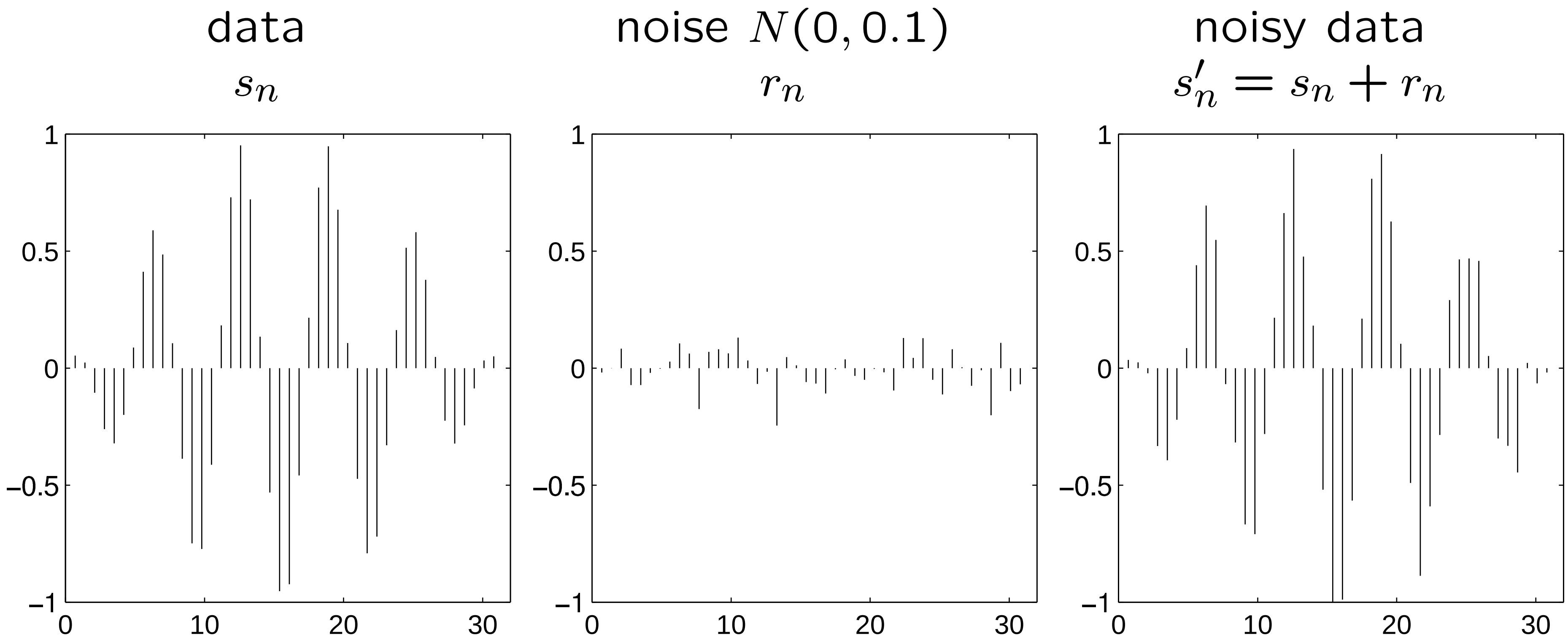
Chapter 3: Data Preprocessing

1. Error Types and Handling
2. Filtering
3. Standardization and Transformation
4. Data Merging

 Copyright © 2020. All rights reserved.

Random Errors

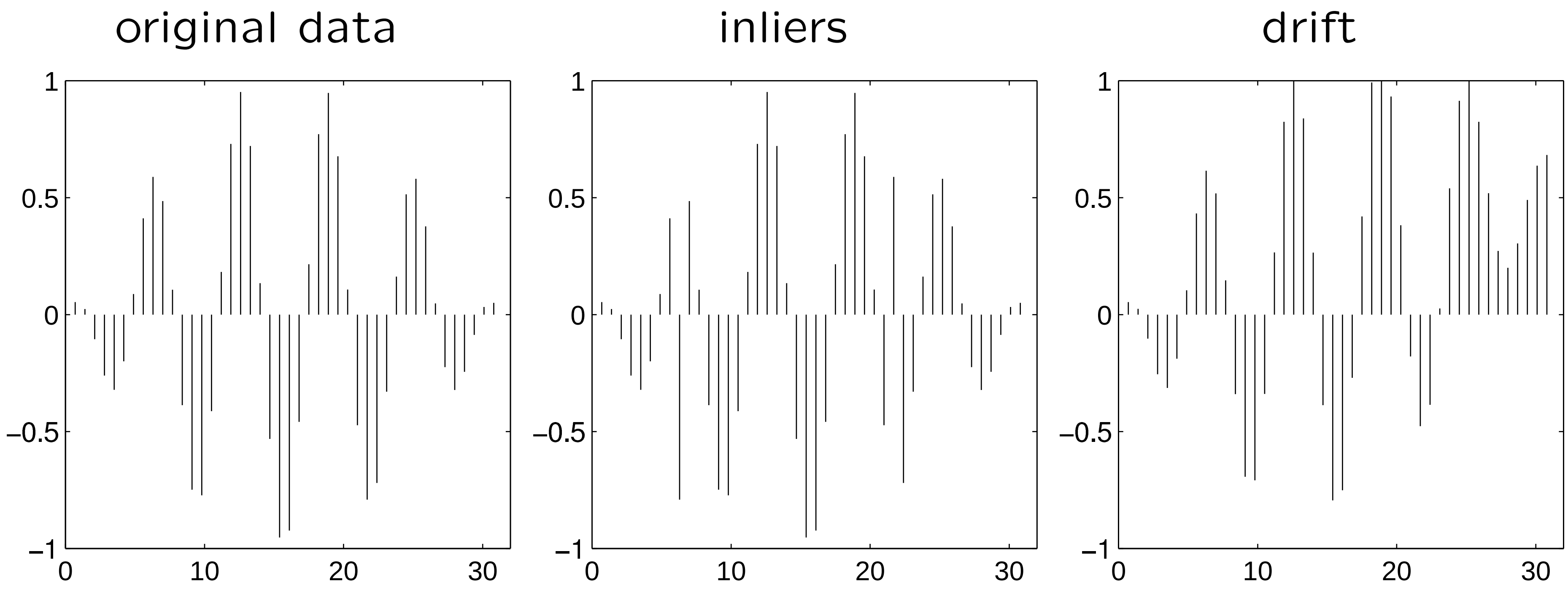
- measurement and transmission errors
- modeling as additive noise



 Copyright © 2020. All rights reserved.

Inliers, Outliers, and Drift

- single incorrect data: inliers/outliers
- processing errors: permuted or wrong data (e.g. 1.000/1,000)
- measurement errors: offset, scaling, drift



 Copyright © 2020. All rights reserved.

Outlier Detection

- comparison with range limits

$$\left(x_k^{(i)} < x_{\text{min}}^{(i)}\right) \vee \left(x_k^{(i)} > x_{\text{max}}^{(i)}\right)$$

- range limits $x_{\text{min}}^{(i)}$, $x_{\text{max}}^{(i)}$ e.g. given by:
  1. sign (price, temperature, time)
  2. sensor range, considered time interval
  3. defined or physically plausible values
- 2–sigma rule

$$\left|\frac{x_k^{(i)} - \bar{x}^{(i)}}{s_x^{(i)}}\right| > 2$$

- problem:
  unusual but valuable data can not be distinguished from incorrect data (real outliers)

 Copyright © 2020. All rights reserved.

Error Handling

- invalidity list
- invalidity value

$$x_k^{(i)} = \textit{NaN (not a number)}$$

implementation in 64 bit IEEE floating point format

`NaN=$7FFFFFFF`

- replace by mean, median, minimum, or maximum of the valid feature data x^i
- replace by nearest neighbor $x_k^{(i)} = x_j^{(i)}$

$$\|x_j - x_k\|_{\neg i} = \min_{l \in \{1,\ldots,n\}} \|x_l - x_k\|_{\neg i}$$

where $\|.\|_{\neg i}$ ignores feature i and invalid or missing data

Copyright © 2020. All rights reserved.

Error Handling

- linear interpolation for equidistant time series

$$x_k^{(i)} = \frac{x_{k-1}^{(i)} + x_{k+1}^{(i)}}{2}$$

- linear interpolation for non–equidistant time series

$$x_k^{(i)} = \frac{x_{k-1}^{(i)} \cdot (t_{k+1} - t_k) + x_{k+1}^{(i)} \cdot (t_k - t_{k-1})}{t_{k+1} - t_{k-1}}$$

- nonlinear interpolation, e.g. splines
- model-based estimation by regression
- filtering
- outlier removal: the complete vector x_k is removed
- feature removal: the complete feature $x^{(i)}$ is removed

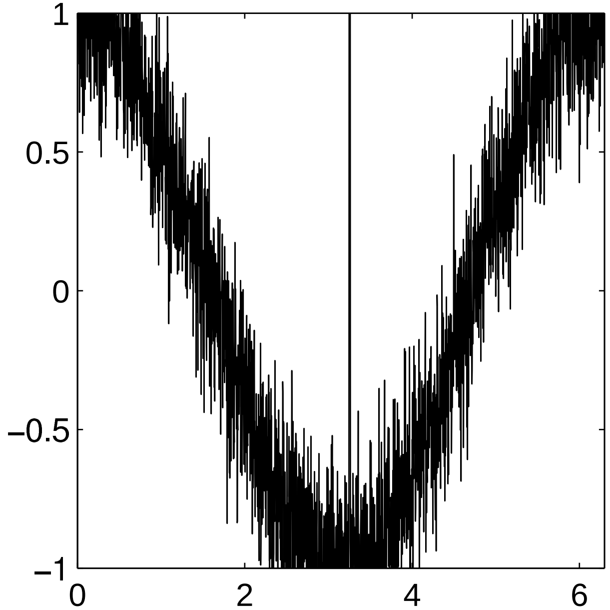
Prof. Dr. Thomas A. Runkler
Copyright © 2020. All rights reserved.

Moving Average

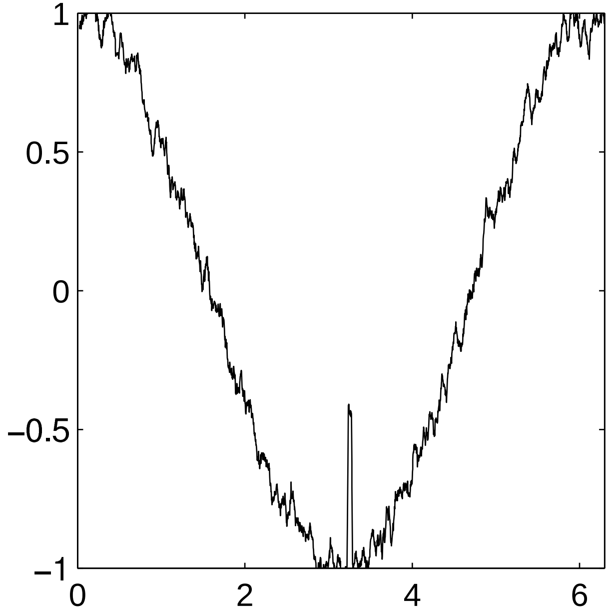
- moving average of (even) order $q \in \{2, 4, 6, \ldots\}$

$$y_k = \frac{1}{q+1} \sum_{i=k-\frac{q}{2}}^{k+\frac{q}{2}} x_i \qquad\qquad y_k = \frac{1}{q+1} \sum_{i=k-q}^{k} x_i$$

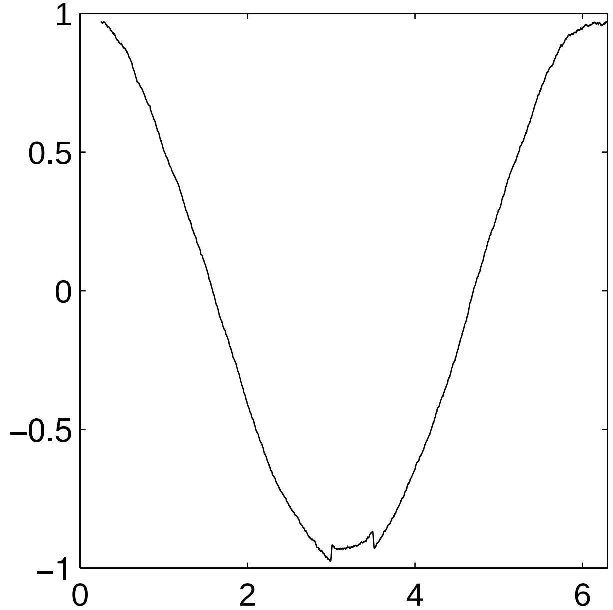
original data



filtered with $q = 20$



filtered with $q = 200$



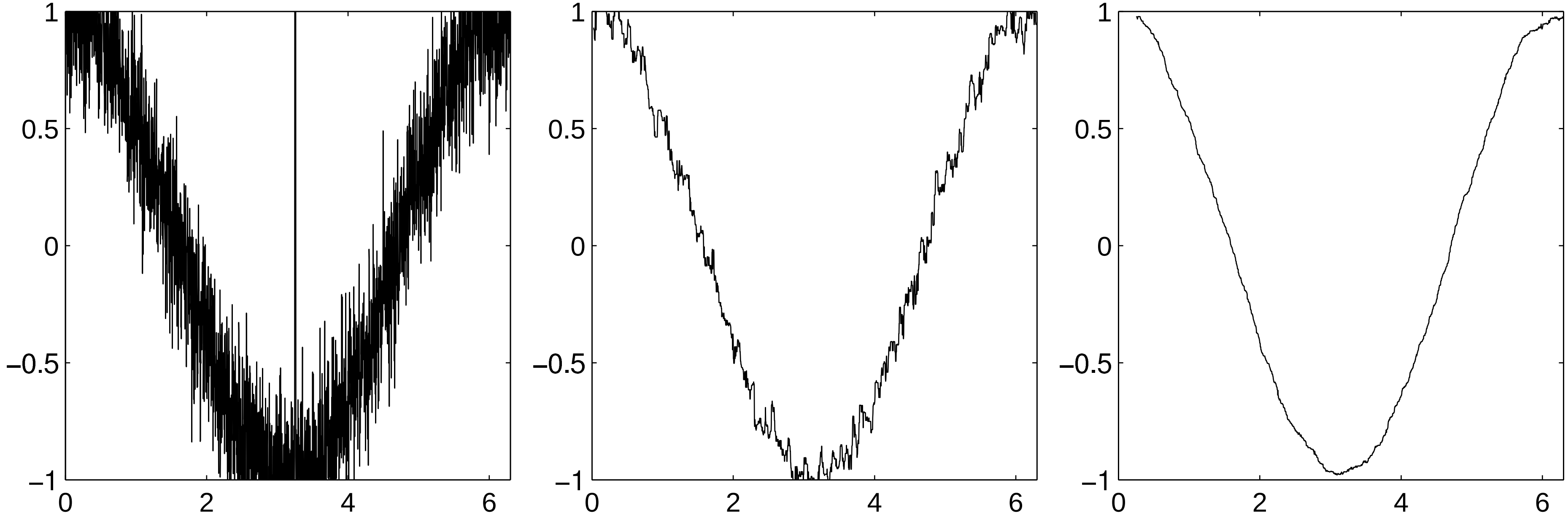
Prof. Dr. Thomas A. Runkler Copyright © 2020. All rights reserved.

Moving Median

- median $m_{kq} \in w_{kq} = \{x_{k-\frac{q}{2}}, \ldots, x_{k+\frac{q}{2}}\}$ or $\{x_{k-q}, \ldots, x_k\}$:

$$|\{x_i \in w_{kq} \mid x_i < m_{kq}\}| = |\{x_i \in w_{kq} \mid x_i > m_{kq}\}|$$

original data | filtered with $q = 20$ | filtered with $q = 200$



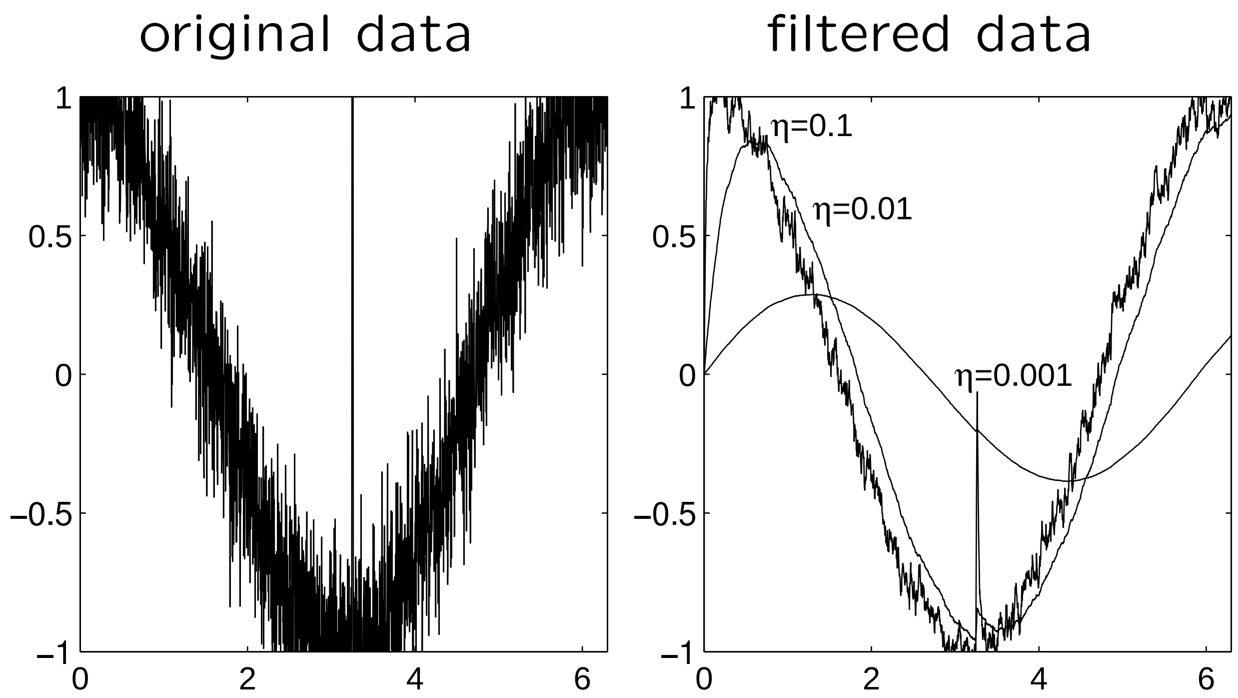
Prof. Dr. Thomas A. Runkler Copyright © 2020. All rights reserved.

Exponential Filter

- simple filter that forgets exponentially

$$y_k = y_{k-1} + \eta \cdot (x_k - y_{k-1}), \quad k = 2, \ldots, n \quad \eta \in [0, 1]$$

- initialization (standardized data) $y_0 = (0, \ldots, 0)$



Copyright © 2020. All rights reserved.

Discrete Linear Filter

- difference equation for linear filters of order q

$$y_k = \sum_{i=0}^{q} \frac{b_i}{a_0} \cdot x_{k-i} - \sum_{i=1}^{q} \frac{a_i}{a_0} \cdot y_{k-i}$$

- filter properties specified by coefficient vectors

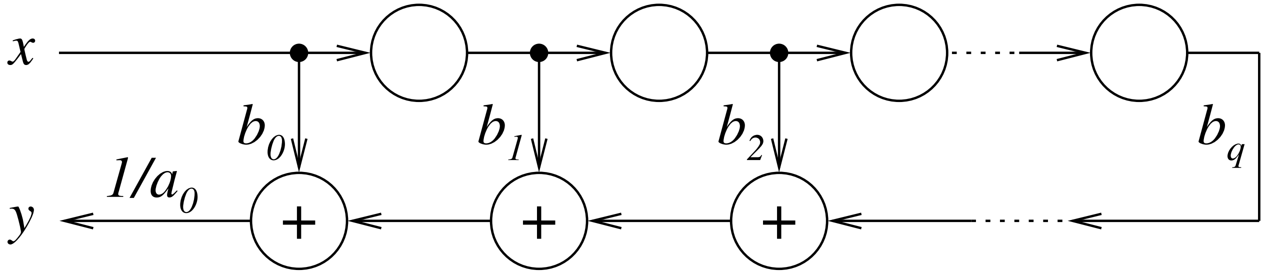
$$a = (a_0, \ldots, a_q), \quad b = (b_0, \ldots, b_q)$$

- finite impulse response (FIR): $a_1 = \ldots = a_q = 0$,
  otherwise infinite impulse response (IIR)
- exponential filter: $a = (1, \eta - 1)$, $b = (\eta, 0)$
- first order FIR low pass
  $=$ second order moving average
  $=$ first order Butterworth low pass with limit frequency 0.5:
  $a = (1)$, $b = (0.5, 0.5)$

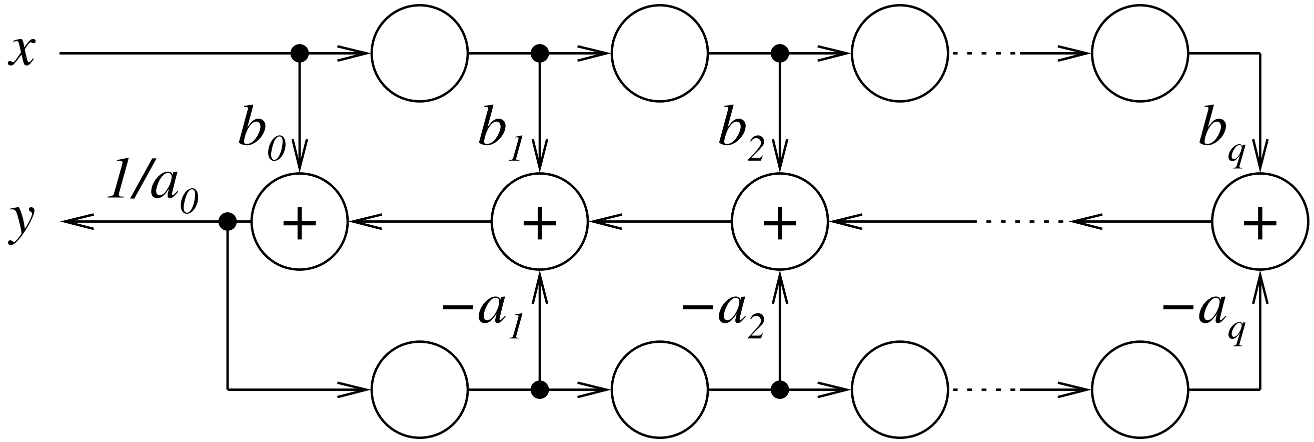
 Copyright © 2020. All rights reserved.

Discrete Linear Filter

- FIR filter

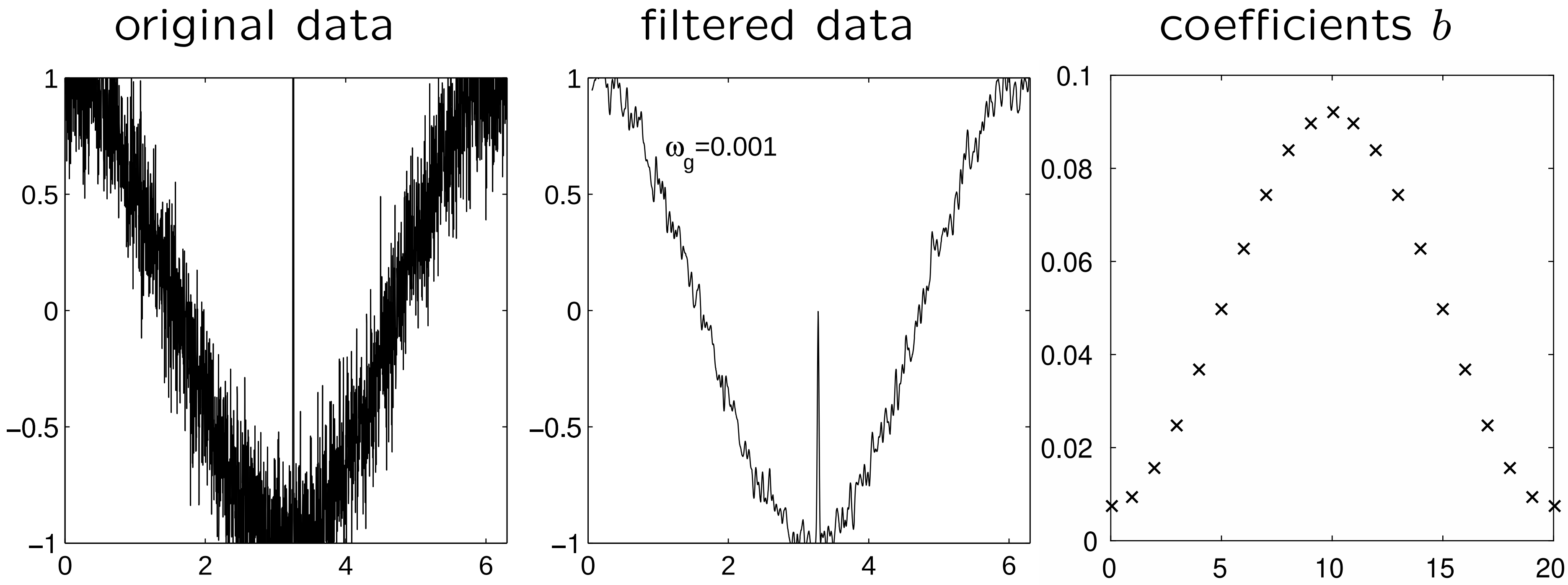


- IIR filter



Prof. Dr. Thomas A. Runkler Copyright © 2020. All rights reserved.

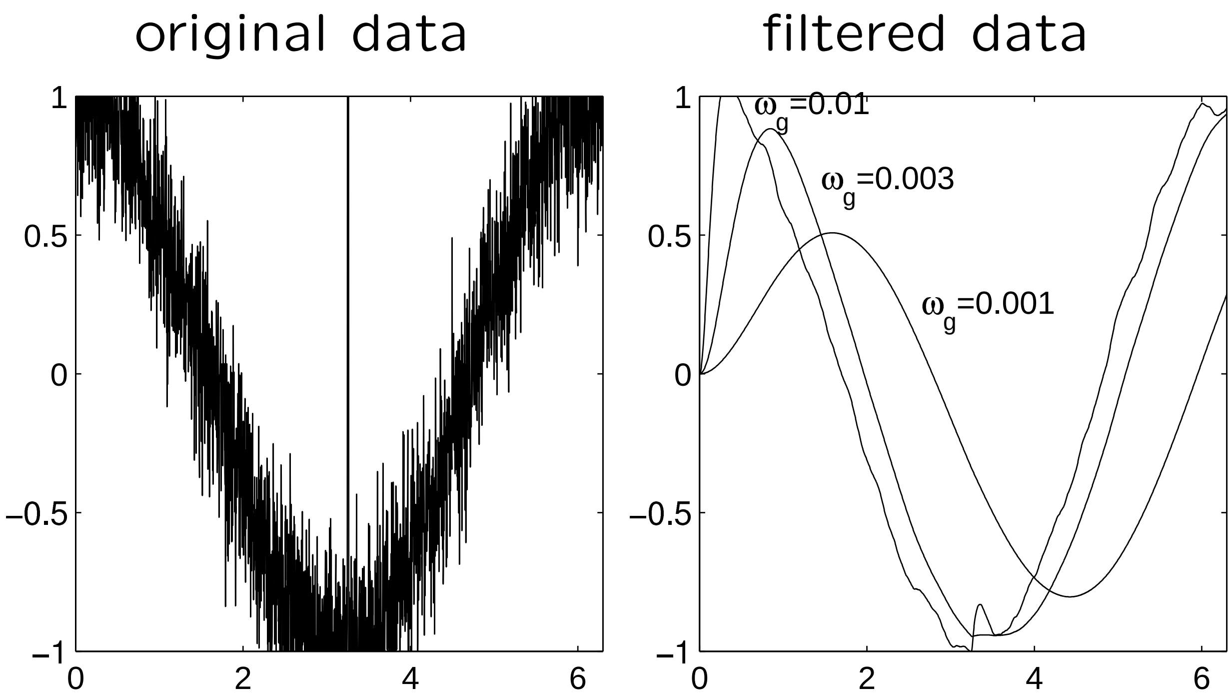
FIR low pass of order 20



- coefficient $a = (1)$
- coefficients b symmetric, $\sum_{i=0}^{q} b_i = 1$, $b_i > 0\ \forall i = 0, \ldots, q$

 Copyright © 2020. All rights reserved.

Second Order Butterworth Low Pass



ω_g	a_0	a_1	a_2	b_0	b_1	b_2
0.01	1	-1.96	0.957	$2.41 \cdot 10^{-4}$	$4.83 \cdot 10^{-4}$	$2.41 \cdot 10^{-4}$
0.003	1	-1.99	0.987	$2.21 \cdot 10^{-5}$	$4.41 \cdot 10^{-5}$	$2.21 \cdot 10^{-5}$
0.001	1	-2	0.996	$2.46 \cdot 10^{-6}$	$4.92 \cdot 10^{-6}$	$2.46 \cdot 10^{-6}$

 Copyright © 2020. All rights reserved.

Standardization

- problem: multi–dimensional data with considerably different component ranges
- observed hypercube

$$[x_{\text{min}}^{(1)}, x_{\text{max}}^{(1)}] \times \ldots \times [x_{\text{min}}^{(p)}, x_{\text{max}}^{(p)}]$$

- limits are arbitrary

$$x_{\text{min}}^{(i)} \neq \min_{k=1,\ldots,n} x_k^{(i)}, \quad x_{\text{max}}^{(i)} \neq \max_{k=1,\ldots,n} x_k^{(i)}$$

- hypercube standardization

$$y_k^{(i)} = \frac{x_k^{(i)} - x_{\text{min}}^{(i)}}{x_{\text{max}}^{(i)} - x_{\text{min}}^{(i)}}$$

Copyright © 2020. All rights reserved.

μ–σ Standardization

- mean

$$\bar{x}^{(i)} = \frac{1}{n} \sum_{k=1}^{n} x_k^{(i)}$$

- standard deviation

$$s_x^{(i)} = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k^{(i)} - \bar{x}^{(i)})^2} = \sqrt{\frac{1}{n-1} \left(\sum_{k=1}^{n} \left(x_k^{(i)} \right)^2 - n \left(\bar{x}^{(i)} \right)^2 \right)}$$

- μ–σ standardization

$$y_k^{(i)} = \frac{x_k^{(i)} - \bar{x}^{(i)}}{s_x^{(i)}}$$

 Copyright © 2020. All rights reserved.

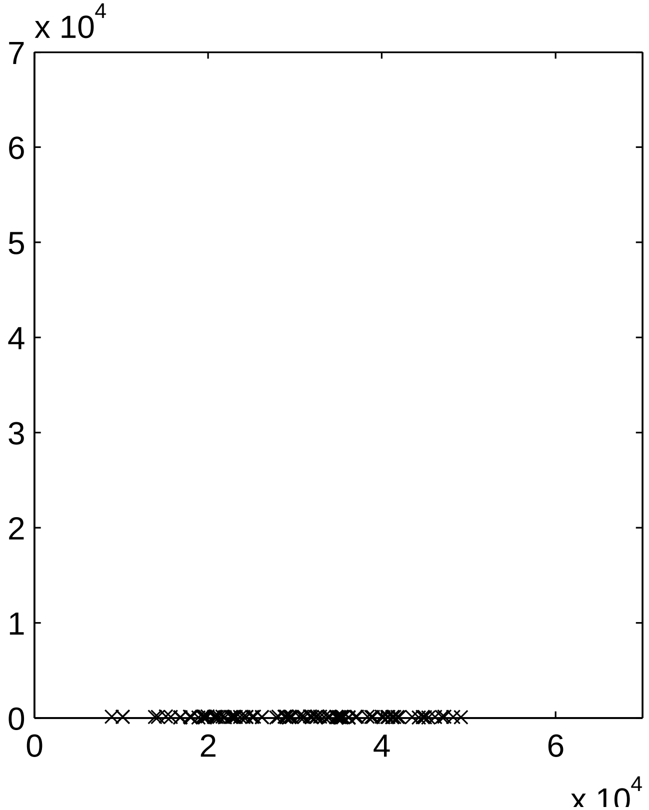
Example Standardization

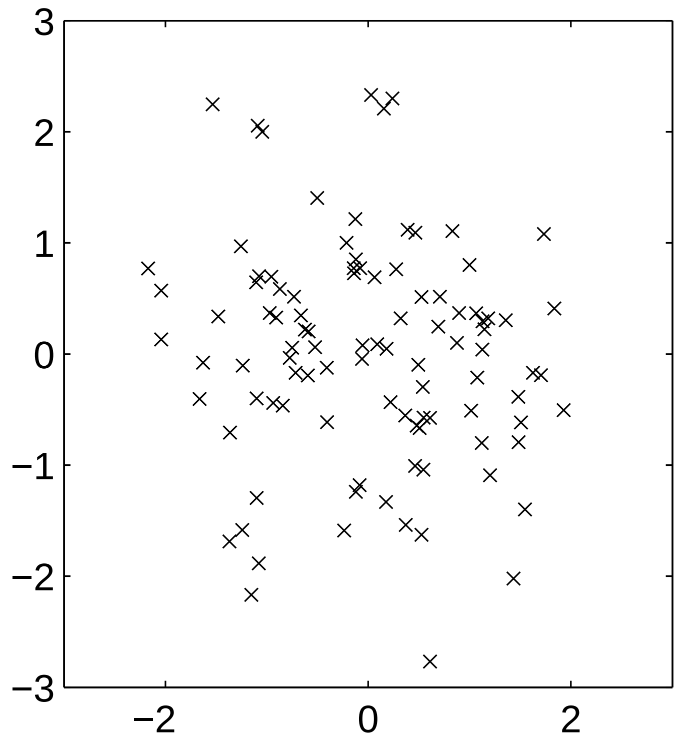
original data

$x_1, \ldots, x_{100}$

standardized data

$y_1, \ldots, y_{100}$





$$s_x = (9000, 30)$$

Prof. Dr. Thomas A. Runkler
Copyright © 2020. All rights reserved.

Data Transformations

- inverse transformation $f : R\backslash\{0\} \to R\backslash\{0\}$

$$f(x) = f^{-1}(x) = \frac{1}{x}$$

- root transformation $f : (c, \infty) \to R^+$

$$f(x) = \sqrt[b]{x - c}, \quad f^{-1}(x) = x^b + c, \quad c \in R,\ b > 0$$

- logarithmic transformation $f : (c, \infty) \to R$

$$f(x) = \log_b(x - c), \quad f^{-1}(x) = b^x + c, \quad c \in R,\ b > 0$$

- Fisher-Z transformation $f : (-1, 1) \to R$

$$f(x) = \operatorname{artanh}\ x = \frac{1}{2} \cdot \ln \frac{1 + x}{1 - x}, \quad f^{-1}(x) = \tanh\ x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Copyright © 2020. All rights reserved.

Data Merging

label	data x
$\vdots$	$\vdots$

$+$

label	data y
$\vdots$	$\vdots$

$=$

label	data x	data y
$\vdots$	$\vdots$	$\vdots$

- labels:
  - code, e.g. person, item
  - (relative) time, e.g. in sequential processes
  - (relative) location, e.g. position on an item
- problems:
  - similar labels considered equivalent
  - labels that do not match all data
  - labels that match multiple data

 Copyright © 2020. All rights reserved.

Chapter 4: Visualization

1. Diagrams
2. Principal Component Analysis
3. Multi Dimensional Scaling
4. Sammon Mapping
5. Auto–Encoder
6. Histograms
7. Spectral Analysis

 Copyright © 2020. All rights reserved.