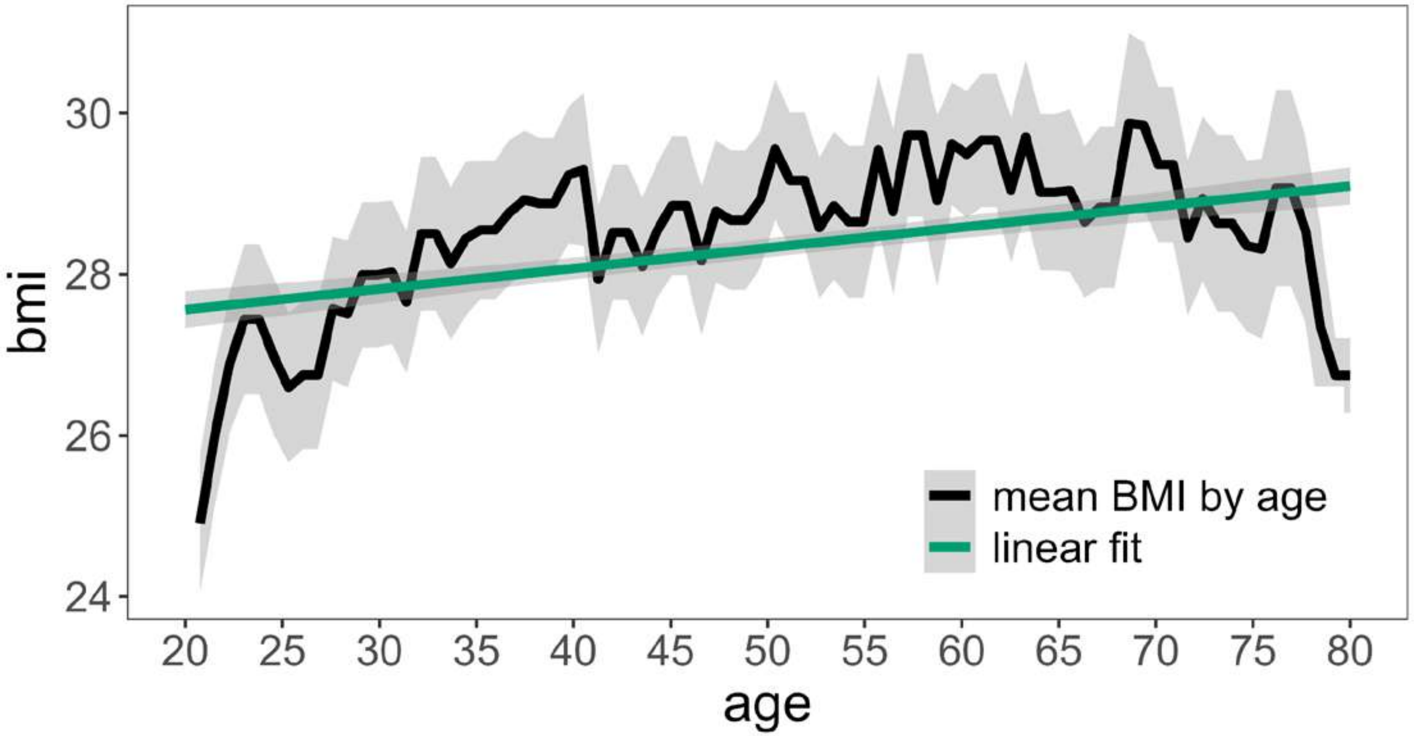
A categorization and comparison of performance measures for estimated non-linear associations with an outcome

Theresa Ullmann, Georg Heinze, Michal Abrahamowicz, Aris Perperoglou, Willi Sauerbrei, Matthias Schmid, Daniela Dunkler, for TG2 of the STRATOS initative

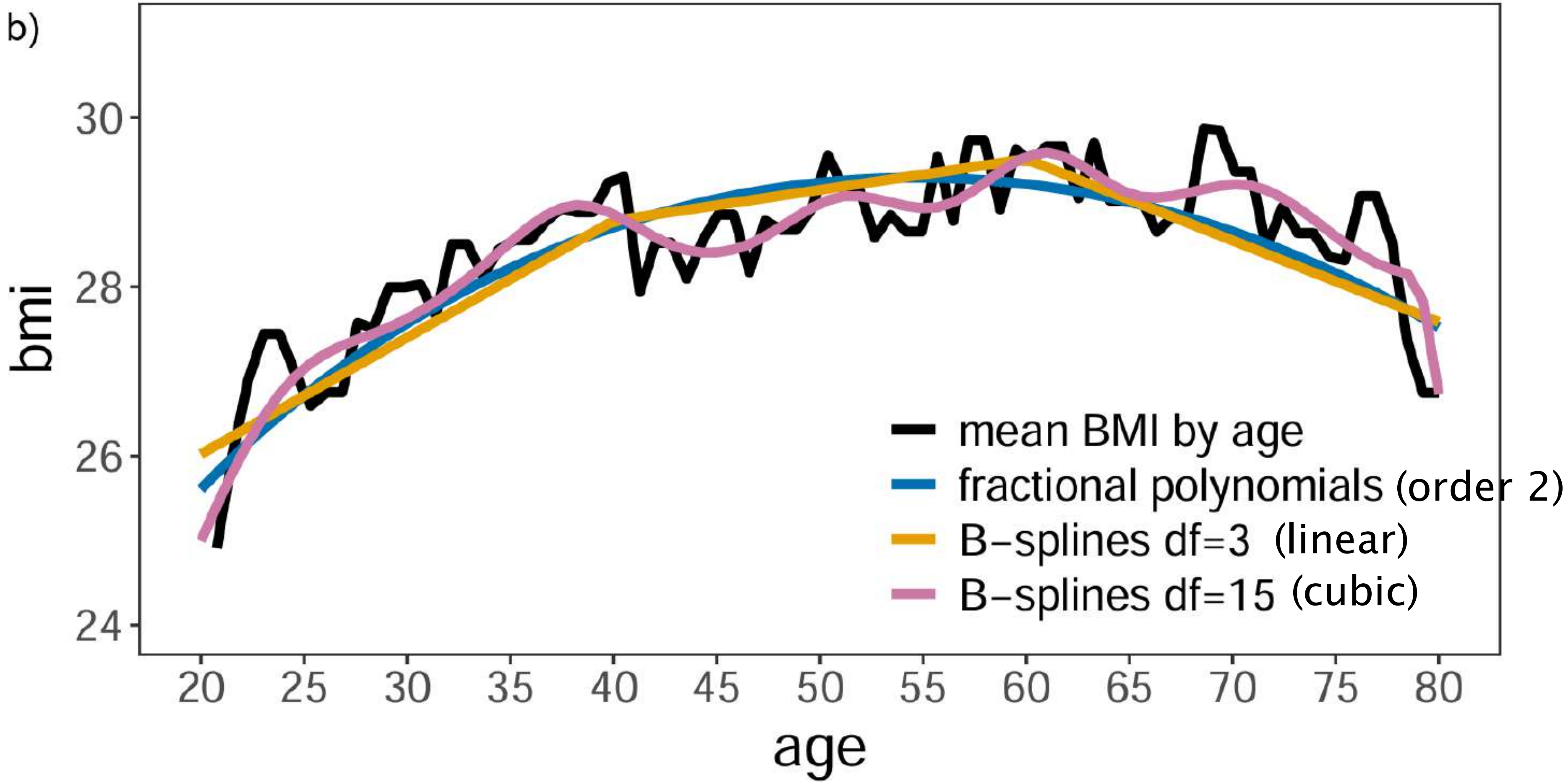
Presenter: Georg Heinze
Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna

Background & motivation

- Consider the association of BMI with age (NHANES)
- How to separate systematic from unsystematic variation?
- Linear model probably a poor smoother

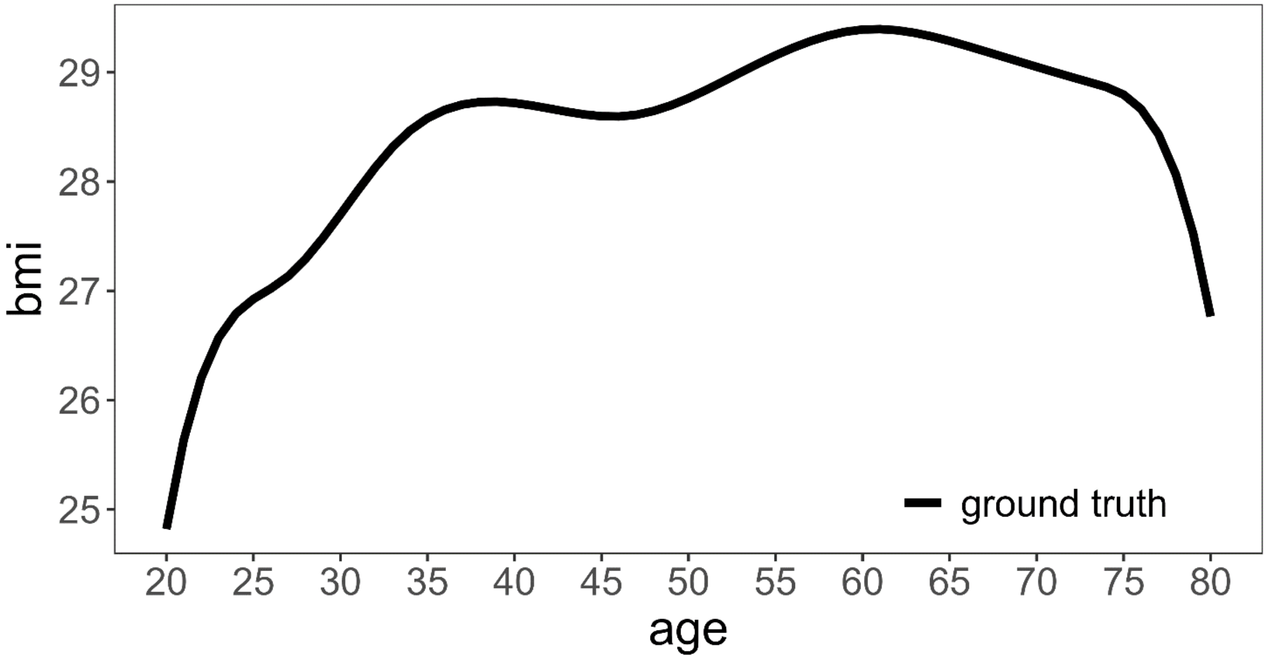


More smoothers...



A simulation study to compare methods?

- Aim: to compare the performance of different methods of nonlinear modeling
- Data generation mechanism:

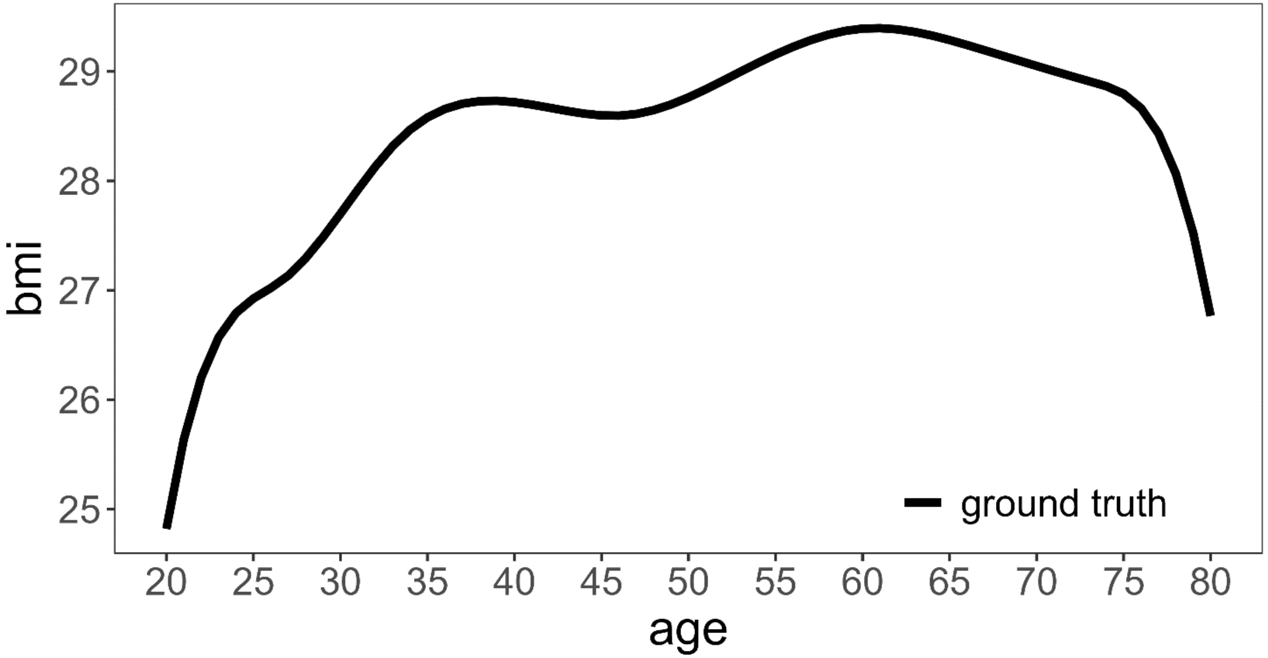


- Estimand: predicted BMI

ADEMP: Morris et al, StatMed 2019

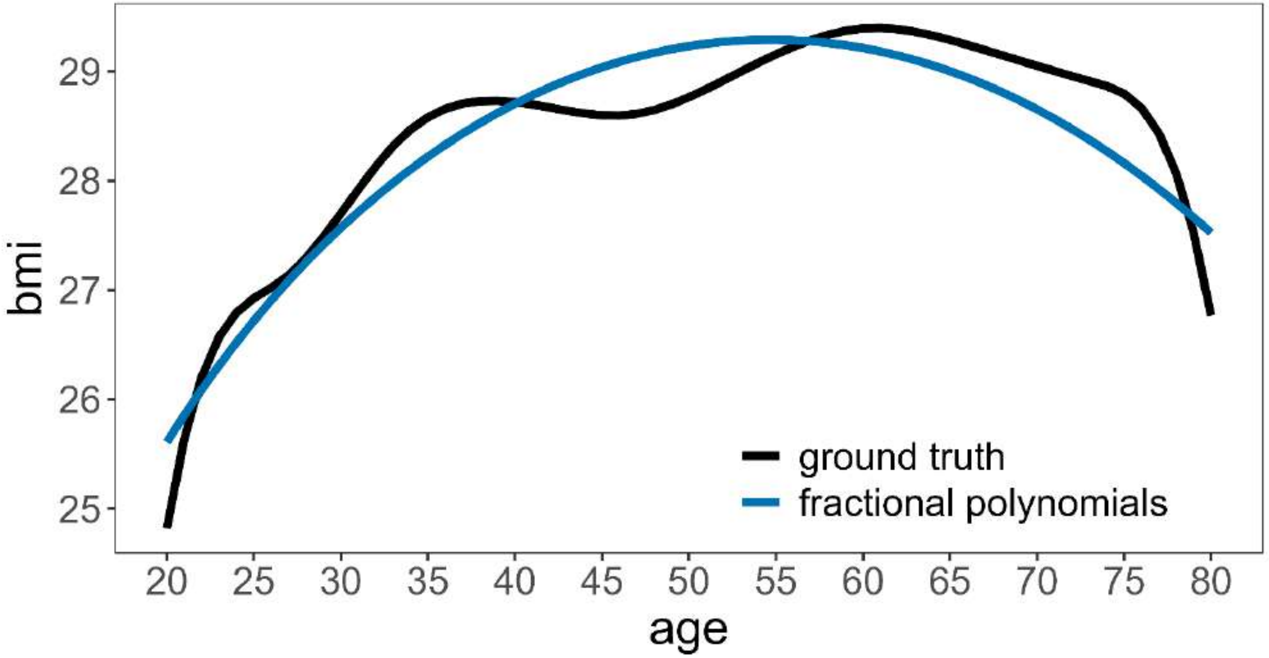
A simulation study to compare methods?

- Methods:



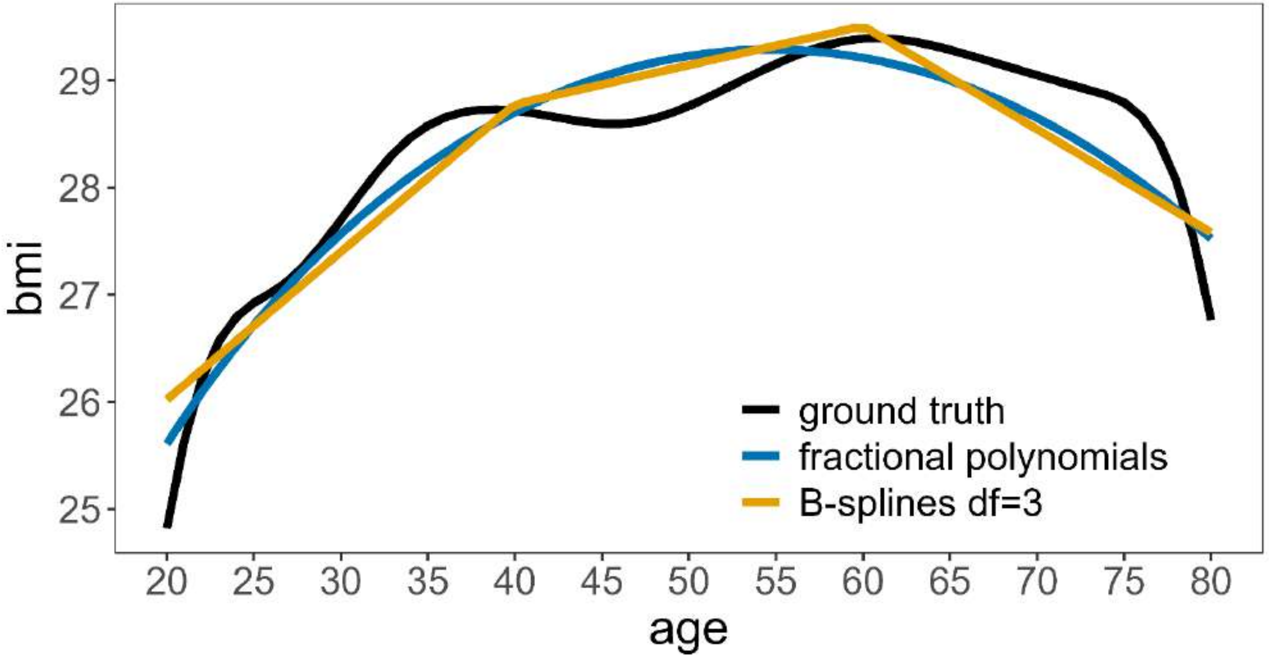
A simulation study to compare methods?

- Methods:



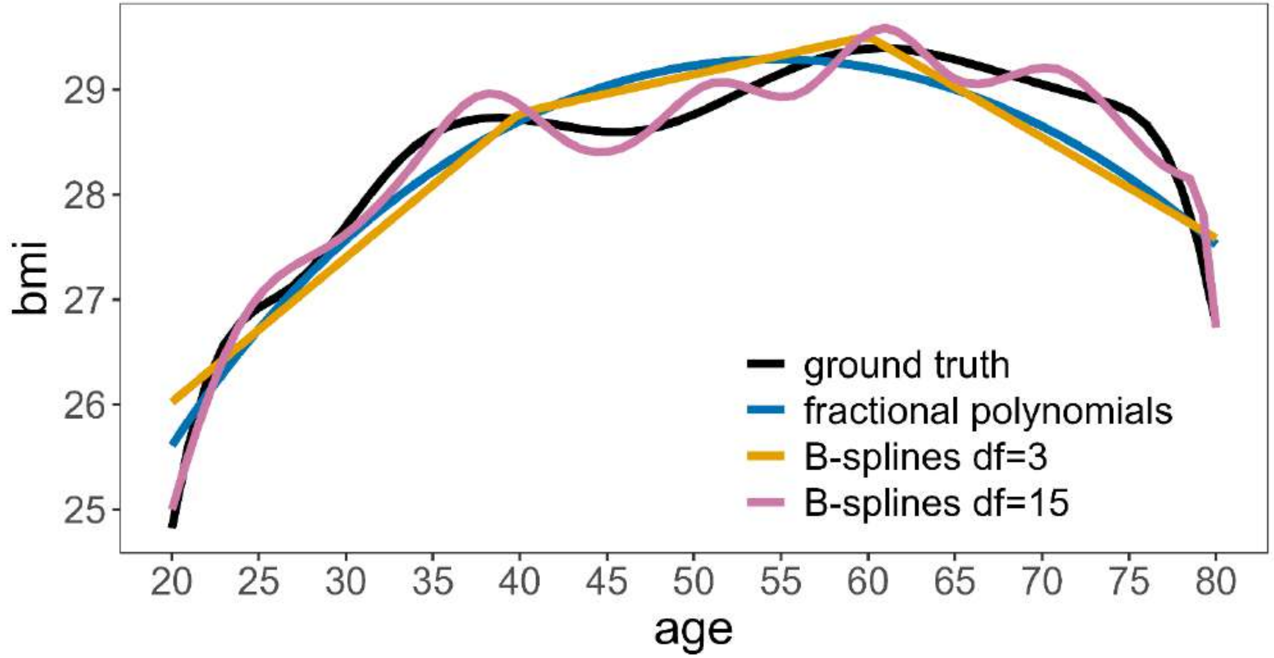
A simulation study to compare methods?

- Methods:



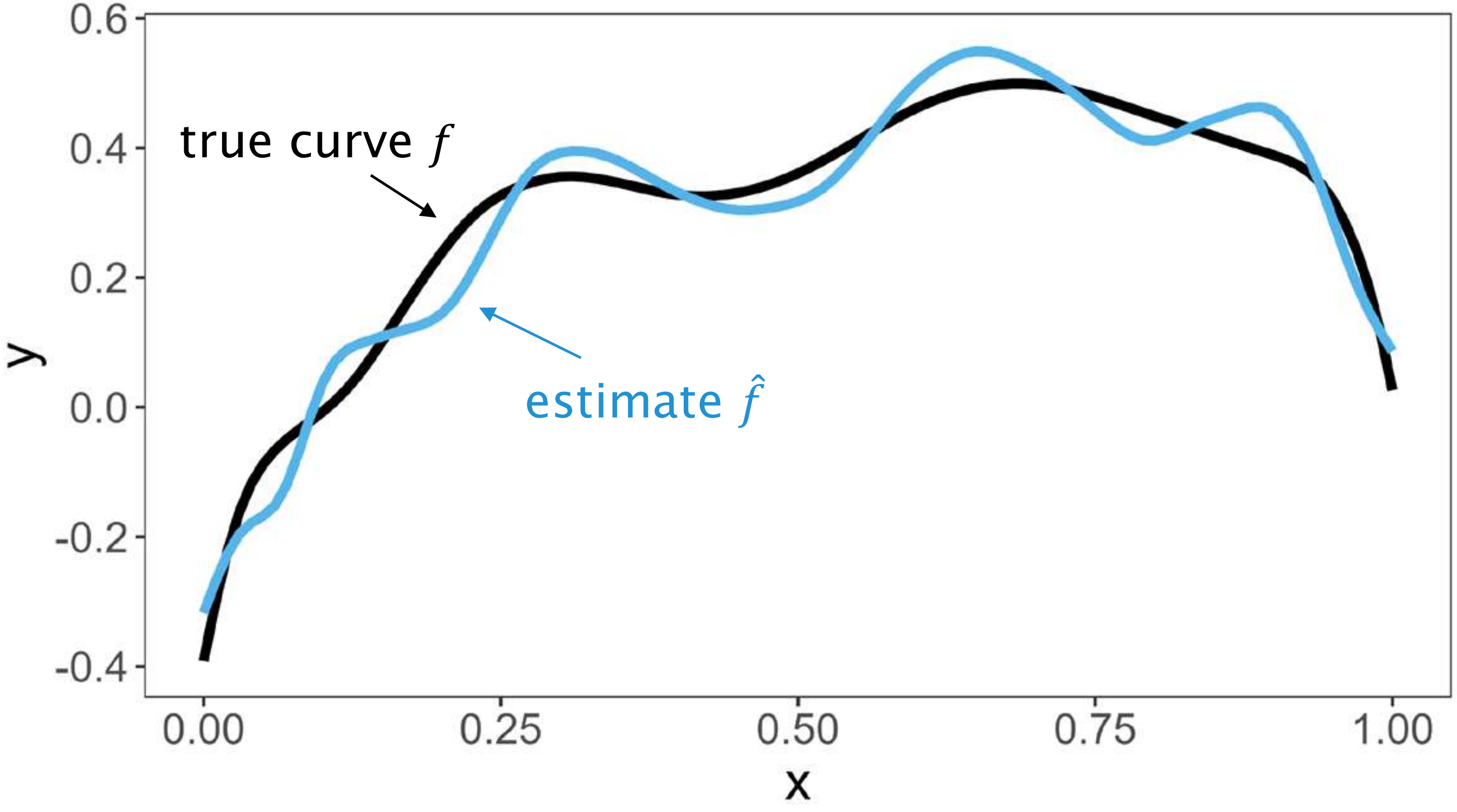
A simulation study to compare methods?

- Methods:



A simulation study to compare methods?

- Performance measures: compare estimated with true curve



A simulation study to compare methods?

- Performance measures:

$$\int_{F_X^{-1}(0.01)}^{F_X^{-1}(0.99)} |\hat{f}(x) - f(x)|\hat{p}(x)\mathrm{d}x$$

Buchholz et al. (2014) (see also Govindarajulu et al., 2007)

A simulation study to compare methods?

- Performance measures:

$$\int_{F_X^{-1}(0.01)}^{F_X^{-1}(0.99)} |\hat{f}(x) - f(x)|\hat{p}(x)\mathrm{d}x$$

Buchholz et al. (2014) (see also Govindarajulu et al., 2007)

$$\int_{F_X^{-1}(0.05)}^{F_X^{-1}(0.95)} \left(\hat{f}'(x) - f'(x)\right)^2 \mathrm{d}F_X(x)$$

Binder et al. (2011)

A simulation study to compare methods?

- Performance measures:

Region of interest: 1st to 99th percentile of F_X

$$\int_{F_X^{-1}(0.01)}^{F_X^{-1}(0.99)} |\hat{f}(x) - f(x)|\hat{p}(x)\mathrm{d}x$$

Buchholz et al. (2014) (see also Govindarajulu et al., 2007)

Region of interest: 5th to 95th percentile of F_X

$$\int_{F_X^{-1}(0.05)}^{F_X^{-1}(0.95)} \left(\hat{f}'(x) - f'(x)\right)^2 \mathrm{d}F_X(x)$$

Binder et al. (2011)

A simulation study to compare methods?

- Performance measures:

Absolute loss

$$\int_{F_X^{-1}(0.01)}^{F_X^{-1}(0.99)} |\hat{f}(x) - f(x)|\hat{p}(x)\mathrm{d}x$$

Buchholz et al. (2014) (see also Govindarajulu et al., 2007)

$$\int_{F_X^{-1}(0.05)}^{F_X^{-1}(0.95)} \left(\hat{f}'(x) - f'(x)\right)^2 \mathrm{d}F_X(x)$$

Binder et al. (2011)

Quadratic loss

A simulation study to compare methods?

- Performance measures:

function

$$\int_{F_X^{-1}(0.01)}^{F_X^{-1}(0.99)} |\hat{f}(x) - f(x)|\hat{p}(x)\mathrm{d}x$$

Buchholz et al. (2014) (see also Govindarajulu et al., 2007)

$$\int_{F_X^{-1}(0.05)}^{F_X^{-1}(0.95)} \left(\hat{f}'(x) - f'(x)\right)^2 \mathrm{d}F_X(x)$$

Binder et al. (2011)

first derivative

A simulation study to compare methods?

- Performance measures:

Integral weighted with precision

$$\int_{F_X^{-1}(0.01)}^{F_X^{-1}(0.99)} |\hat{f}(x) - f(x)|\hat{p}(x)\mathrm{d}x$$

Buchholz et al. (2014) (see also Govindarajulu et al., 2007)

$$\int_{F_X^{-1}(0.05)}^{F_X^{-1}(0.95)} \left(\hat{f}'(x) - f'(x)\right)^2 \mathrm{d}F_X(x)$$

Binder et al. (2011)

Integral over distribution of X

A simulation study to compare methods?

- Performance measures:

Region of interest × Loss × Functional characteristic × Aggregation × …

A simulation study to compare methods?

- Performance measures:

?

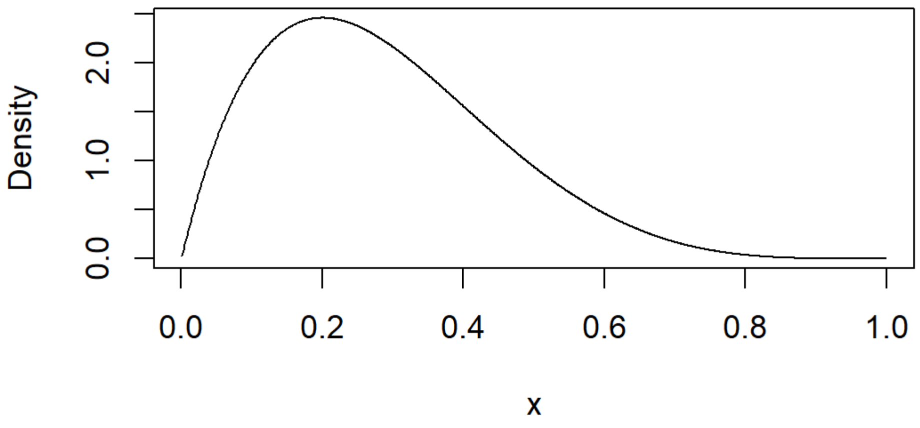
Region of interest × Loss × Function characteristic × Aggregation × …

Aims of this project

- To provide a comprehensive characterization of performance measures to be used in methods comparison studies
  - Define aspects of such measures
  - Suggest sensible combinations of choices for each of the aspects
- To demonstrate with simple illustrative examples and some hypothetical ‚methods‘
  - How the resulting performance measures behave
  - That different performance measures capture different aspects of behaviour

The aspects:

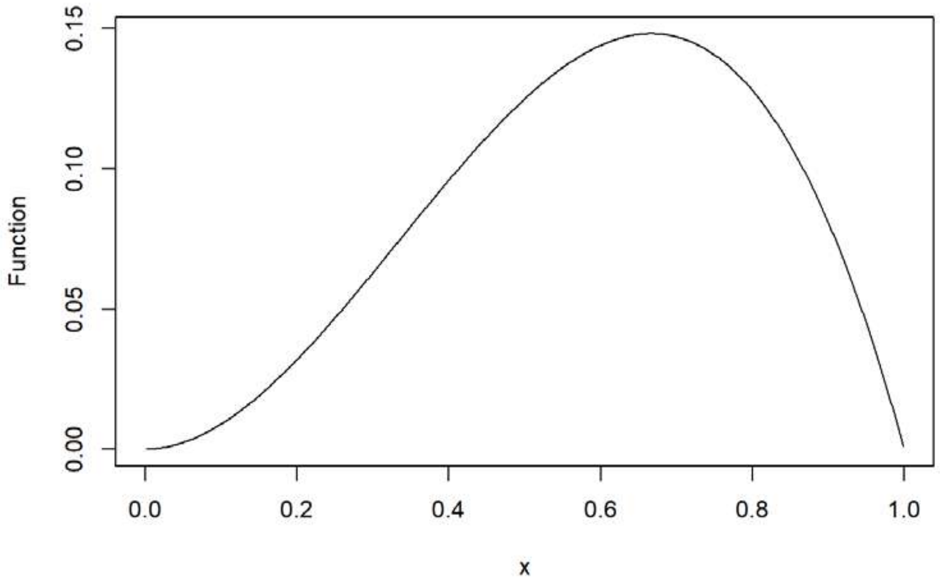
- Localization: Where are we looking at?

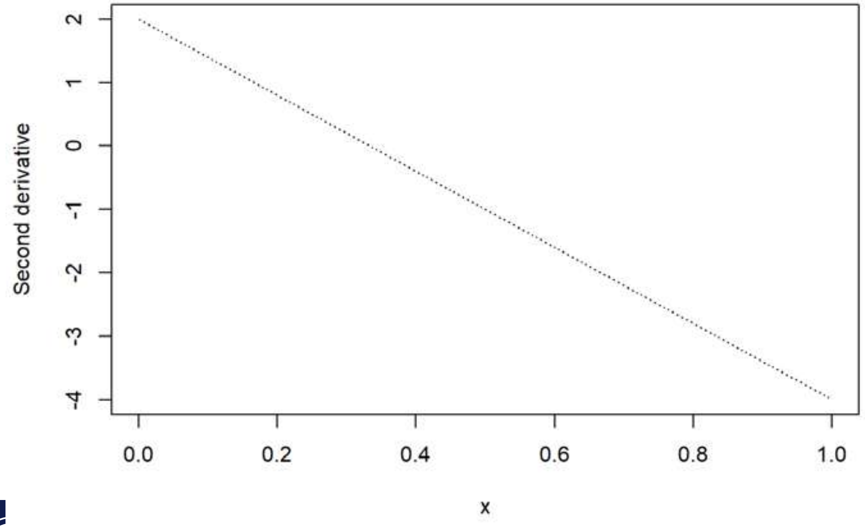


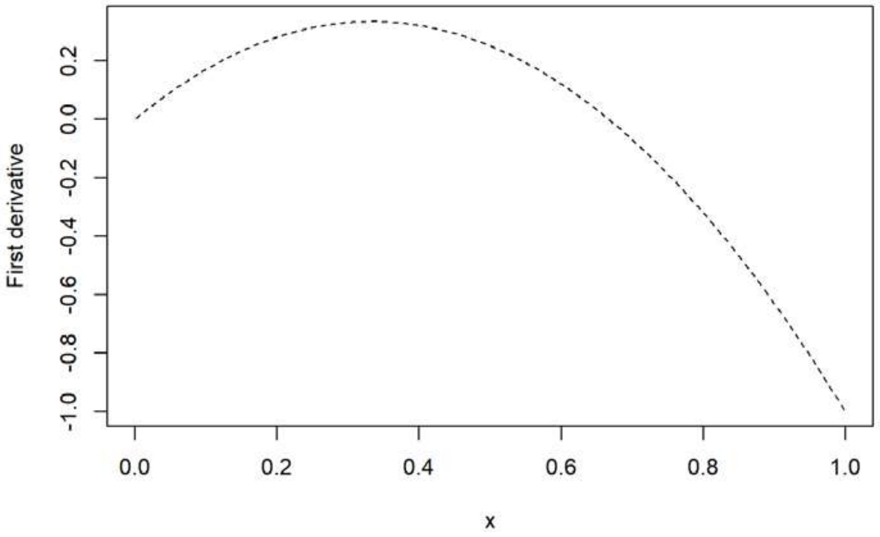
  - The full range of values (global)
  - A subrange (region)
  - A single value (point)

The aspects:

- Functional characteristic:
  - The function itself
  - First derivative
  - Second derivative







The aspects:

- Type of loss:
  - Difference: $m(x) = \hat{f}(x) - f(x)$
  - Absolute difference: $m(x) = |\hat{f}(x) - f(x)|$
  - Quadratic difference: $m(x) = \left(\hat{f}(x) - f(x)\right)^2$
  - ϵ-level accuracy: $m(x) = \mathrm{I}(\left|\hat{f}(x) - f(x)\right| \leq \epsilon)$

If we consider the range or a region:

- Axis of aggregation:
  - Y
    - Integration over dx: $\int m(x)\, dx$
    - Integration over dF(x) [=expected value): $\int m(x)\, dF(x)$
  - X
    - Location of maximum/minimum f(x) (=$argmax\left(\hat{f}(x)\right), argmin\left(\hat{f}(x)\right)$)
    - Number of roots (e.g. of $\hat{f}'(x)$)

Combining these aspects

Select the performance measure

Localization:

- [x] Range
- [ ] Point

Functional characteristic:

- [x] f(x)
- [ ] f'(x)
- [ ] f''(x)

Loss:

- [x] Difference
- [ ] Absolute
- [ ] Squared
- [ ] Epsilon-level accuracy

Axis of aggregation:

- [x] Y
- [ ] X

Type of aggregation:

- [x] Integration over dx
- [ ] Expectation over dF_X
- [ ] Quantile with respect to F_X
- [ ] Maximum
- [ ] Minimum

Scope of aggregration:

- [x] whole range $[0, 1]$
- [ ] subrange $[F_X^{-1}(0.05), F_X^{-1}(0.95)]$

$$= \int \left(\hat{f}(x) - f(x)\right) dx$$

„mean deviation“

Combining these aspects

Select the performance measure

Localization:
- (●) Range
- ( ) Point

Functional characteristic:
- (●) f(x)
- ( ) f'(x)
- ( ) f''(x)

Loss:
- ( ) Difference
- (●) Absolute
- ( ) Squared
- ( ) Epsilon-level accuracy

Axis of aggregation:
- (●) Y
- ( ) X

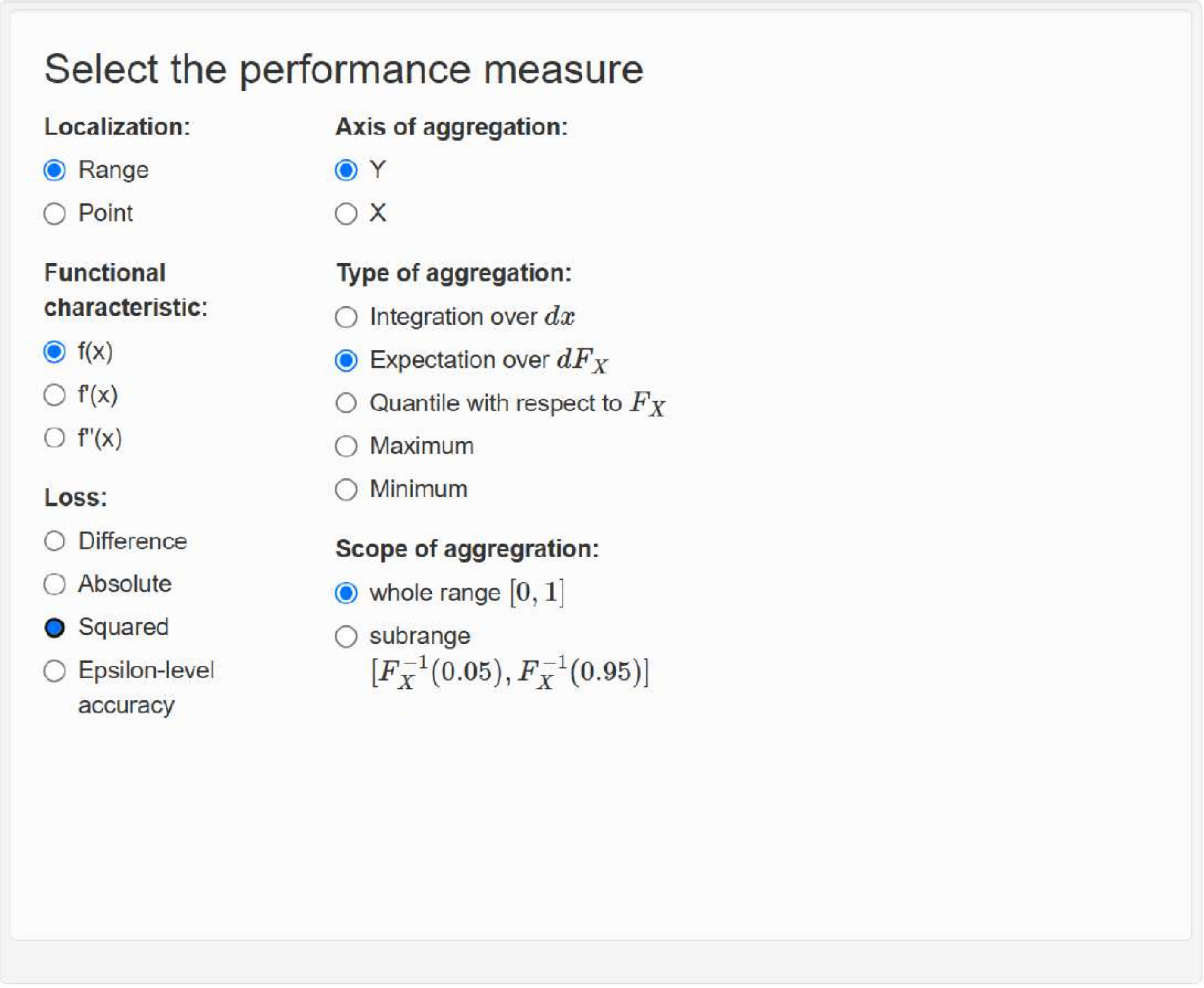
Type of aggregation:
- (●) Integration over dx
- ( ) Expectation over dF_X
- ( ) Quantile with respect to F_X
- ( ) Maximum
- ( ) Minimum

Scope of aggregration:
- (●) whole range $[0, 1]$
- ( ) subrange $[F_X^{-1}(0.05), F_X^{-1}(0.95)]$

$$=\int |\hat{f}(x) - f(x)| dx$$

„mean absolute deviation“

Combining these aspects



Select the performance measure

Localization:

- (●) Range
- ( ) Point

Functional characteristic:

- (●) f(x)
- ( ) f'(x)
- ( ) f''(x)

Loss:

- ( ) Difference
- ( ) Absolute
- (●) Squared
- ( ) Epsilon-level accuracy

Axis of aggregation:

- (●) Y
- ( ) X

Type of aggregation:

- ( ) Integration over dx
- (●) Expectation over dF_X
- ( ) Quantile with respect to F_X
- ( ) Maximum
- ( ) Minimum

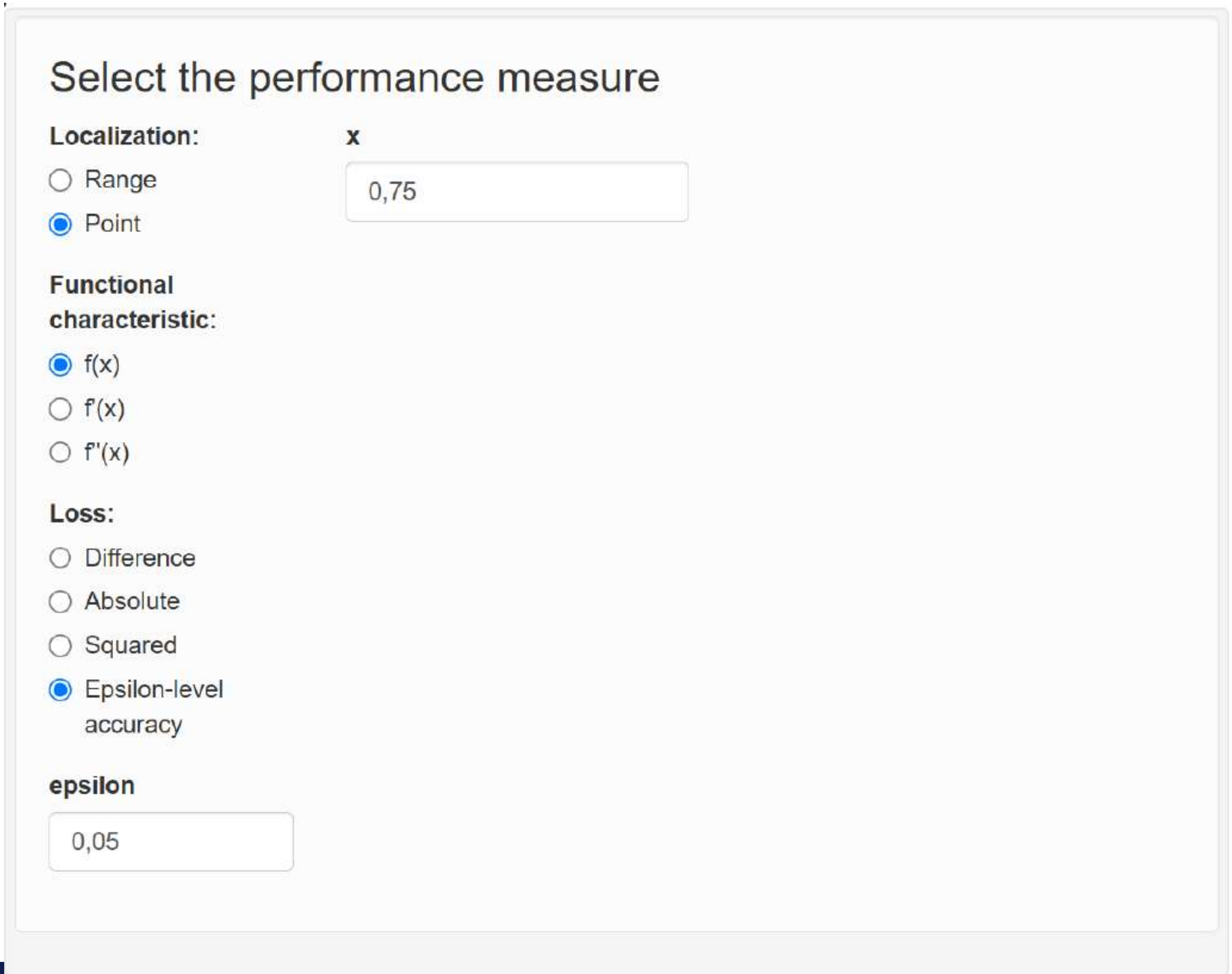
Scope of aggregration:

- (●) whole range $[0, 1]$
- ( ) subrange $[F_X^{-1}(0.05), F_X^{-1}(0.95)]$

$$= \int \left(\hat{f}(x) - f(x)\right)^2 dF(x)$$

„expected (over $F(x)$) squared deviation“

Combining these aspects



$$= I\left(\left|\hat{f}(0.75) - f(0.75)\right| \leq 0.05\right)$$

„within $f(x) \pm 0.05$ at $x = 0.75$“

Combining these aspects

Select the performance measure

Localization:

- (●) Range
- ( ) Point

Functional characteristic:

- ( ) f(x)
- ( ) f'(x)
- (●) f''(x)

Loss:

- ( ) Difference
- ( ) Absolute
- (●) Squared
- ( ) Epsilon-level accuracy

Axis of aggregation:

- (●) Y
- ( ) X

Type of aggregation:

- (●) Integration over dx
- ( ) Expectation over dF_X
- ( ) Quantile with respect to F_X
- ( ) Maximum
- ( ) Minimum

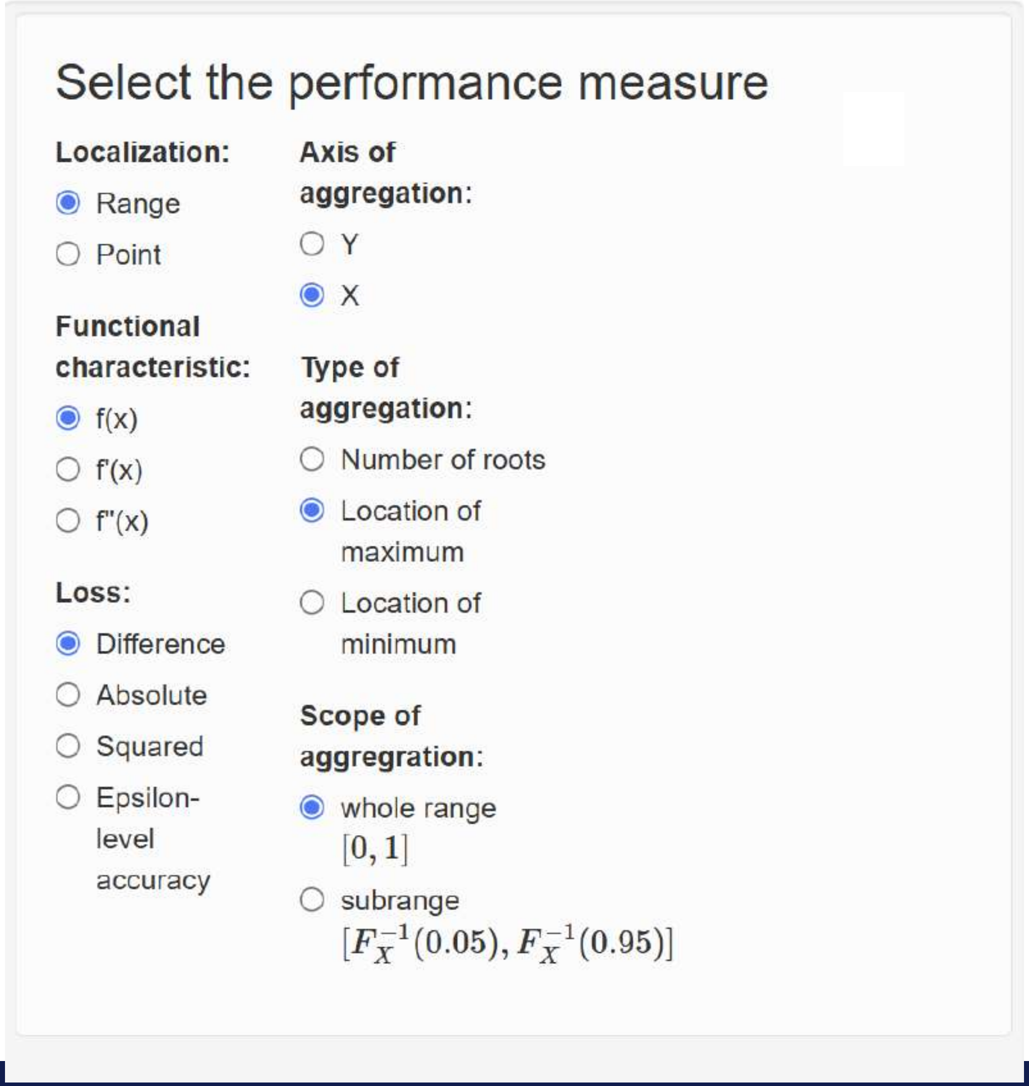
Scope of aggregration:

- ( ) whole range $[0, 1]$
- (●) subrange $[F_X^{-1}(0.05), F_X^{-1}(0.95)]$

$$=\int_{Q05}^{Q95}\left(\hat{f}''(x)-f''(x)\right)^2 dx$$

„wiggliness“

Combining these aspects



Select the performance measure

Localization:
- (●) Range
- ( ) Point

Functional characteristic:
- (●) f(x)
- ( ) f'(x)
- ( ) f"(x)

Loss:
- (●) Difference
- ( ) Absolute
- ( ) Squared
- ( ) Epsilon-level accuracy

Axis of aggregation:
- ( ) Y
- (●) X

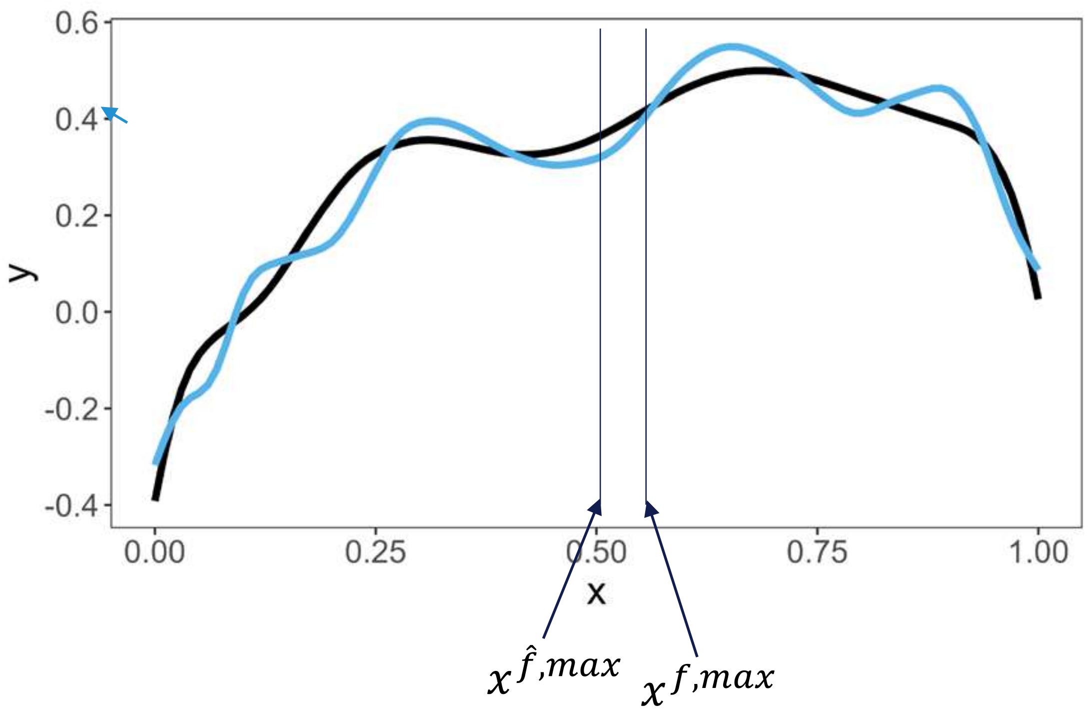
Type of aggregation:
- ( ) Number of roots
- (●) Location of maximum
- ( ) Location of minimum

Scope of aggregration:
- (●) whole range $[0, 1]$
- ( ) subrange $[F_X^{-1}(0.05), F_X^{-1}(0.95)]$

„Deviation of location of maximum“:

$$x^{\hat{f},max} - x^{f,max}$$

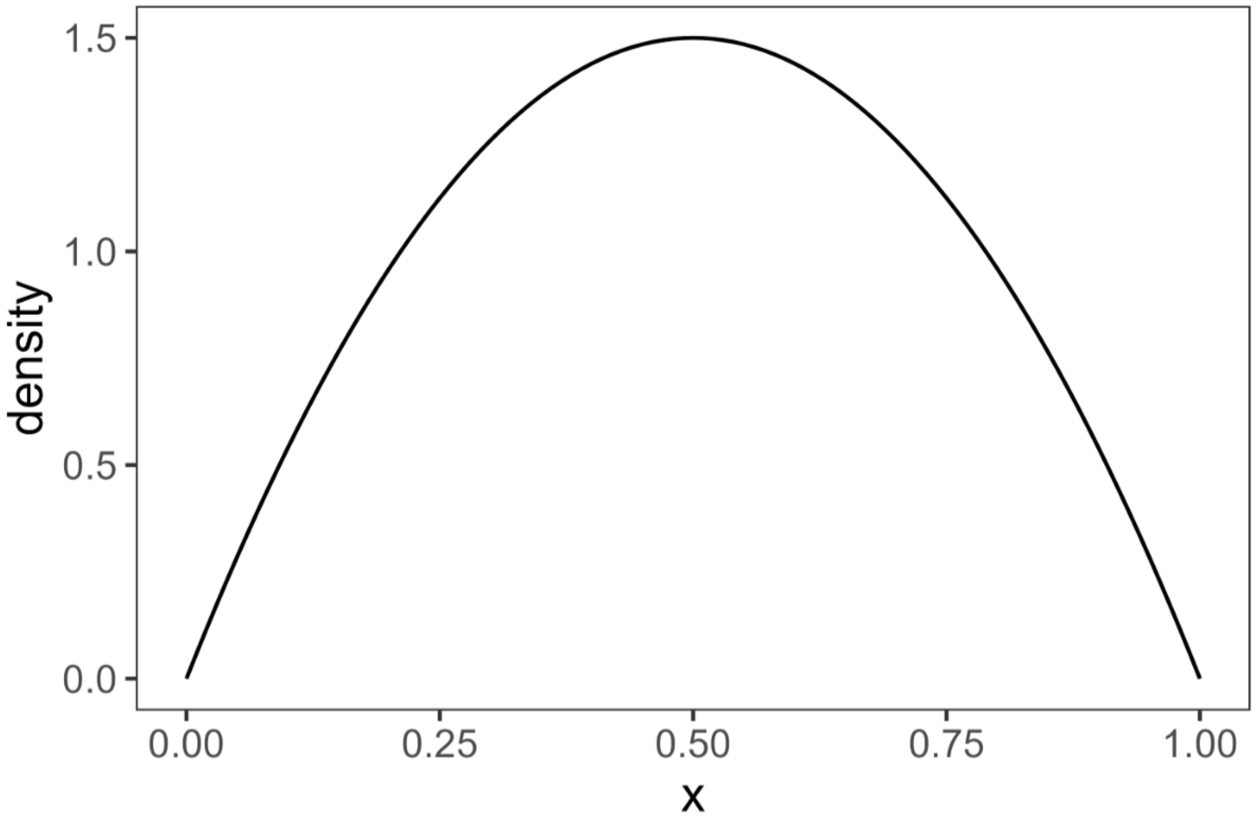
true curve f estimate $\hat{f}$

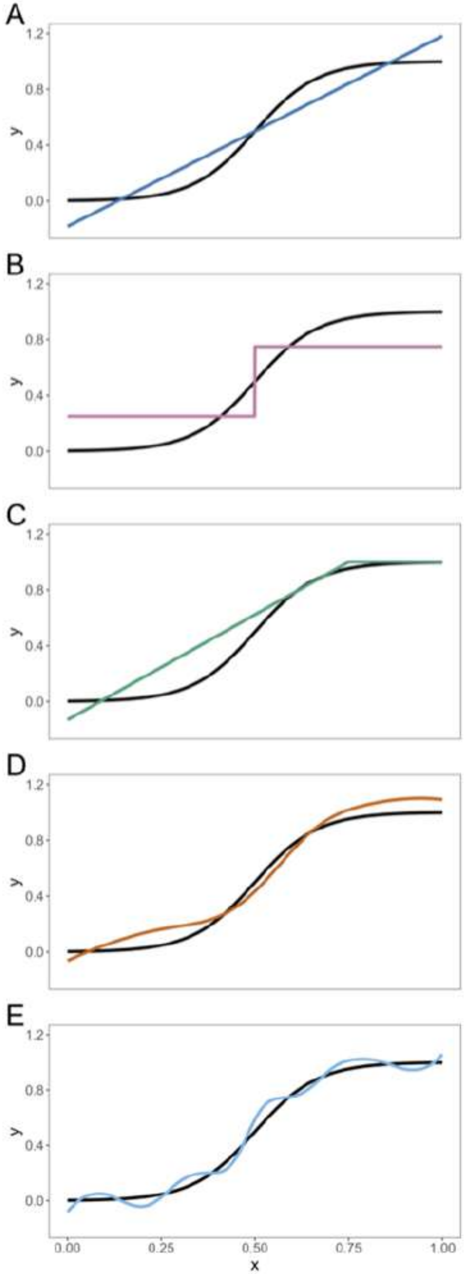


Some examples

- In these examples, we consider x distributed as Beta(2,2)

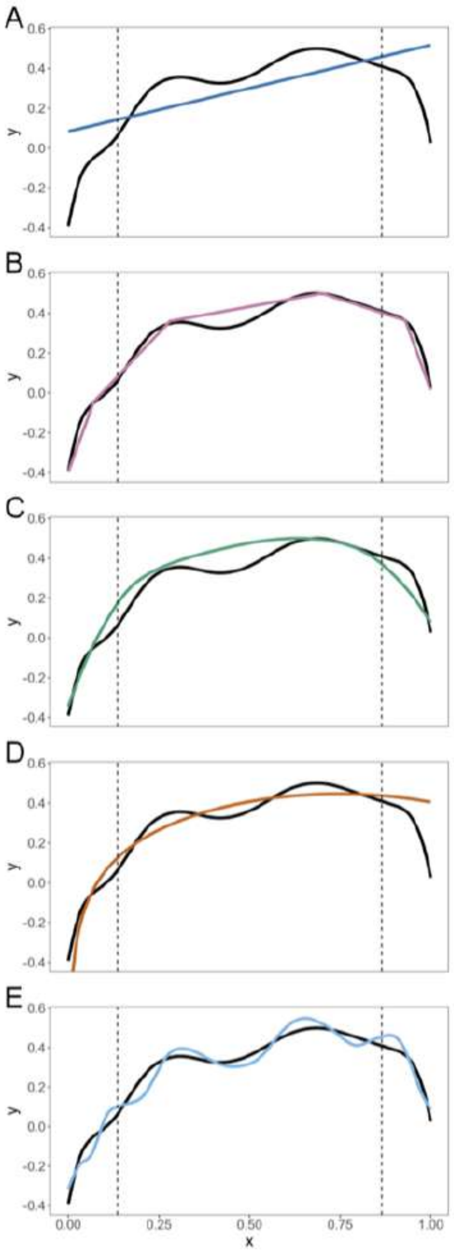
- In some examples, we will nevertheless perform the integration over dx
- In others
  we will integrate over
  dF(x)



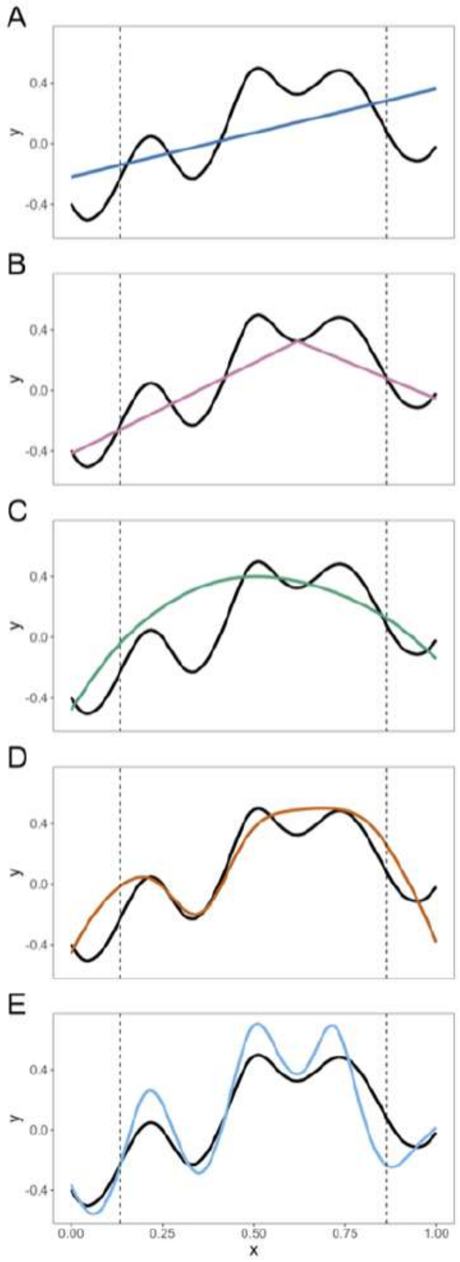


Estimate	Rank according to performance measure...		
	$\int_{\mathcal{X}} \lvert\hat{f}(x) - f(x)\rvert \mathrm{d}x$	$\int_{\mathcal{X}} \lvert\hat{f}'(x) - f'(x)\rvert \mathrm{d}x$	$\int_{\mathcal{X}} \lvert\hat{f}''(x) - f''(x)\rvert \mathrm{d}x$
A	4 (0.10)	3 (0.99)	3 (5.94)
B	5 (0.18)	4 (0.10)	3 (5.94)
C	3 (0.09)	2 (0.75)	3 (5.94)
D	2 (0.07)	1 (0.56)	1 (5.31)
E	1 (0.04)	5 (1.15)	5 (31.64)

Estimate	Rank according to performance measure... $\int_{\mathcal{X}} \lvert\hat{f}''(x) - f''(x)\rvert \mathrm{d}x$	$\int_{F_X^{-1}(0.05)}^{F_X^{-1}(0.95)} \lvert\hat{f}''(x) - f''(x)\rvert \mathrm{d}x$
A	2.5 (25.08)	2.5 (6.98)
B	2.5 (25.08)	2.5 (6.98)
C	1 (23.87)	4 (7.28)
D	5 (∞)	1 (6.58)
E	4 (43.57)	5 (15.36)

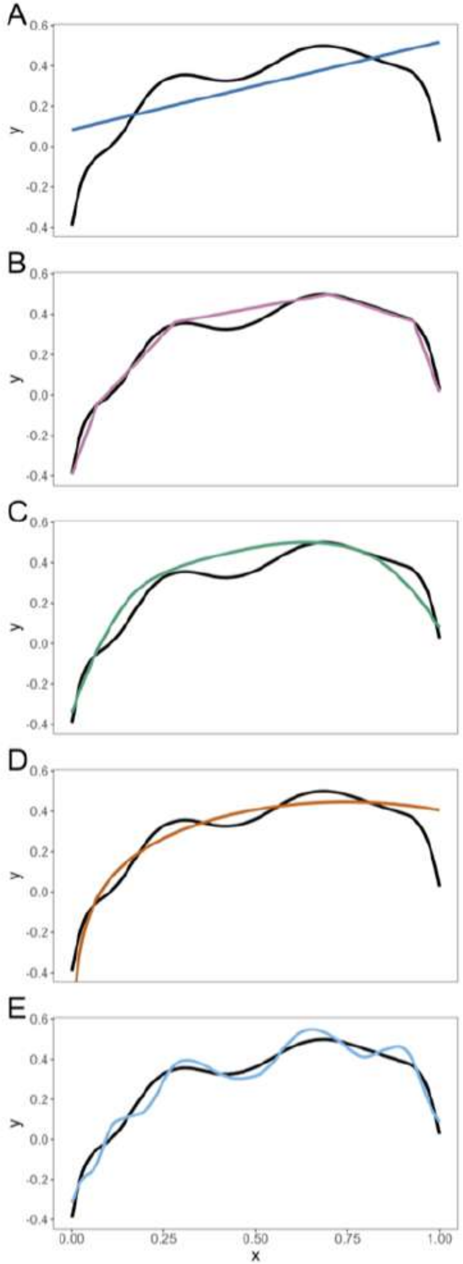


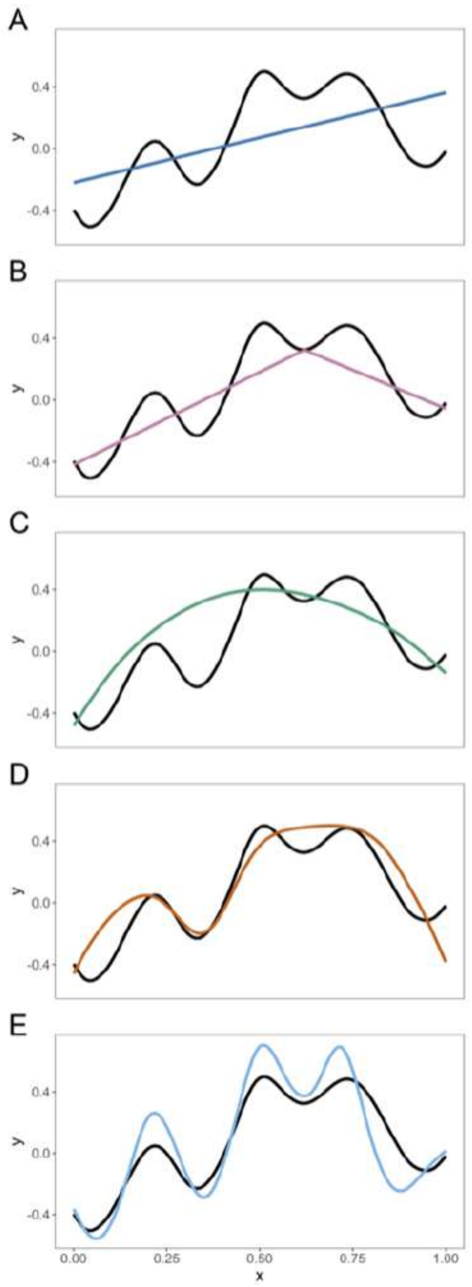
Rank according to performance measure...



Estimate	$\max_{x \in \mathcal{X}} \lvert \hat{f}(x) - f(x) \rvert$	$\max_{x \in [F_X^{-1}(0.05), F_X^{-1}(0.95)]} \lvert \hat{f}(x) - f(x) \rvert$
A	4 (0.45)	4 (0.42)
B	1 (0.31)	2 (0.31)
C	5 (0.54)	5 (0.54)
D	3 (0.37)	1 (0.21)
E	2 (0.35)	3 (0.35)

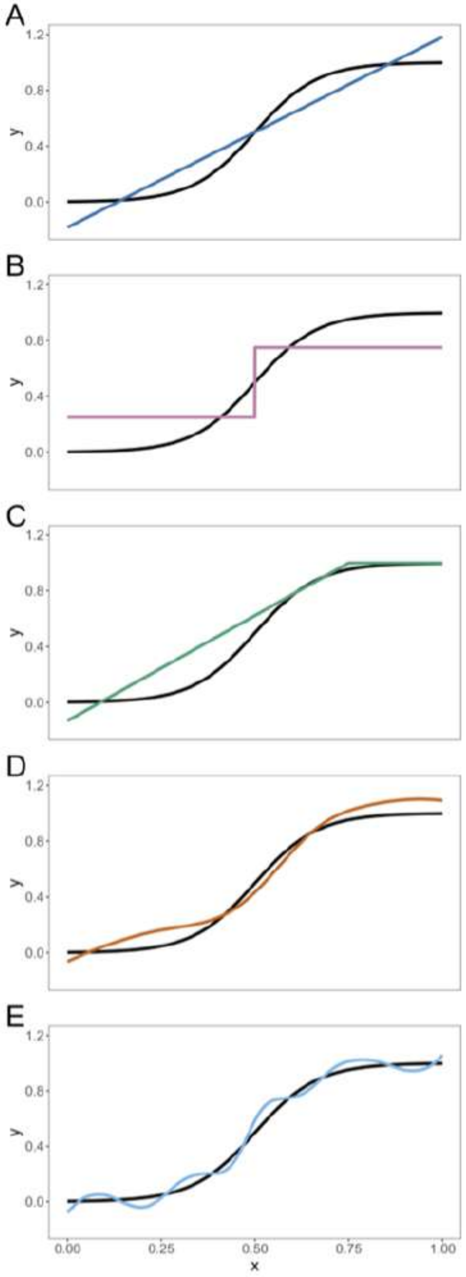
Estimate	Rank according to performance measure...	
	$\int_{\mathcal{X}} \left(\hat{f}(x) - f(x)\right)^2 \mathrm{d}x$	$\int_{\mathcal{X}} \left(\hat{f}(x) - f(x)\right)^2 \mathrm{d}F_X(x)$
A	5 (0.019)	5 (0.010)
B	1 (0.002)	2 (0.002)
C	3 (0.005)	4 (0.005)
D	4 (0.006)	3 (0.002)
E	2 (0.002)	1 (0.002)



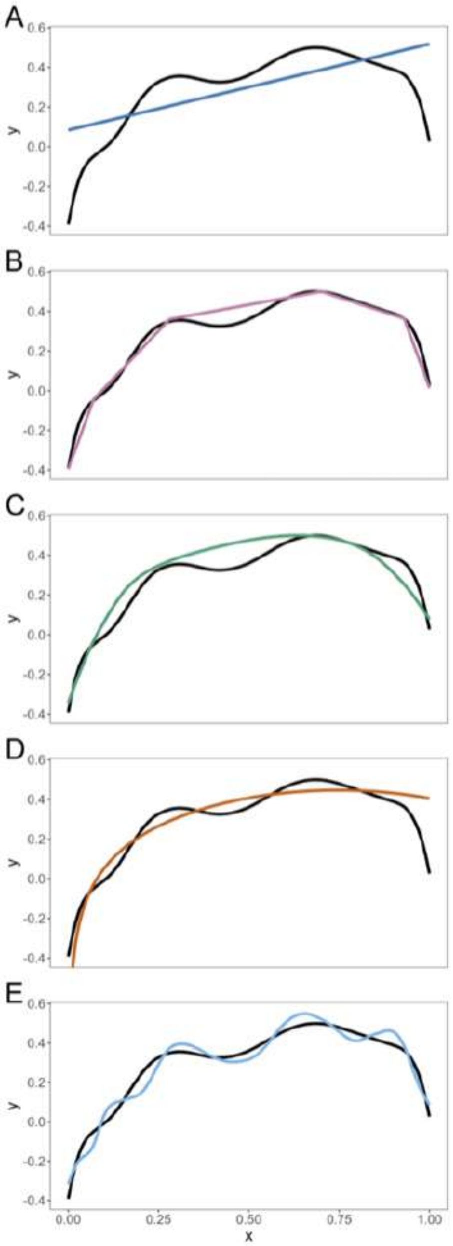


Estimate	Rank according to performance measure... $\int_{\mathcal{X}} \lvert\hat{f}(x) - f(x)\rvert \mathrm{d}x$	$\max_{x\in\mathcal{X}} \lvert\hat{f}(x) - f(x)\rvert$
A	5 (0.22)	4 (0.45)
B	3 (0.14)	1 (0.31)
C	4 (0.16)	5 (0.54)
D	1 (0.10)	3 (0.37)
E	2 (0.12)	2 (0.35)

Estimate	Rank according to performance measure...	
	$\int_{\mathcal{X}} \hat{f}(x) - f(x)\mathrm{d}x$	$\int_{\mathcal{X}} \lvert\hat{f}(x) - f(x)\rvert\mathrm{d}x$
A	1.5 (0)	4 (0.10)
B	1.5 (0)	5 (0.18)
C	5 (0.08)	3 (0.09)
D	4 (0.04)	2 (0.07)
E	3 (0.01)	1 (0.04)



Estimate	Rank according to performance measure...	
	$\int_{\mathcal{X}} \lvert \hat{f}'(x) - f'(x) \rvert \mathrm{d}F_X(x)$	$\int_{\mathcal{X}} \left(\hat{f}'(x) - f'(x) \right)^2 \mathrm{d}F_X(x)$
A	4 (0.81)	3 (1.24)
B	1 (0.46)	1 (0.42)
C	2 (0.59)	2 (0.55)
D	3 (0.66)	5 (∞)
E	5 (0.91)	4 (1.37)



How many measures are there?

According to our categorization, there are…

m = 228
performance measures

If we consider one option for the type of aggregation = "quantile with respect to F_X" (e.g., the median) and two options for the scope of aggregation ($\mathcal{X}$ and 5%-95% quantile of F_X)

m = 216
measures with
localization = "range"

m = 12
measures with
localization = "point"

→How to choose a **smaller set** of performance measures for a simulation study?

→ Select those that capture different features (see examples!)

Aggregation over simulated data sets

- Our performance measures will summarize the quality of the fitted line in 1 simulated data set
- The analyst still has to decide whether
  - Expected value of the performance measure
  - Variance of the performance measure
  - or other population quantity is of interest (e.g., median, p^{th} quantile etc.)
- If there is a clear optimum value (e.g. expected difference [=bias] should be 0), one could also construct a combination of bias + variance
  - Obvious: MSE = $bias^2$ + variance

Applications

- Univariate models: unadjusted association
- Models where the association of interest is adjusted for a (fixed) set of adjustment variables (descriptive-associational)
- Evaluation over a two-dimensional grid on X_1, X_2
- Prediction/calibration:
    - agreement of predicted and observed values
    - agreement of predicted and true linear predictor values
- Extensions: comparison to ‚null' instead of true $f(x)$
    - Number of roots
    - General wiggliness

Preprint is available on Arxiv

Ullmann, T., Heinze, G., Abrahamowicz, M., Perperoglou, A., Sauerbrei, W., Schmid, M., Dunkler, D., for TG2 of the Stratos initiative. (2025). A categorization of performance measures for estimated non-linear associations between an outcome and continuous predictors (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2503.16981

References

Binder, H., Sauerbrei, W., & Royston, P. (2011). Multivariable model-building with continuous covariates: 1. Performance measures and simulation design. Technical report.

Buchholz, A., Sauerbrei, W., & Royston, P. (2014). A measure for assessing functions of time-varying effects in survival analysis. Open Journal of Statistics, 4(11), 977998

Govindarajulu, U. S., Spiegelman, D., Thurston, S. W., Ganguli, B., & Eisen, E. A. (2007). Comparing smoothing techniques in Cox models for exposure–response relationships. Statistics in Medicine, 26(20), 3735–3752.

Morris, T.P., White, I.R., Crowther, M.J., 2019. Using simulation studies to evaluate statistical methods. Statistics in Medicine 38, 2074–2102. https://doi.org/10.1002/sim.8086