle containing the text '① Running mean'. A white arrow points down from this box to a red circle in the center containing the text 'Dilemma of setting goal'. From the red circle, two white arrows point outwards to two blue rounded rectangles. The left blue rectangle contains the text 'Greater size of running window' and the right one contains 'Smaller size of running window'. From each of these blue rectangles, a white arrow points down to a yellow rounded rectangle. Each yellow rectangle contains a large black question mark '?'.

**Dilemma
of setting
goal**

**Greater size of
running window**

**Smaller size of
running window**

?

?

① Running mean

```
graph TD; A[① Running mean] --> B((Dilemma of setting goal)); B --> C[Greater size of running window]; B --> D[Smaller size of running window]; C --> E[GREAT risk of process distortion, but EFFECTIVE filtration of measurement errors]; D --> F[SMALL risk of process distortion, but LESS EFFECTIVE filtration of measurement errors];
```

**Dilemma
of setting
goal**

**Greater size of
running window**

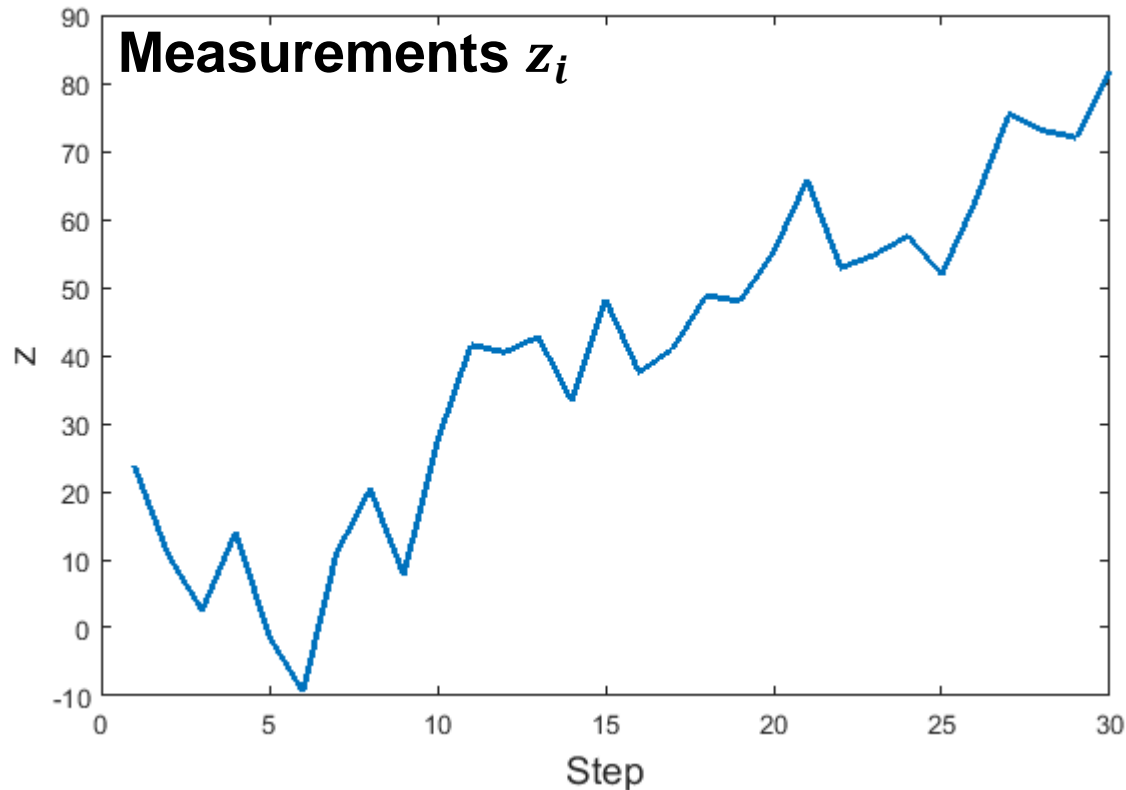
**GREAT risk of process
distortion, but
EFFECTIVE filtration
of measurement errors**

**Smaller size of
running window**

**SMALL risk of process
distortion, but LESS
EFFECTIVE filtration
of measurement errors**

② Exponential smoothing

Let's assume that values of process X are characterized by sudden change



2 Exponential smoothing

Let's assume that values of process X are characterized by sudden change



② Exponential smoothing

$$\hat{X}_i = \alpha z_i + (1 - \alpha)\hat{X}_{i-1}$$

$$\hat{X}_i$$

Smoothed
estimate
at time i

$$\alpha$$

Smoothing
constant
 $\alpha \in (0; 1)$

$$z_i$$

Measurements
at time i

$$\hat{X}_{i-1}$$

Smoothed
estimate
at time $i - 1$

② Exponential smoothing

$$\hat{X}_i = \alpha z_i + (1 - \alpha)\hat{X}_{i-1}$$

$$\hat{X}_i$$

Smoothed
estimate
at time i

$$\alpha$$

Smoothing
constant
 $\alpha \in (0; 1)$

$$z_i$$

Measurements
at time i

$$\hat{X}_{i-1}$$

Smoothed
estimate
at time $i - 1$

$$\hat{X}_i = \alpha z_i + \alpha(1 - \alpha)z_{i-1} + \alpha(1 - \alpha)^2 z_{i-2} + \cdots + \alpha(1 - \alpha)^i z_0$$

The weight of measurements
decreases according to geometric
progression or exponential law

② Exponential smoothing: Dilemma of setting goal

$$\hat{X}_i = \hat{X}_{i-1} + \alpha(z_i - \hat{X}_{i-1})$$

The diagram illustrates the exponential smoothing equation. A central blue box contains the equation $\hat{X}_i = \hat{X}_{i-1} + \alpha(z_i - \hat{X}_{i-1})$. Two white arrows point from this equation to two separate grey boxes below. The left arrow points to a box containing \hat{X}_{i-1} and the text 'Previous estimate'. The right arrow points to a box containing $(z_i - \hat{X}_{i-1})$ and the text 'Residual – mismatch between measurement and previous estimate'.

$$\hat{X}_{i-1}$$

Previous estimate

$$(z_i - \hat{X}_{i-1})$$

**Residual – mismatch between
measurement and previous estimate**

② Exponential smoothing: Dilemma of setting goal

$$\hat{X}_i = \hat{X}_{i-1} + \alpha(z_i - \hat{X}_{i-1})$$

$$\hat{X}_{i-1}$$

Previous estimate

$$(z_i - \hat{X}_{i-1})$$

Residual – mismatch between
measurement and previous estimate

**SMALLER α , GREATER
confidence to the latest
estimate, SLOWER
reaction to changes**

**But EFFECTIVE
filtration of
measurement errors**

**Choice
of α**

**GREATER α , GREATER
confidence to the latest
measurement, FASTER
reaction to changes**

**But less EFFECTIVE
filtration of
measurement errors**

Comparison of smoothing methods

1 Running mean

$$\hat{X}_i = \frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i+\frac{M-1}{2}} z_k$$

**Last M
measurements
are used**

2 Exponential mean

$$\hat{X}_i = \hat{X}_{i-1} + \alpha(z_i - \hat{X}_{i-1})$$

**All previous
measurements
are used**

Comparison of smoothing methods

1 Running mean

$$\hat{X}_i = \frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i+\frac{M-1}{2}} z_k$$

Equal weights of
measurements

2 Exponential mean

$$\hat{X}_i = \hat{X}_{i-1} + \alpha(z_i - \hat{X}_{i-1})$$

The weight of measurements
decreases according
to exponential law

Comparison of smoothing methods

1 Running mean

$$\hat{X}_i = \frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i+\frac{M-1}{2}} z_k$$

Delay of estimation
on $\frac{M-1}{2}$ steps

2 Exponential mean

$$\hat{X}_i = \hat{X}_{i-1} + \alpha(z_i - \hat{X}_{i-1})$$

Estimation is
obtained at last
available time moment

3

Comparison of smoothing methods

1 Running mean

$$\hat{X}_i = \frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i+\frac{M-1}{2}} z_k$$

$$M = 1$$

No filtration
of errors

2 Exponential mean

$$\hat{X}_i = \hat{X}_{i-1} + \alpha(z_i - \hat{X}_{i-1})$$

$$\alpha = 1$$

4

Estimates of both
smoothing methods
are the same

Comparison of smoothing methods

1 Running mean

$$\hat{X}_i = \frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i+\frac{M-1}{2}} z_k$$

$M \rightarrow \infty$

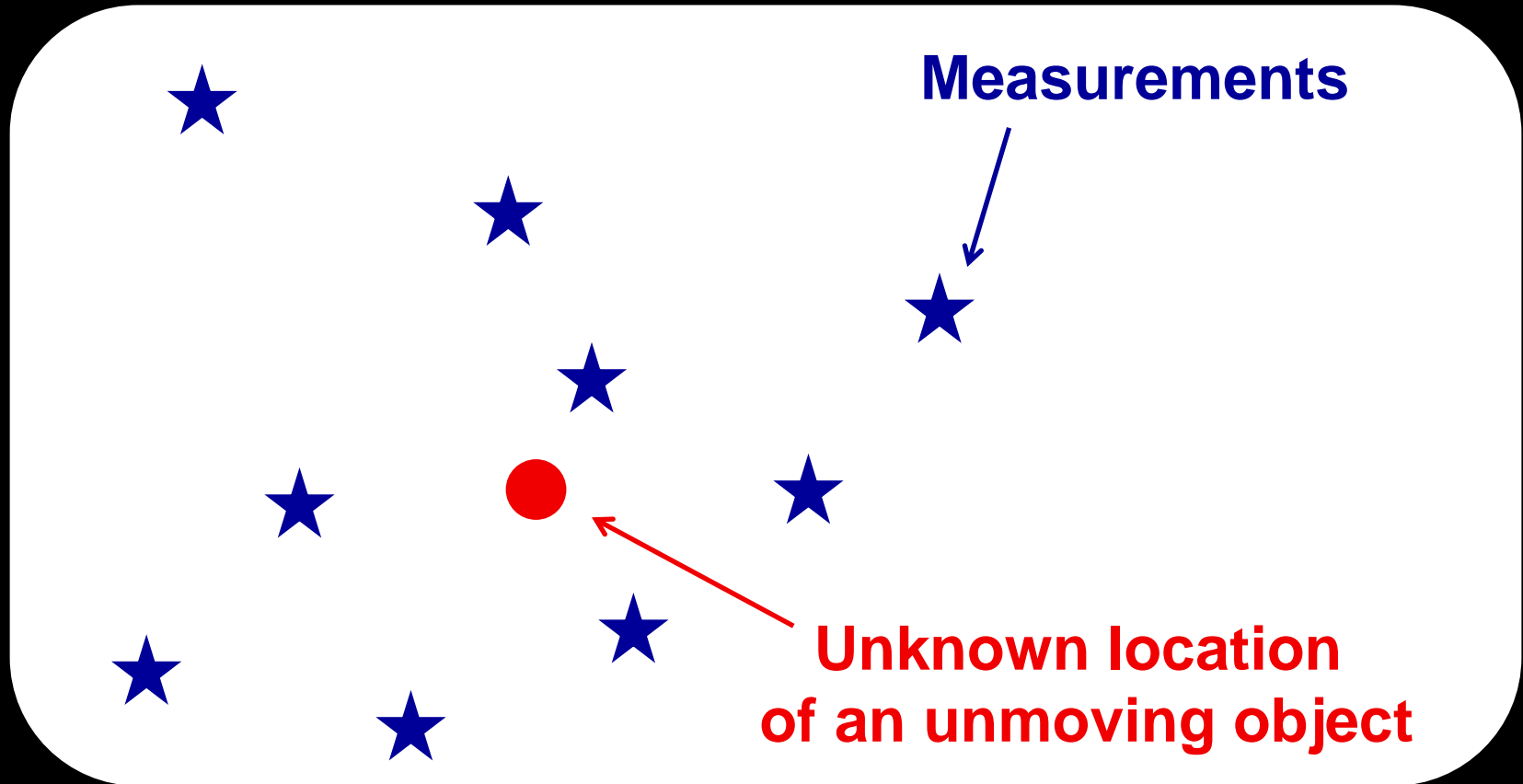
Effective filtration
of errors. But no
reaction to changes
in dynamics

2 Exponential mean

$$\hat{X}_i = \hat{X}_{i-1} + \alpha(z_i - \hat{X}_{i-1})$$

$\alpha \rightarrow 0$

Sources of estimation errors

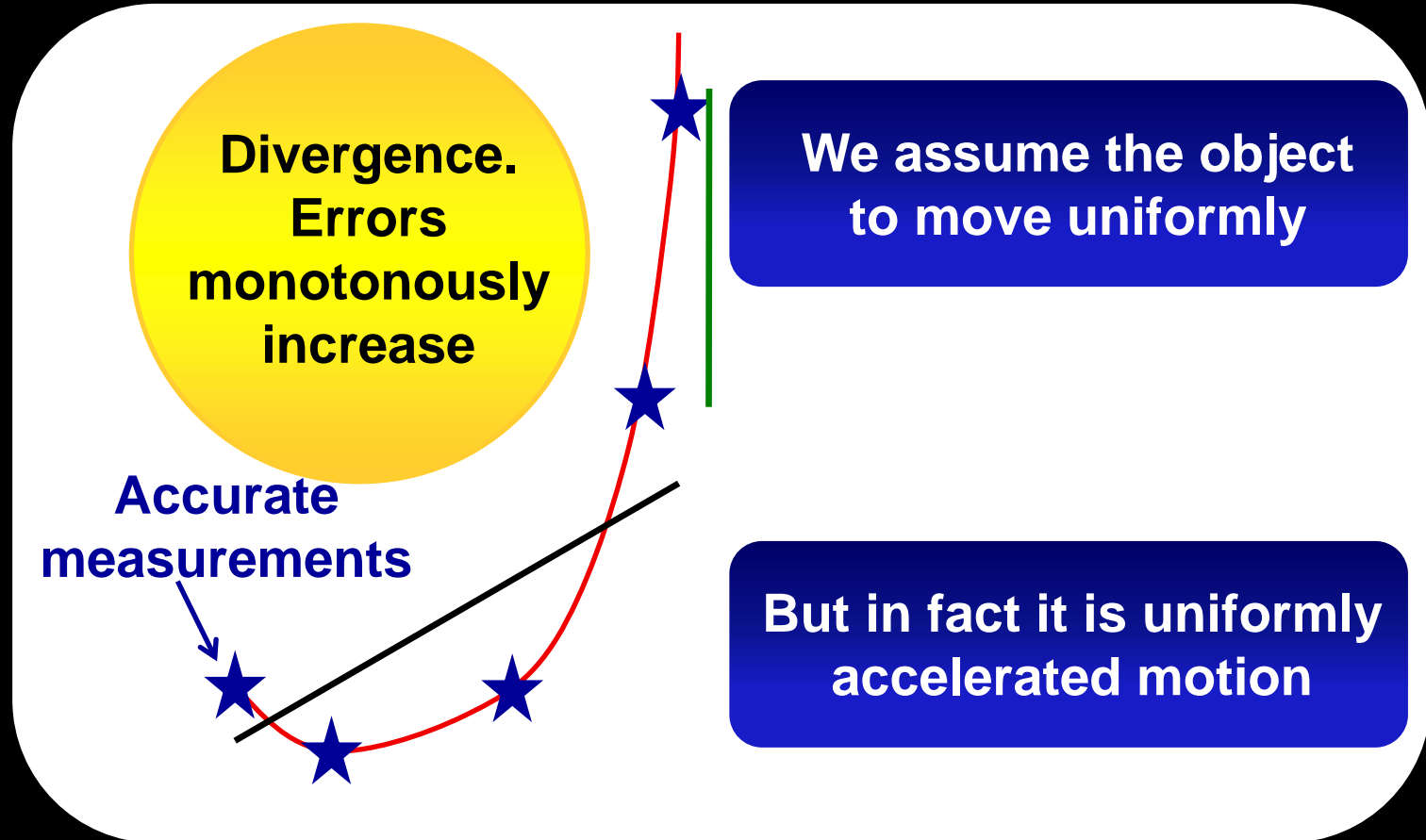


**Source 1:
Measurement
errors**



**Errors of estimation are related
with only measurements errors.
Model of motion is accurate**

Sources of estimation errors



**Source 2:
Methodical
errors**



**Errors of estimation are
related with errors of methods.
Model of motion is inaccurate.**

Source 1: Measurement errors

1 Running mean

$$\hat{X}_i = \frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i+\frac{M-1}{2}} z_k$$

$$\sigma_{\hat{X}}^2 = \frac{1}{M^2} \sum_{k=i-\frac{M-1}{2}}^{i+\frac{M-1}{2}} \sigma_{\eta}^2$$

$$\sigma_{\hat{X}}^2 = \frac{\sigma_{\eta}^2}{M}$$

2 Exponential mean

$$\hat{X}_i = \alpha \sum_{k=0}^{i-1} (1 - \alpha)^k z_{i-k} + (1 - \alpha)^i z_0$$

$$\lim_{i \rightarrow \infty} \sigma_{\hat{X}}^2 = \lim_{i \rightarrow \infty} \left(\alpha^2 \sigma_{\eta}^2 \sum_{k=0}^{i-1} (1 - \alpha)^{2k} \right)$$

$$\sigma_{\hat{X}}^2 = \sigma_{\eta}^2 \frac{\alpha}{2 - \alpha}$$

Source 1: Measurement errors

1 Running mean

$$\sigma_{\hat{X}}^2 = \frac{\sigma_{\eta}^2}{M}$$

$$M = 1$$

No filtration
of errors

$$\alpha = 1$$

2 Exponential mean

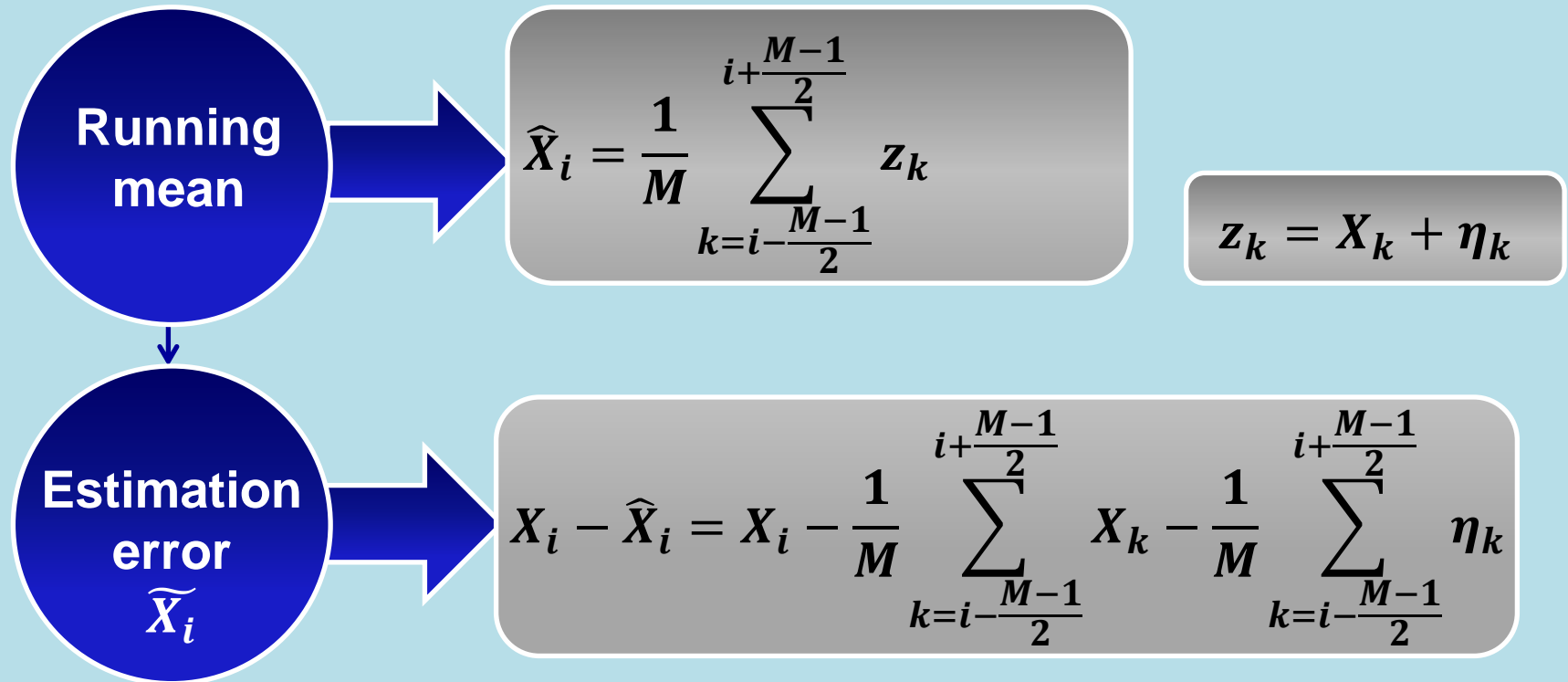
$$\sigma_{\hat{X}}^2 = \sigma_{\eta}^2 \frac{\alpha}{2 - \alpha}$$

$$M \rightarrow \infty$$

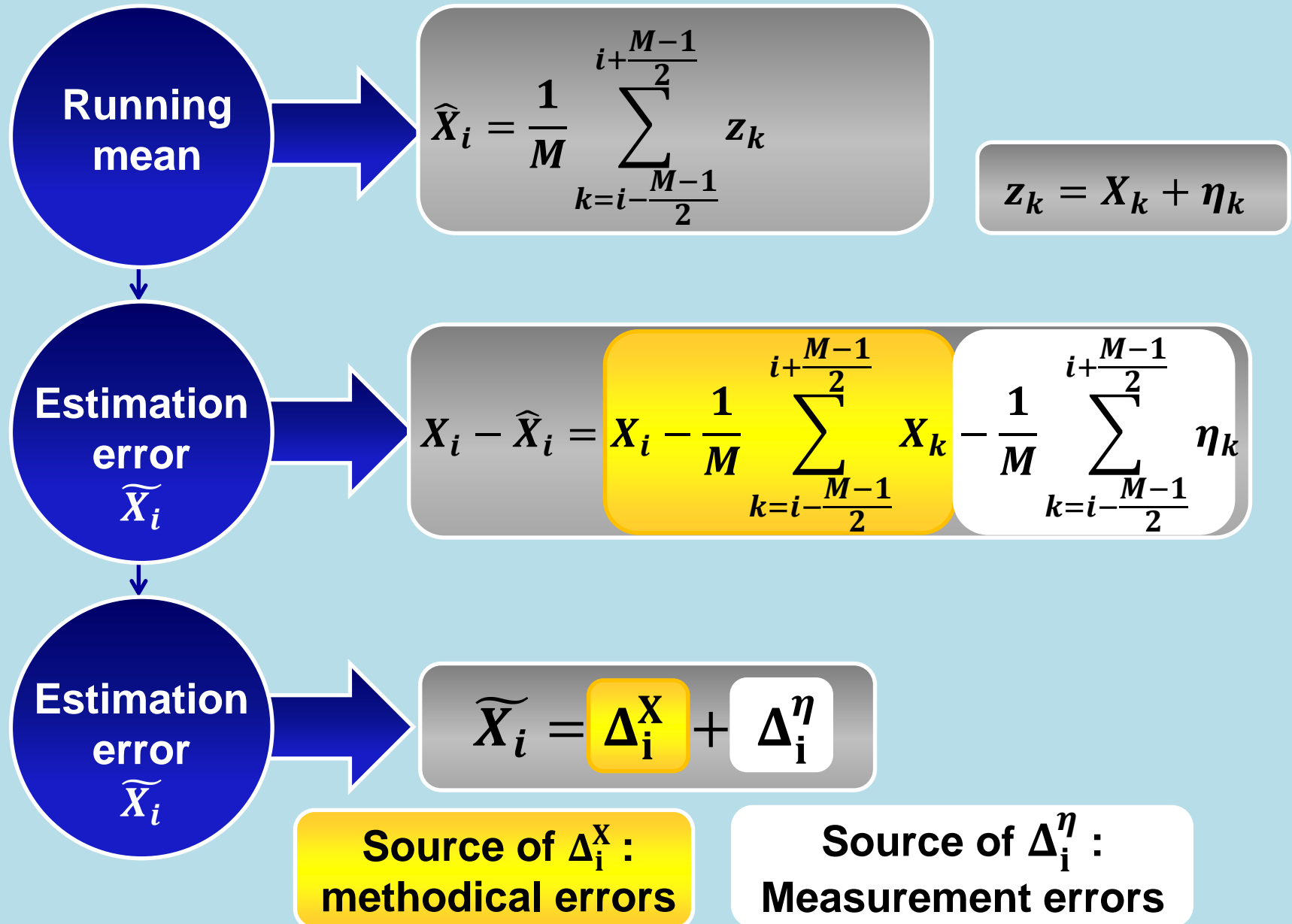
Effective filtration
of errors. But no
reaction to changes
in dynamics

$$\alpha \rightarrow 0$$

Source 2: Methodical errors of running mean



Source 2: Methodical errors of running mean



Source 2: Methodical errors of running mean

$$9 = \sum_{k=i-4}^{i+4} 1$$

$$\Delta_i^X = X_i - \frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i+\frac{M-1}{2}} X_k$$

$$X_i = X_i \frac{1}{M} M = \frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i+\frac{M-1}{2}} X_i$$

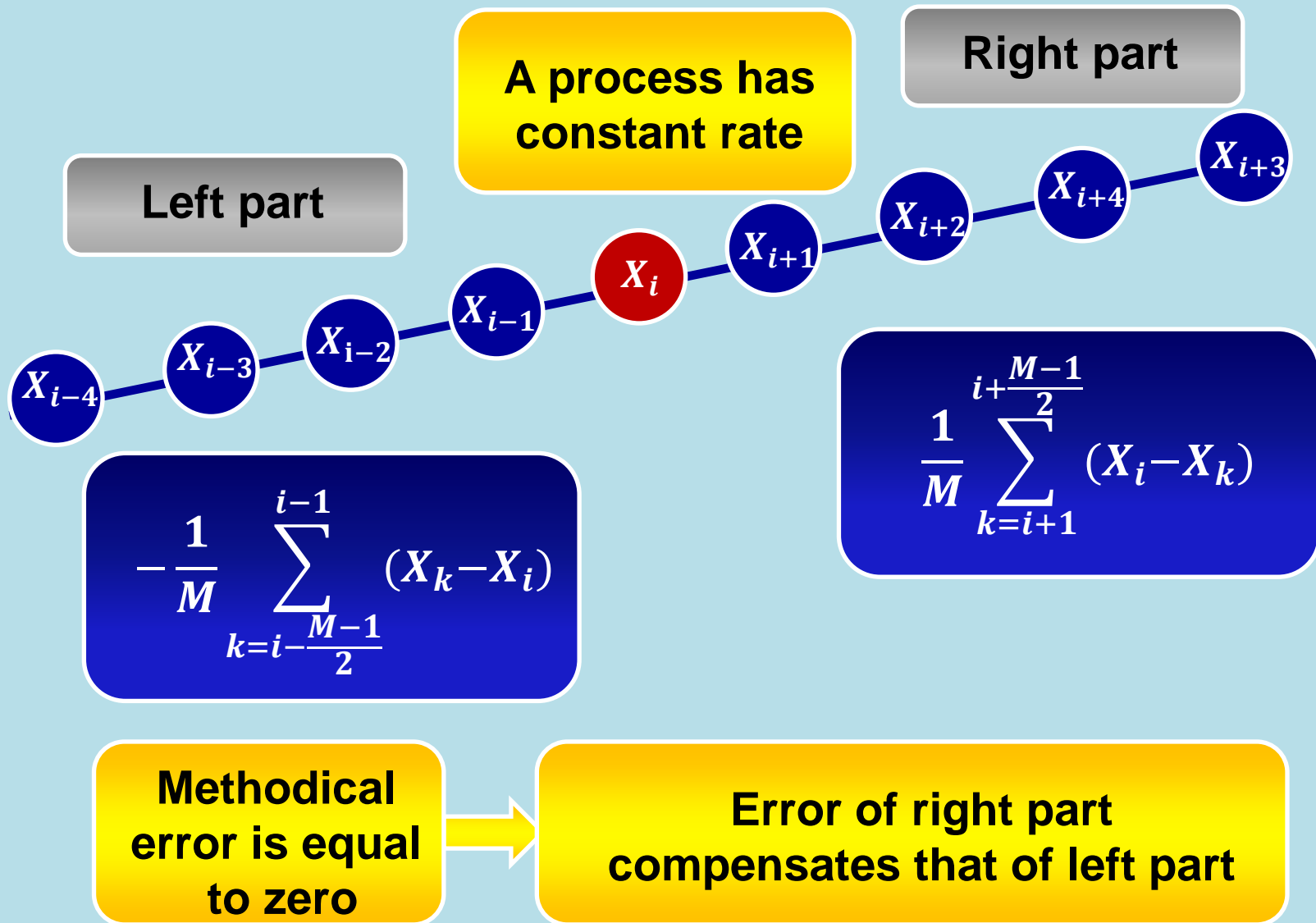
$$\Delta_i^X = \frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i+\frac{M-1}{2}} (X_i - X_k)$$

$$X_i - \hat{X}_i = -\frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i-1} (X_k - X_i) + \frac{1}{M} \sum_{k=i+1}^{i+\frac{M-1}{2}} (X_i - X_k)$$

$$k > i$$

$$k < i$$

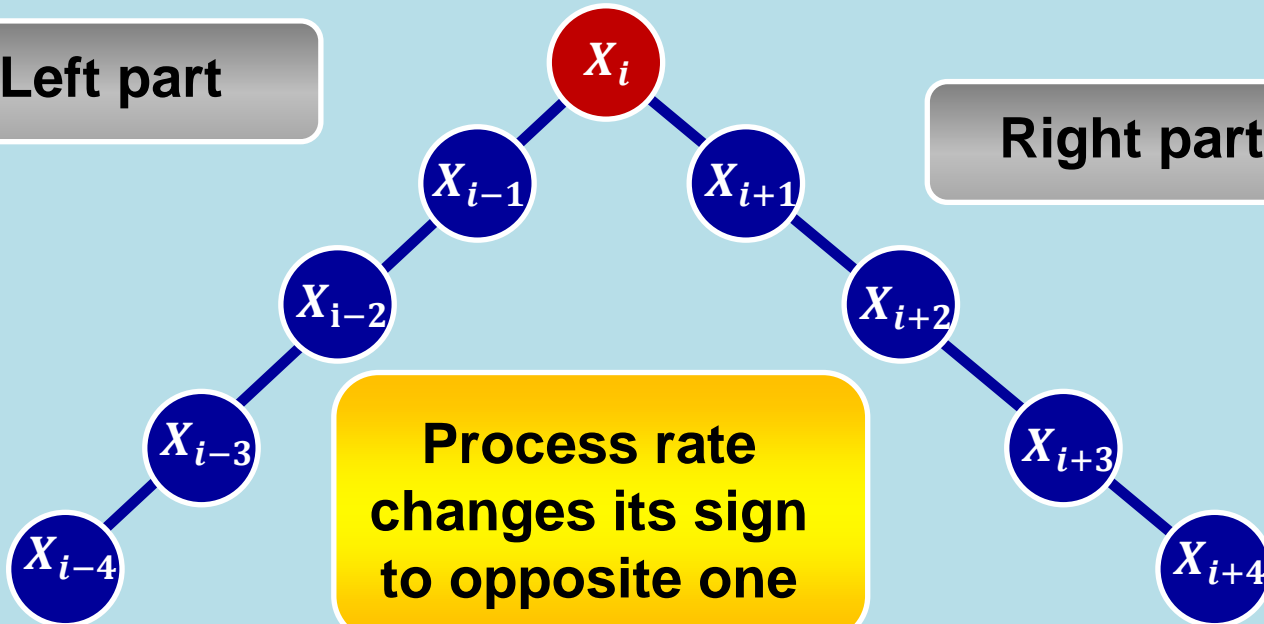
Source 2: Methodical errors of running mean



Source 2: Methodical errors of running mean

Left part

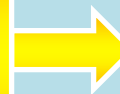
Right part



$$-\frac{1}{M} \sum_{k=i-\frac{M-1}{2}}^{i-1} (X_k - X_i)$$

$$\frac{1}{M} \sum_{k=i+1}^{i+\frac{M-1}{2}} (X_i - X_k)$$

Methodical error is doubled

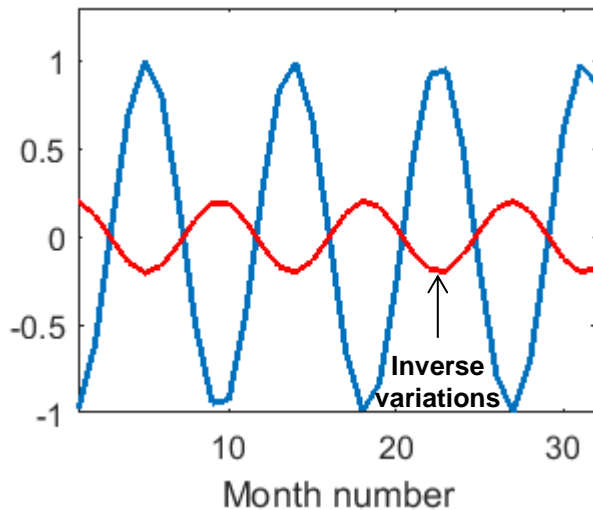


Error of right part doubles that of left part

Analysis of running mean errors

Running mean may significantly distort the dynamics of the process

9-months variation

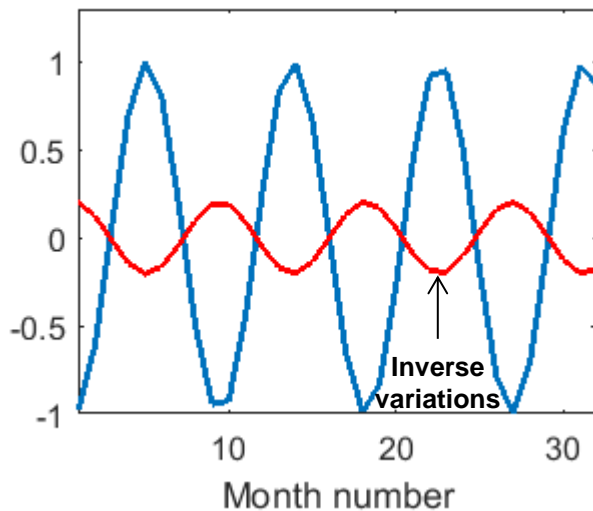


Inverse variations with periods from 6 to 12 months. Convex curve is replaced by concave curve and vice versa

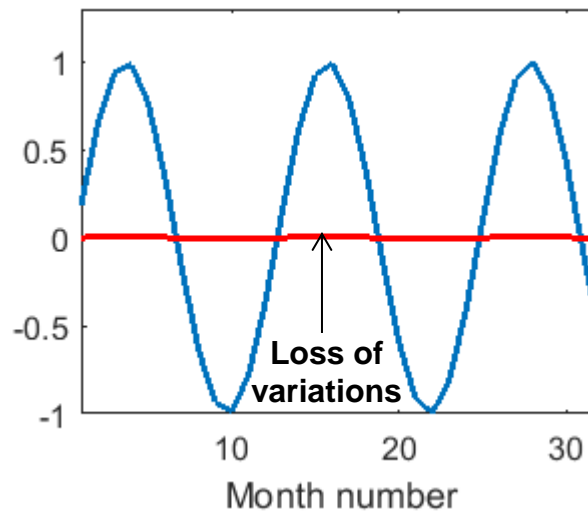
Analysis of running mean errors

Running mean may significantly distort the dynamics of the process

9-months variation



13-months variation



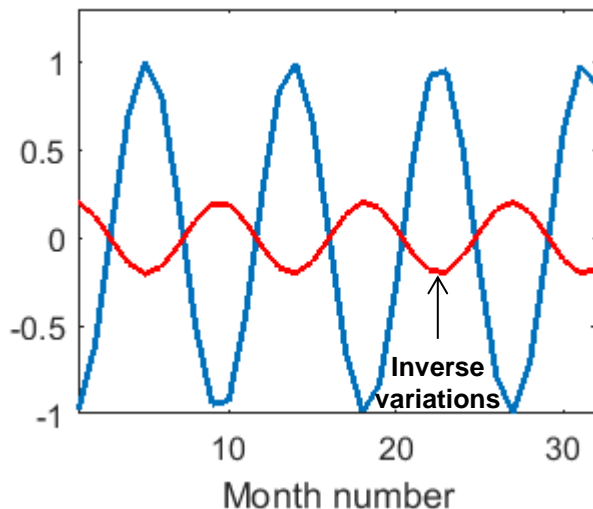
Inverse variations with periods from 6 to 12 months. Convex curve is replaced by concave curve and vice versa

Total loss of 6- and 12-month variations decreasing them to zero

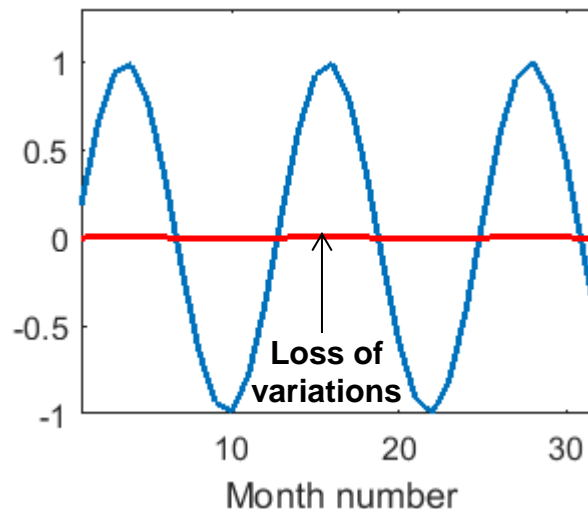
Analysis of running mean errors

Running mean may significantly distort the dynamics of the process

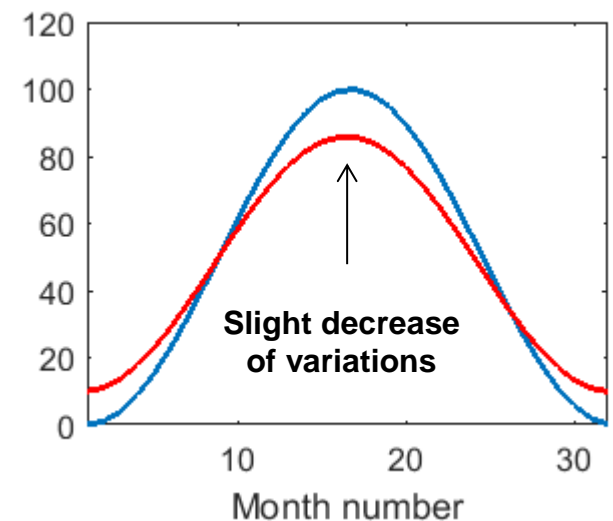
9-months variation



13-months variation



32-months variation

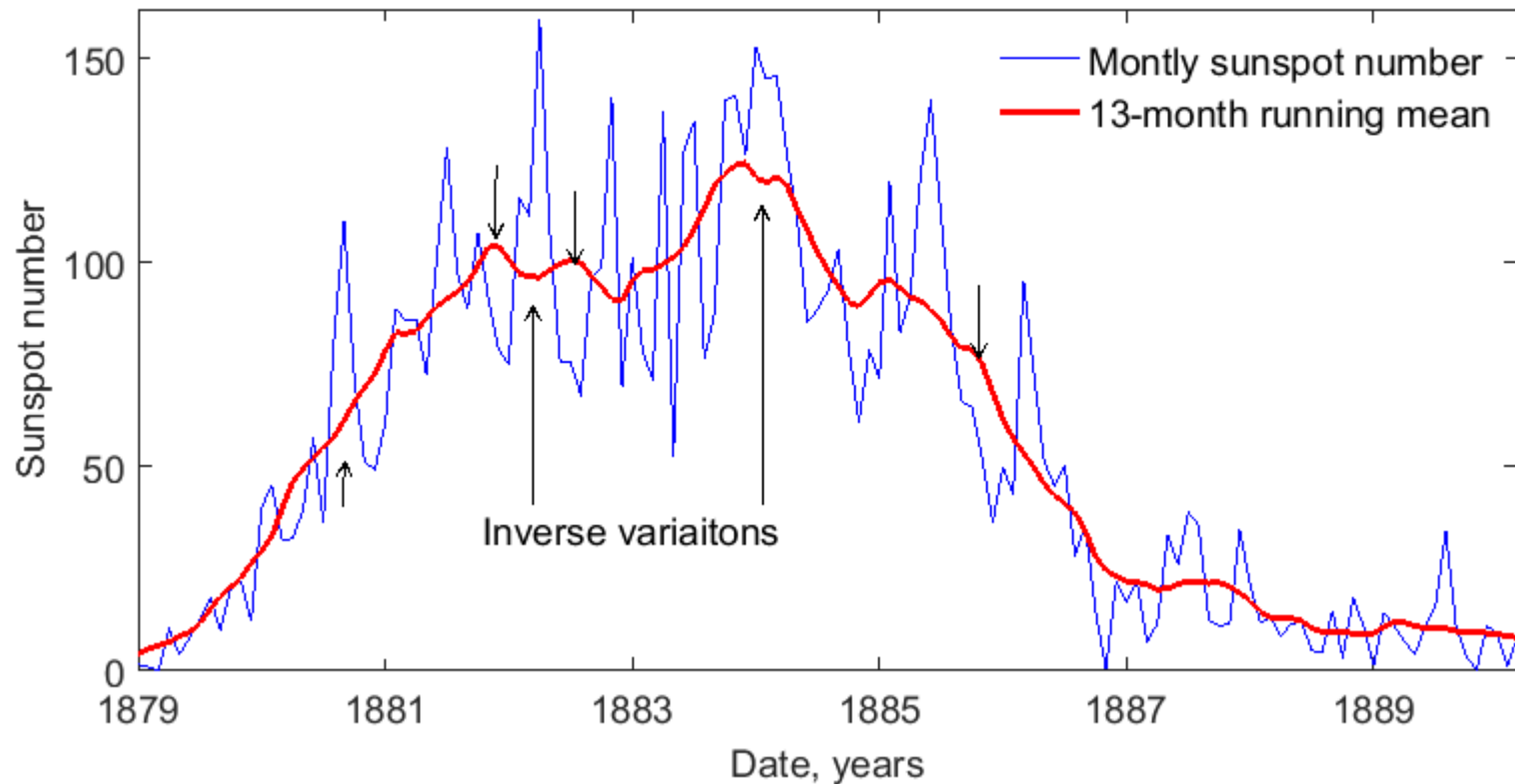


Inverse variations with periods from 6 to 12 months. Convex curve is replaced by concave curve and vice versa

Total loss of 6- and 12-month variations decreasing them to zero

Period greater than running window size (13 months). The process in general is not distorted

Distortion of physics in sunspot cycle 12



Performed analysis allows us to anticipate the errors of smoothing and getting false conclusions

Alternatives in the following topics of course

Methodical errors and modifications of exponential smoothing

Will be considered in laboratory works

Conclusions

Don't apply methods in blind to not fall into the trap leading to false conclusions

Even if implementation is simple, the method itself requires careful analysis