A Comparison of Trend Estimators under Heteroscedasticity

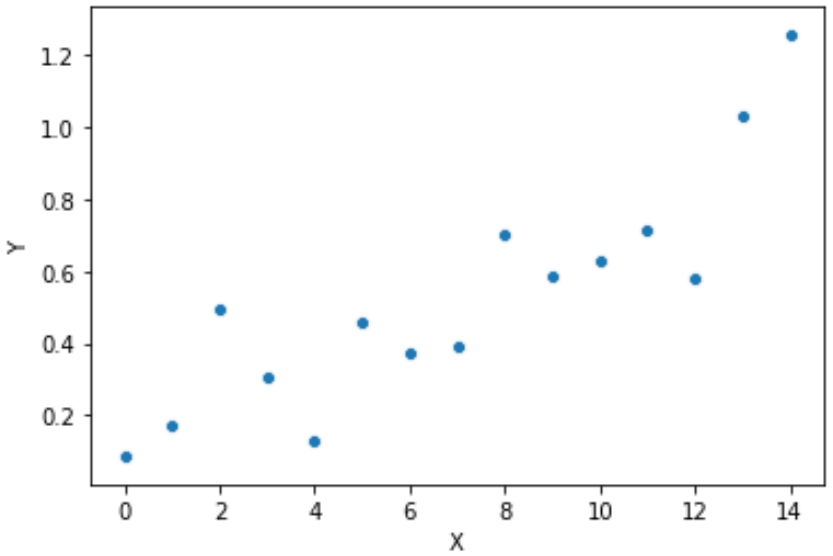
Jan Kalina, Petra Vidnerová, Jan Tichavský

The Czech Academy of Sciences, Institute of Computer Science, Prague

Outline

- Introduction
  - Trend estimation
  - Heteroscedasticity
- Estimators
  - Taut string
  - Neural networks
- Results
- Conclusion

Trend estimation

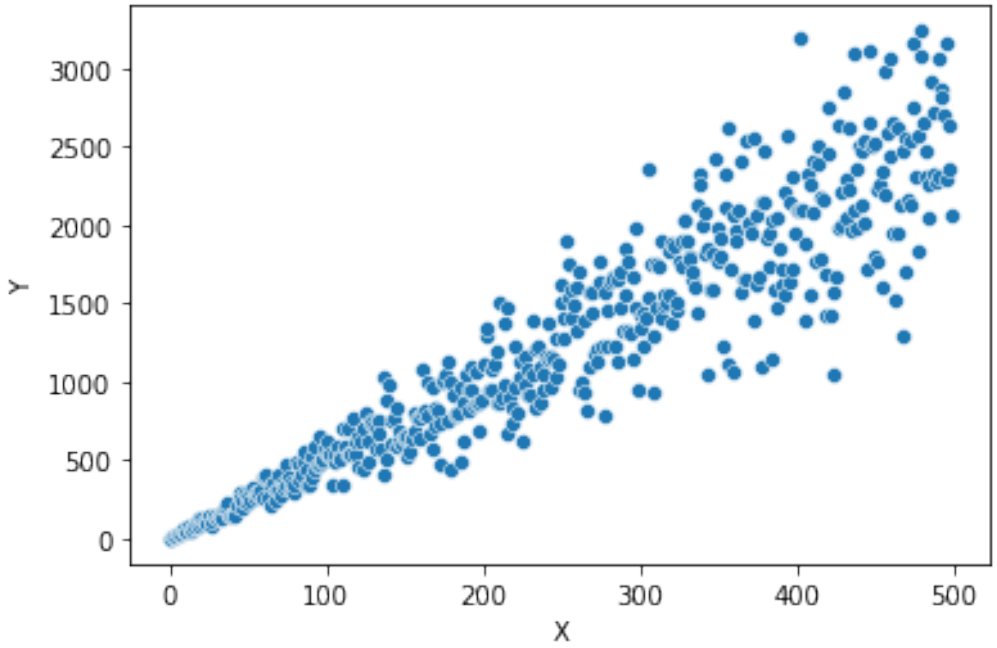


- Estimating and predicting a nonlinear trend of an observed continuous variable
- We consider special case in the form

$$Y_i = f_i + e_i \qquad \text{for } i = 1, \ldots, n$$

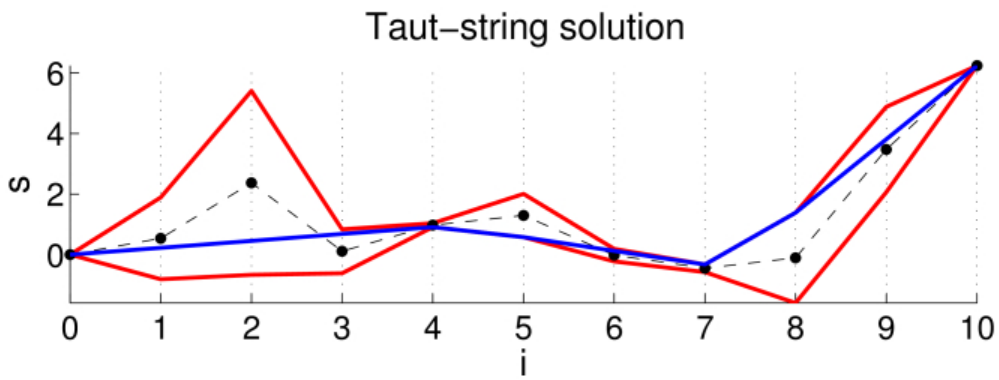
- No regressors

Heteroscedasticity



- Focus on heteroscedastic data
- Different variability of residuals
- Difficult for many ML algorithms

Taut string



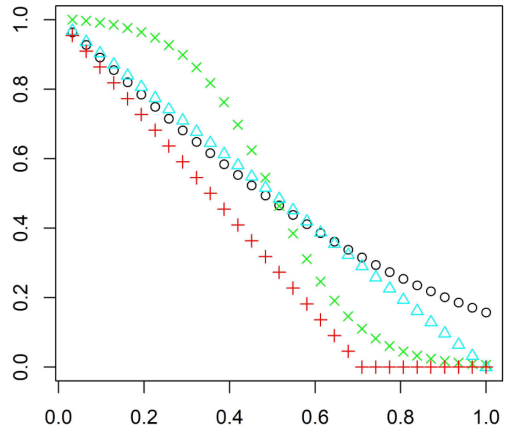
- Non-parametric regression, data approximation without a model
- Piece-wise linear
- No need to estimate hyperparameters
- Easy to use

Davies, P.L., Kovac, A.: *Local extremes, runs, strings and multiresolution.* Ann. Statist. 29, 1–65 (2001)

- Multilayer perceptrons (MLP), RBF network (RBF)
- LWS-MLP, LWS-RBF: least weighted squares variants

$$L = \sum_{i=1}^{n} w_i r_i^2,$$

where r_i are residuals $r_0 < r_1 < \ldots < r_n$ and w_i are weigths



- LTS-MLP, LTS-RBF: least trimmed squares variants
- back-MLP, back-RBF: robust approaches based on backward subsample selection

Aitken model

- Consider regression model in heteroscedastic setup

$$Y_i = f(X_i) + e_i, \qquad \text{var} e_i = \sigma^2 k_i, \text{ for } i = 1, \dots n$$

- $k_1, \dots, k_n$ are known

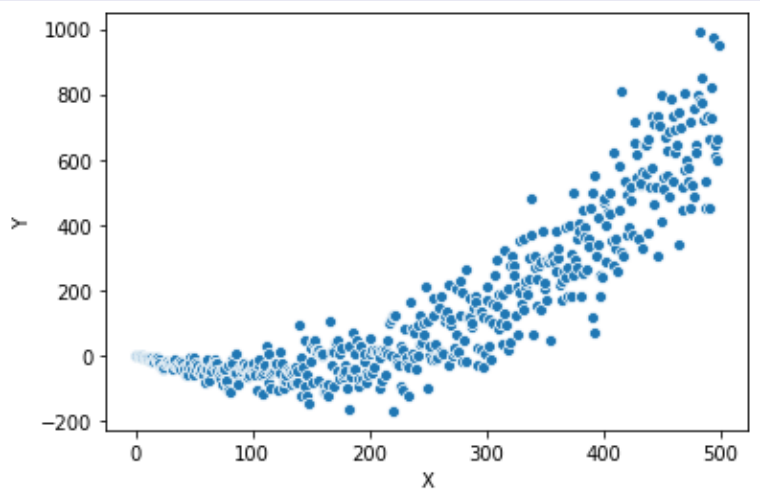
- Alternative model – *Aitken model*

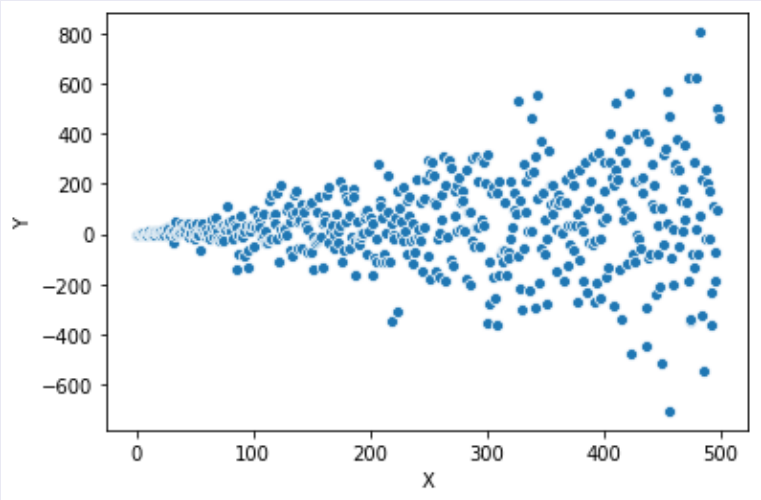
$$\frac{Y_i}{\sqrt{k_i}} = \frac{f(X_i)}{\sqrt{k_i}} + \frac{e_i}{\sqrt{k_i}}, \quad i = 1, \dots, n$$

- Errors are homoscedastic

$$\text{var} \frac{e_i}{\sqrt{k_i}} = \frac{1}{k_i} \text{var}\, e_i = \sigma^2$$

Synthetic datasets





- **LEFT: Dataset A,** $Y = \frac{(i-100)^2}{200} - 50 + e_i$
- **RIGHT: Dataset B,** $Y = \frac{i}{5} + e_i$

- Plus variant with additionaly artificial contamination

Results on Dataset A

Method	Raw data		Contam. data	
	MSE	TMSE	MSE	TMSE
Taut string	76.0	18.6	211.4	46.4
L_1-taut string	56.9	16.8	164.9	45.7
MLP	77.2	23.7	225.1	72.3
LWS-MLP	78.5	23.0	253.6	51.6
LTS-MLP	80.3	22.9	262.1	54.2
Back-MLP	82.4	23.4	270.4	61.8
RBF network	73.6	21.1	203.9	60.8
LWS-RBF	75.4	20.5	247.0	47.7
LTS-RBF	79.8	19.4	252.8	50.3
Back-RBF	80.1	20.8	264.3	52.6

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1}$$

$$\text{TMSE} = \frac{1}{h}\sum_{i=1}^{h} r_{(i)}^2 \text{ (where we use } h = \lfloor 0.8n \rfloor). \tag{2}$$

Crossvalidation on Dataset A

Method	Raw data MSE	Raw data TMSE	Contam. data MSE	Contam. data TMSE
	standard approach			
MLP	79.3	27.8	243.7	83.4
LWS-MLP	81.4	26.1	271.8	56.0
LTS-MLP	83.5	25.8	279.3	57.6
Back-MLP	85.0	26.9	286.4	63.1
RBF network	76.7	25.8	232.2	76.3
LWS-RBF	80.6	25.2	251.8	52.5
LTS-RBF	82.2	24.8	256.1	54.9
Back-RBF	84.9	25.3	263.2	59.4
	Aitken approach			
MLP	75.9	26.7	241.4	74.3
LWS-MLP	79.4	25.2	275.2	31.9
LTS-MLP	80.3	25.0	286.5	33.0
Back-MLP	83.6	27.6	303.6	37.4
RBF network	72.2	24.3	237.0	69.8
LWS-RBF	74.9	25.7	259.3	30.7
LTS-RBF	75.1	24.7	262.1	31.6
Back-RBF	78.5	26.8	280.7	34.7

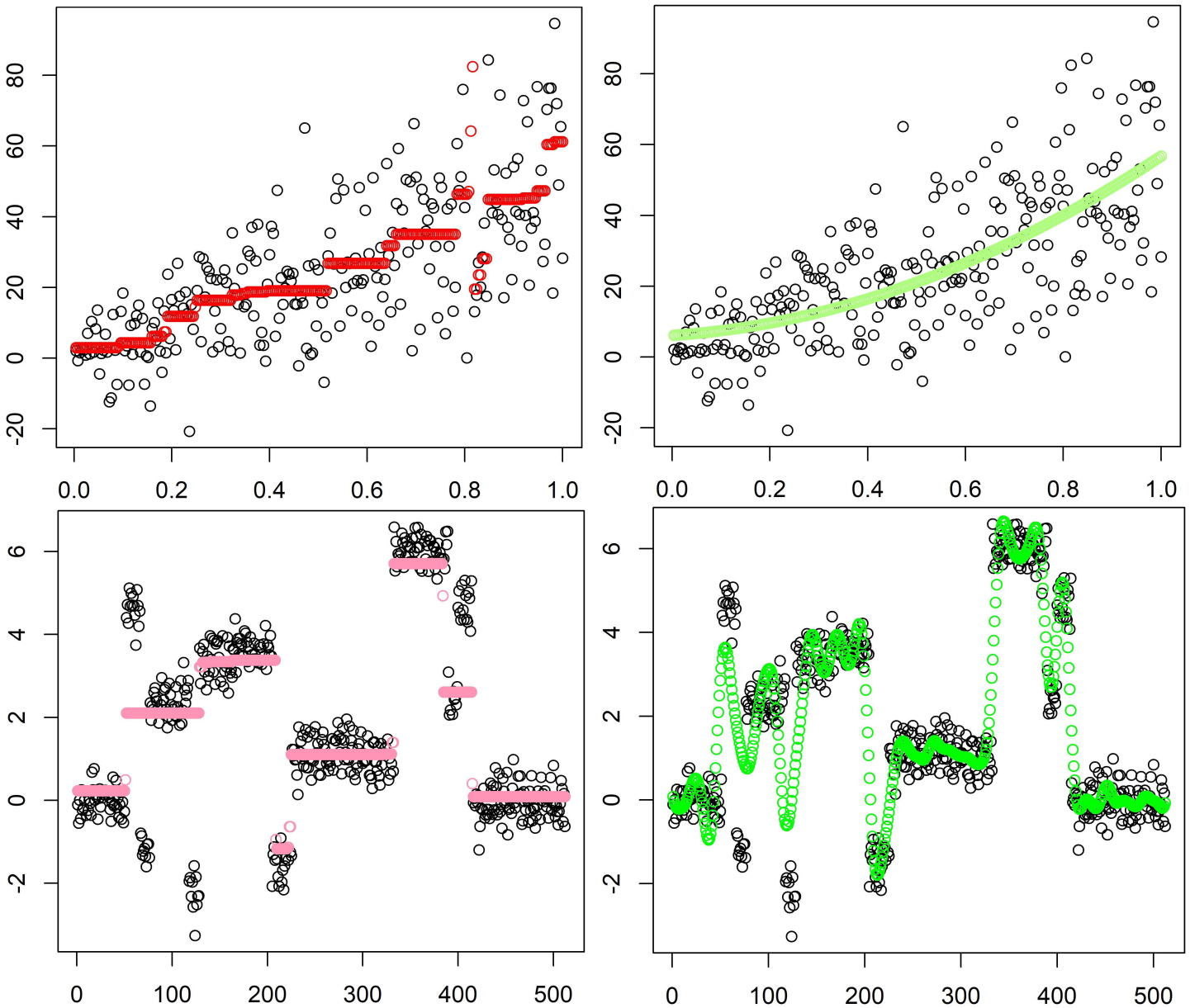
Results ond Dataset B

Method	Raw data		Contam. data	
	MSE	TMSE	MSE	TMSE
Taut string	142.8	34.2	500.2	62.6
L_1-taut string	98.0	27.1	204.2	53.0
MLP	134.4	37.8	480.8	98.7
LWS-MLP	145.1	37.5	494.3	64.1
LTS-MLP	148.9	37.2	491.6	66.4
Back-MLP	156.3	38.1	503.7	70.3
RBF network	125.3	36.1	463.7	97.0
LWS-RBF	136.6	35.7	482.5	61.2
LTS-RBF	138.2	35.3	481.1	63.9
Back-RBF	147.0	36.8	487.8	65.4

Crossvalidation on Dataset B

Method	Raw data MSE	Raw data TMSE	Contam. data MSE	Contam. data TMSE
	standard approach			
MLP	145.1	40.4	516.7	115.6
LWS-MLP	148.2	38.5	541.9	70.5
LTS-MLP	147.6	39.9	538.4	72.6
Back-MLP	153.7	38.8	568.1	78.2
RBF network	136.8	38.8	483.2	109.4
LWS-RBF	139.4	37.0	509.7	65.3
LTS-RBF	138.3	36.2	506.3	67.8
Back-RBF	145.7	38.1	511.0	73.9
	Aitken approach			
MLP	133.2	37.0	452.3	104.6
LWS-MLP	139.1	35.3	490.7	46.1
LTS-MLP	137.6	34.9	494.8	45.3
Back-MLP	145.7	36.1	503.4	49.4
RBF network	128.2	33.5	437.9	99.2
LWS-RBF	132.1	32.2	474.7	42.0
LTS-RBF	130.5	32.8	478.0	44.8
Back-RBF	141.8	33.1	488.1	47.5

Illustration examples



Conclusion

- We tested several trend estimators – taut string, neural networks
- Robust variants were considered
- Heteroscedastic data with possibility of additional contamination

- For contaminated data the robust versions achieve better performance
- The Aitken model brings further improvement
  $\rightarrow$ **use Aitken model with heteroscedastic data**

- Taut string achieves low errors, but it tends to overfit data
- Appealing for very specific situations
- Limited to one dimension

Questions?