Comparing methods for flexible estimation of non-linear associations: the importance of suitable performance measures

Theresa Ullmann

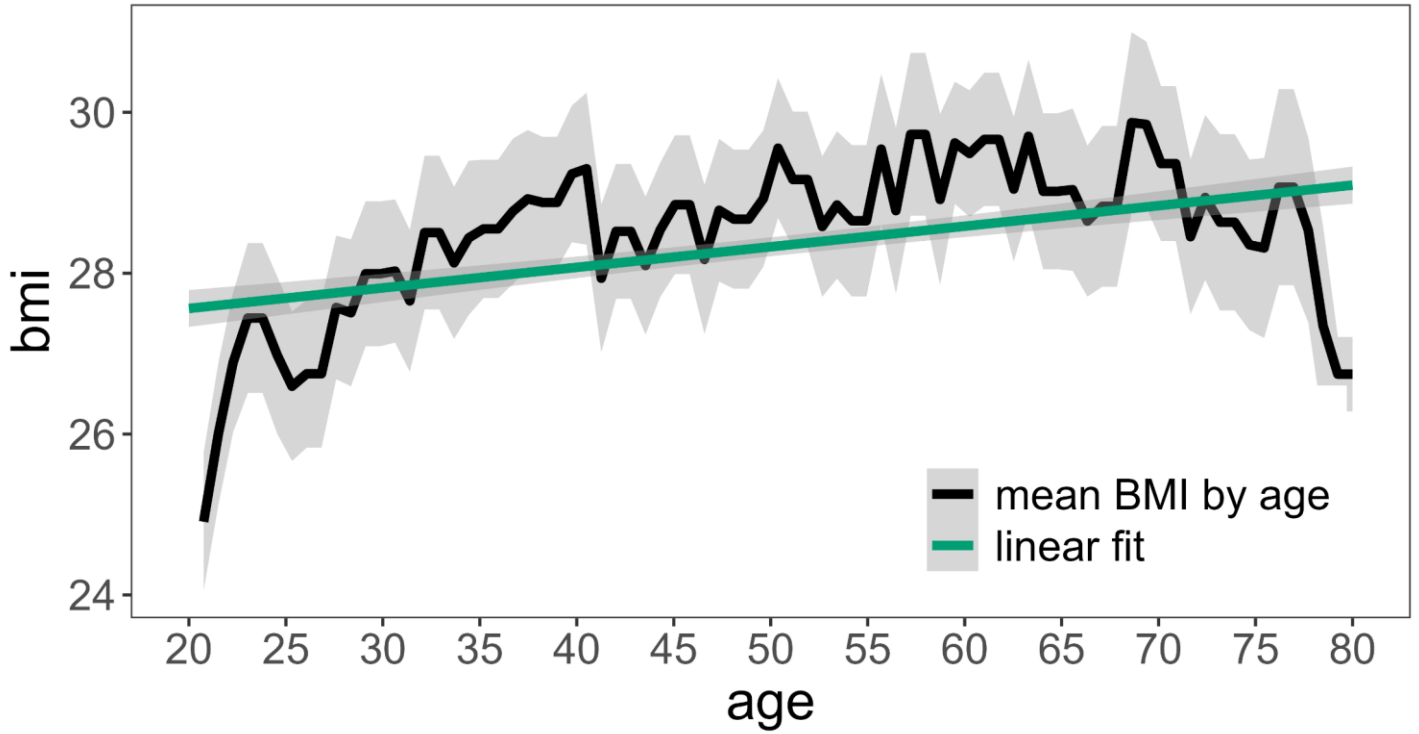
on behalf of TG2 of the STRATOS initiative

Institute of Clinical Biometrics, Center for Medical Data Science,

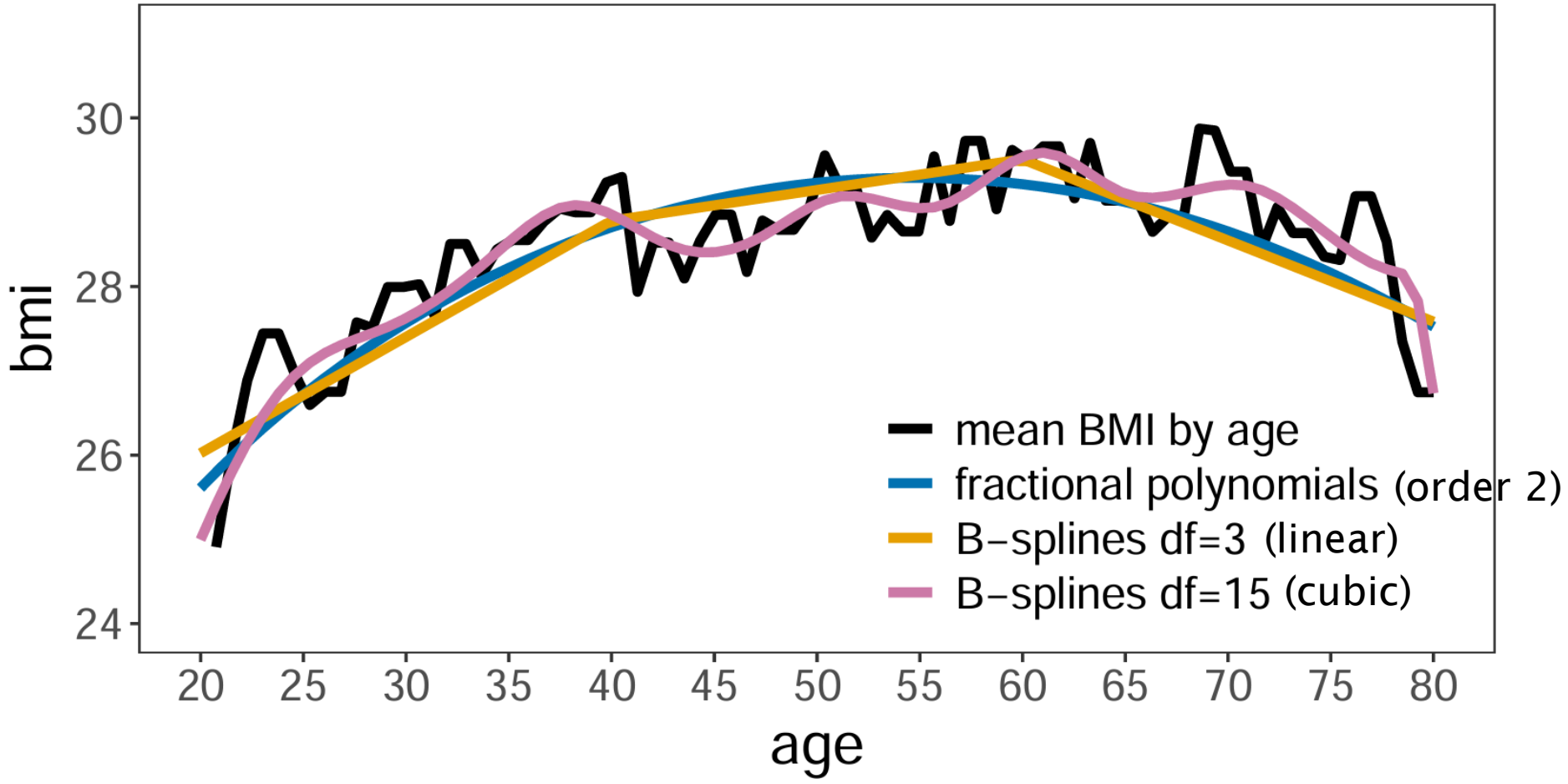
Medical University of Vienna

Background & motivation

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



More smoothers...



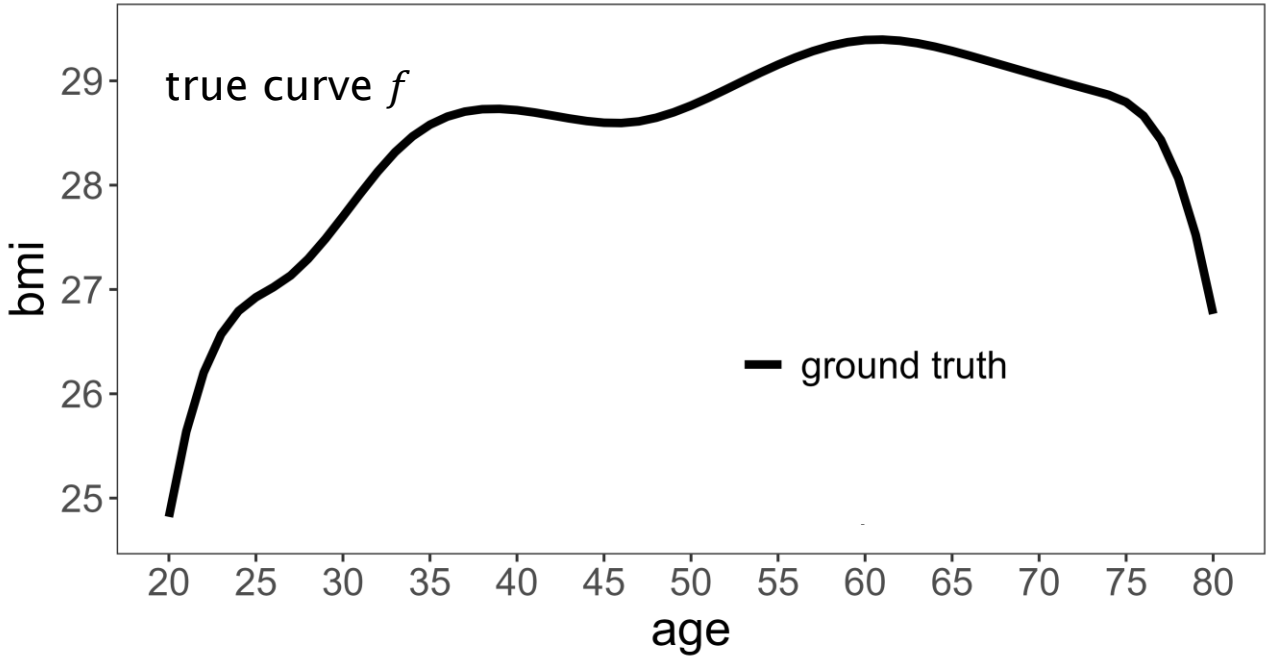
Which method(s) should we use?

- Many methods available for estimating non-linear associations (fractional polynomials, different types of splines,…)
- Which method(s) should we use in which situation?
- Simulation studies which systematically compare methods can help to provide guidance

Suppose we want to perform such a simulation study…

A simulation study to compare methods?

- Aim: to compare the performance of different methods for non-linear modeling
- Data-generating mechanism: sample values x_i from F_X (here X = age), $y_i = f(x_i) + \varepsilon_i$

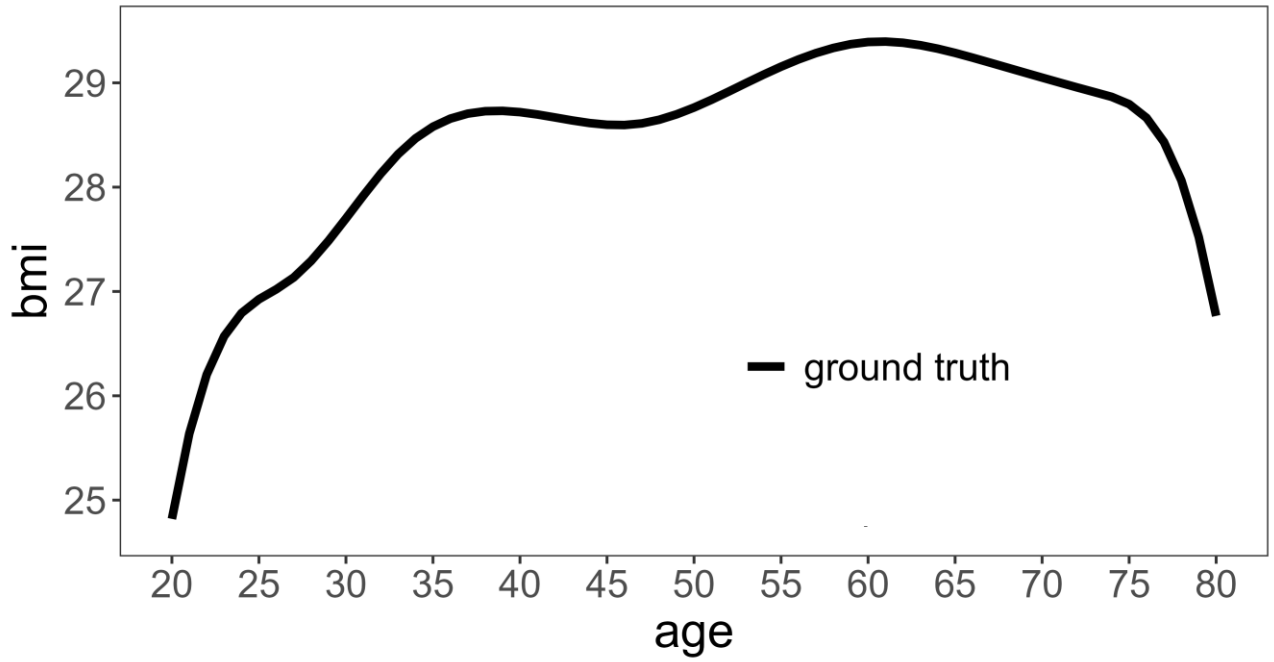


- Estimand: ground truth curve

ADEMP: Morris et al, StatMed 2019

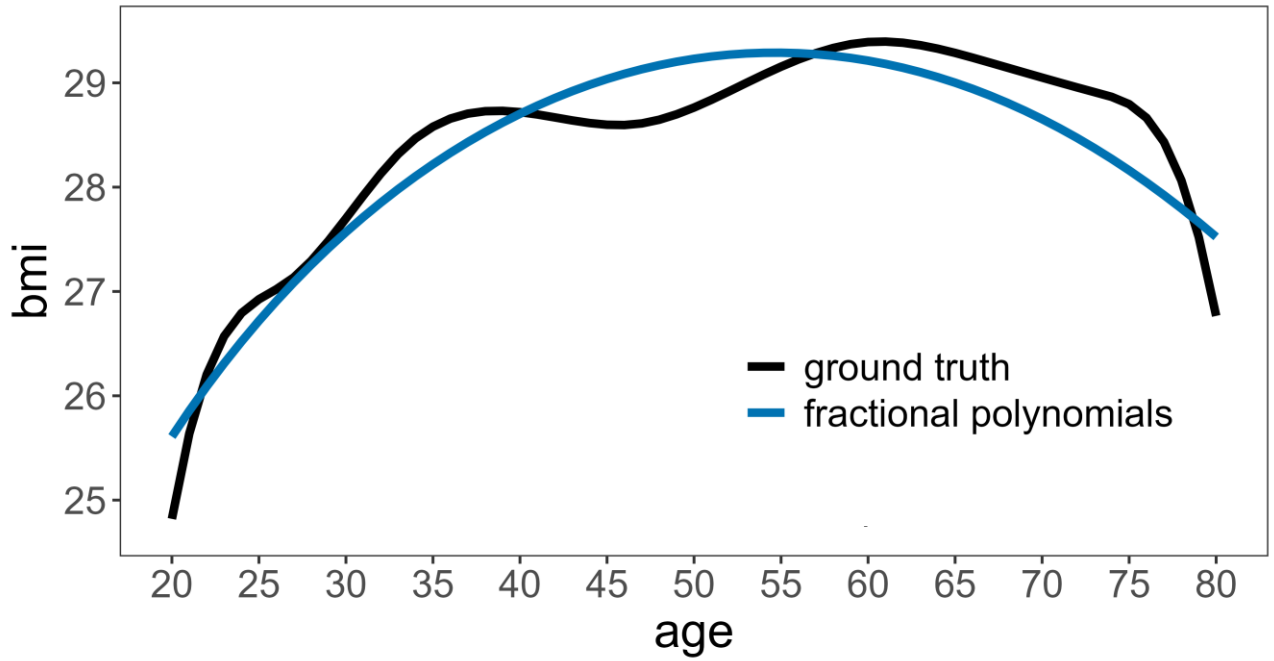
A simulation study to compare methods?

- Methods:



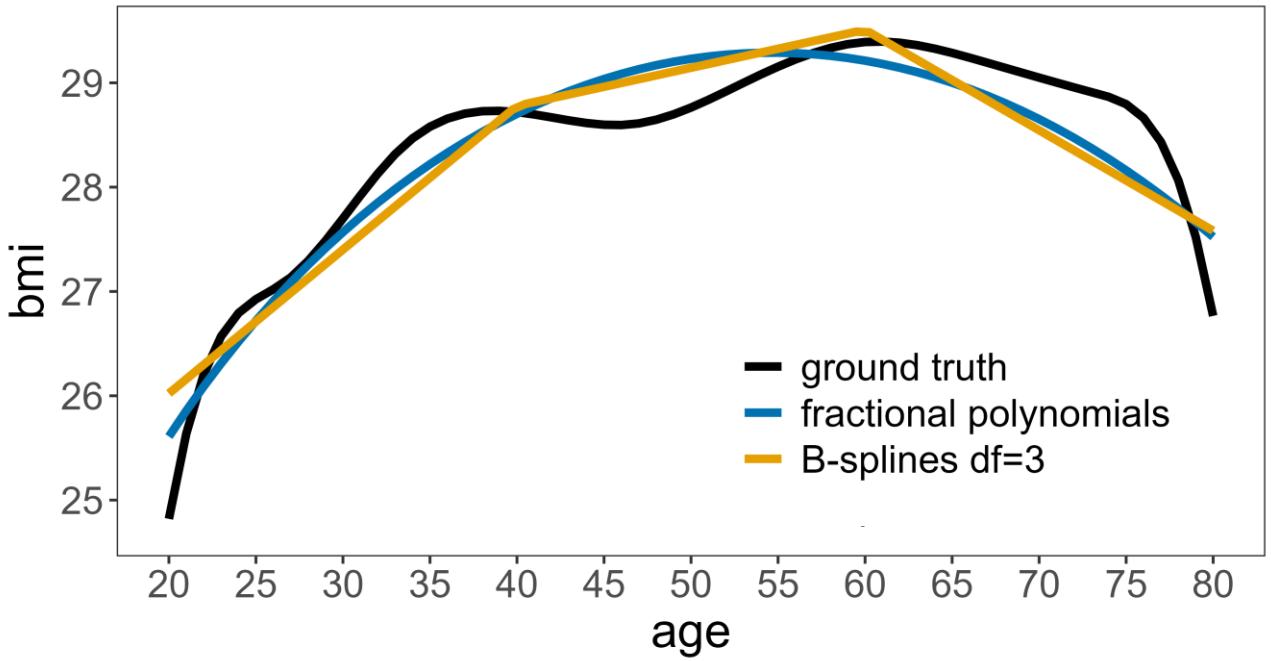
A simulation study to compare methods?

- Methods:



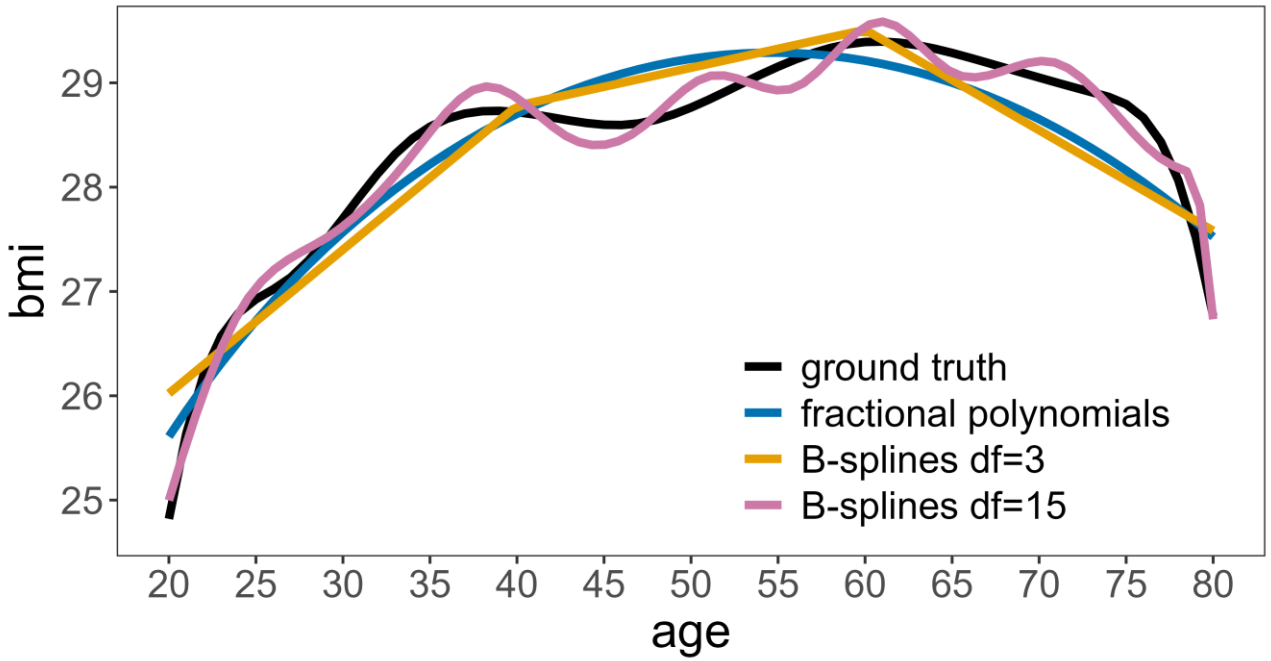
A simulation study to compare methods?

- Methods:



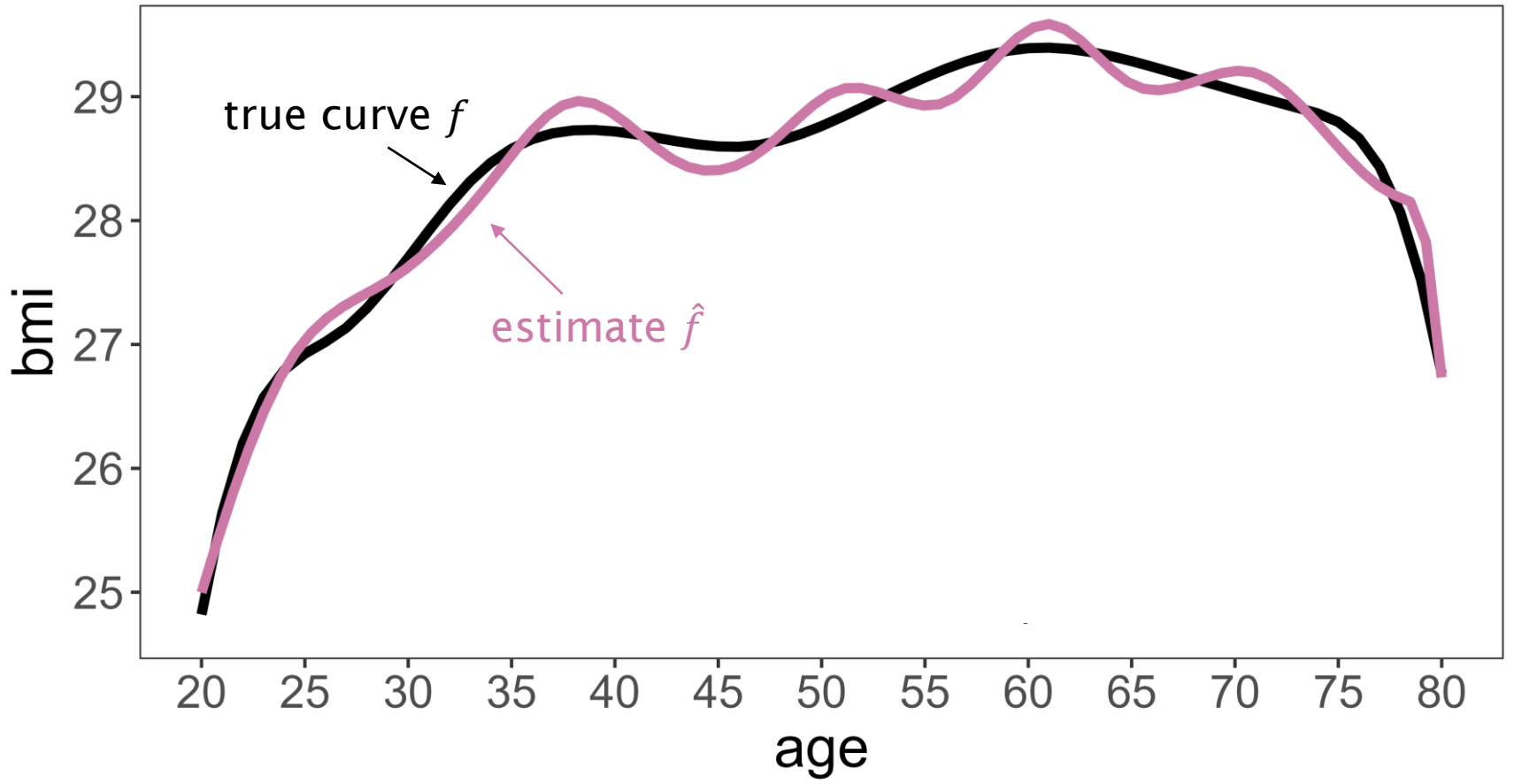
A simulation study to compare methods?

- 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)

$$\int_{F_X^{-1}(0.05)}^{F_X^{-1}(0.95)} \left(\hat{f}'(x) - f'(x)\right)^2 \mathrm{d}F_X(x)$$ Binder et al. (2011)

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

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?

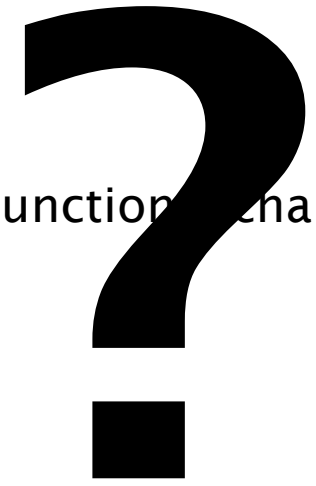
- Performance measures:

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

A simulation study to compare methods?

- Performance measures:

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



Goals of our 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…
  - … how the resulting performance measures behave
  - … that different performance measures capture different aspects of behavior

WIREs COMPUTATIONAL STATISTICS

ADVANCED REVIEW | Open Access

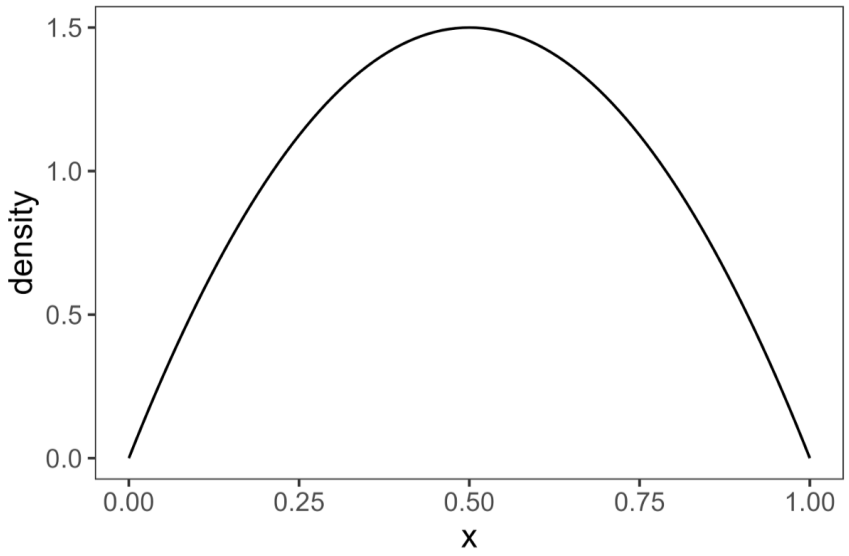
A Systematic Categorization of Performance Measures for Estimated Non-Linear Associations Between an Outcome and Continuous Predictors

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

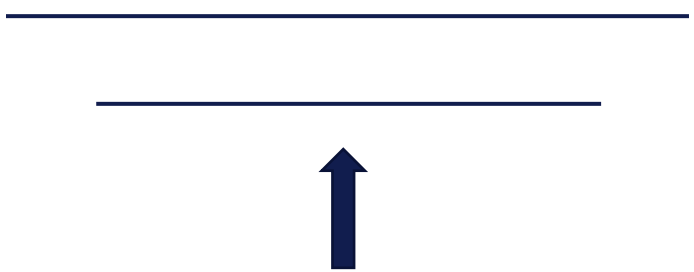
First published: 31 August 2025 | https://doi.org/10.1002/wics.70042

The aspects of the categorization:

Distribution of X

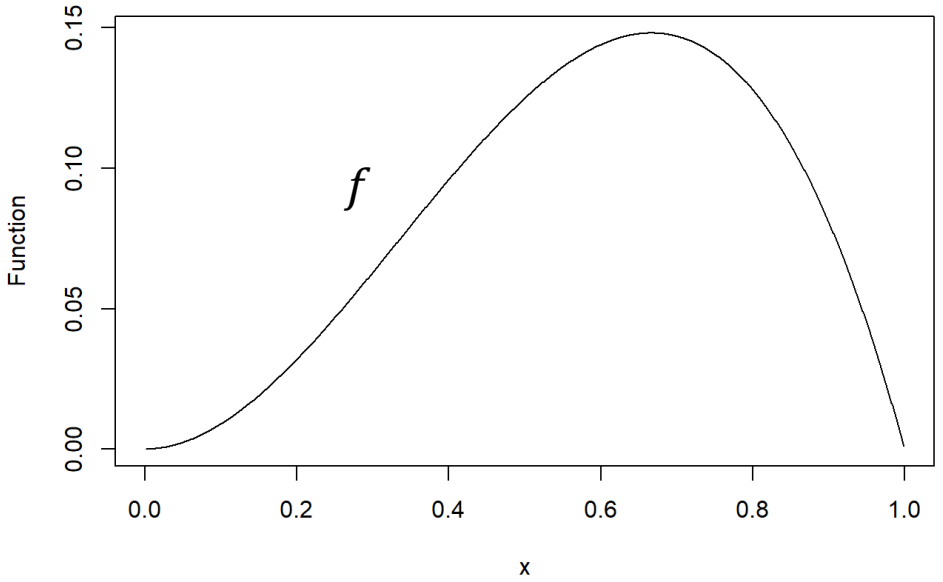


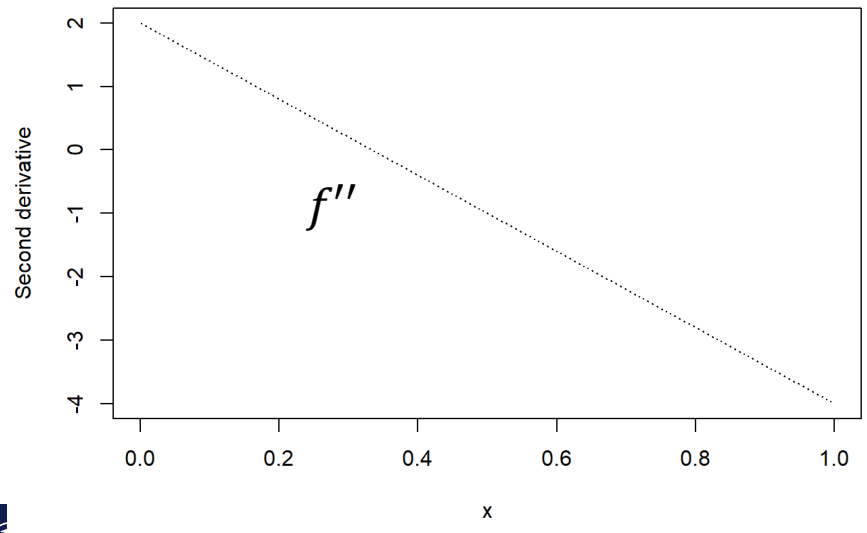
- Localization: Where are we looking at?
  - The whole range of values (global)
  - A subrange (region)
  - A single value (point)

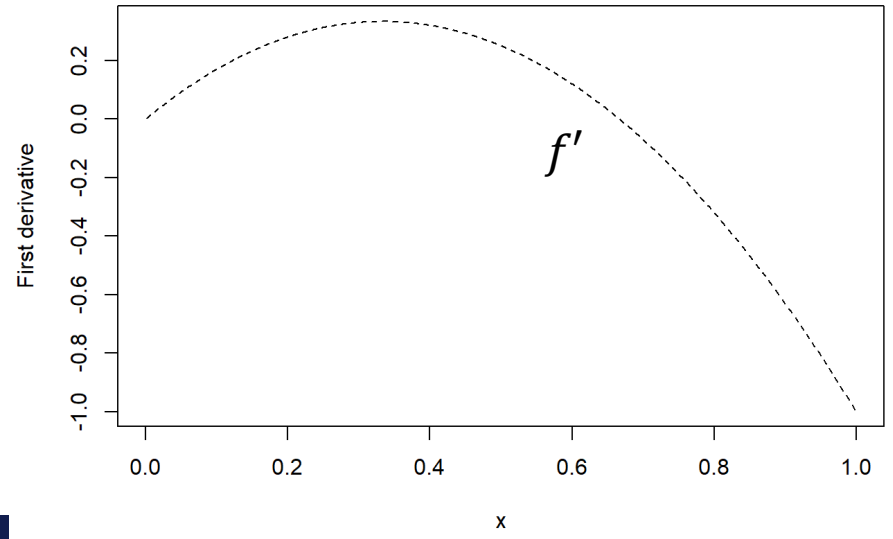


The aspects of the categorization:

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





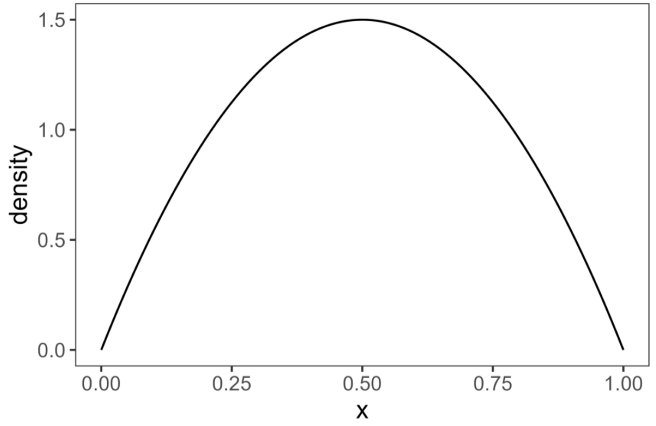


The aspects of the categorization:

- 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: $m(x) = \mathbb{I}\left(\hat{f}(x) - f(x) \leq \epsilon\right)$

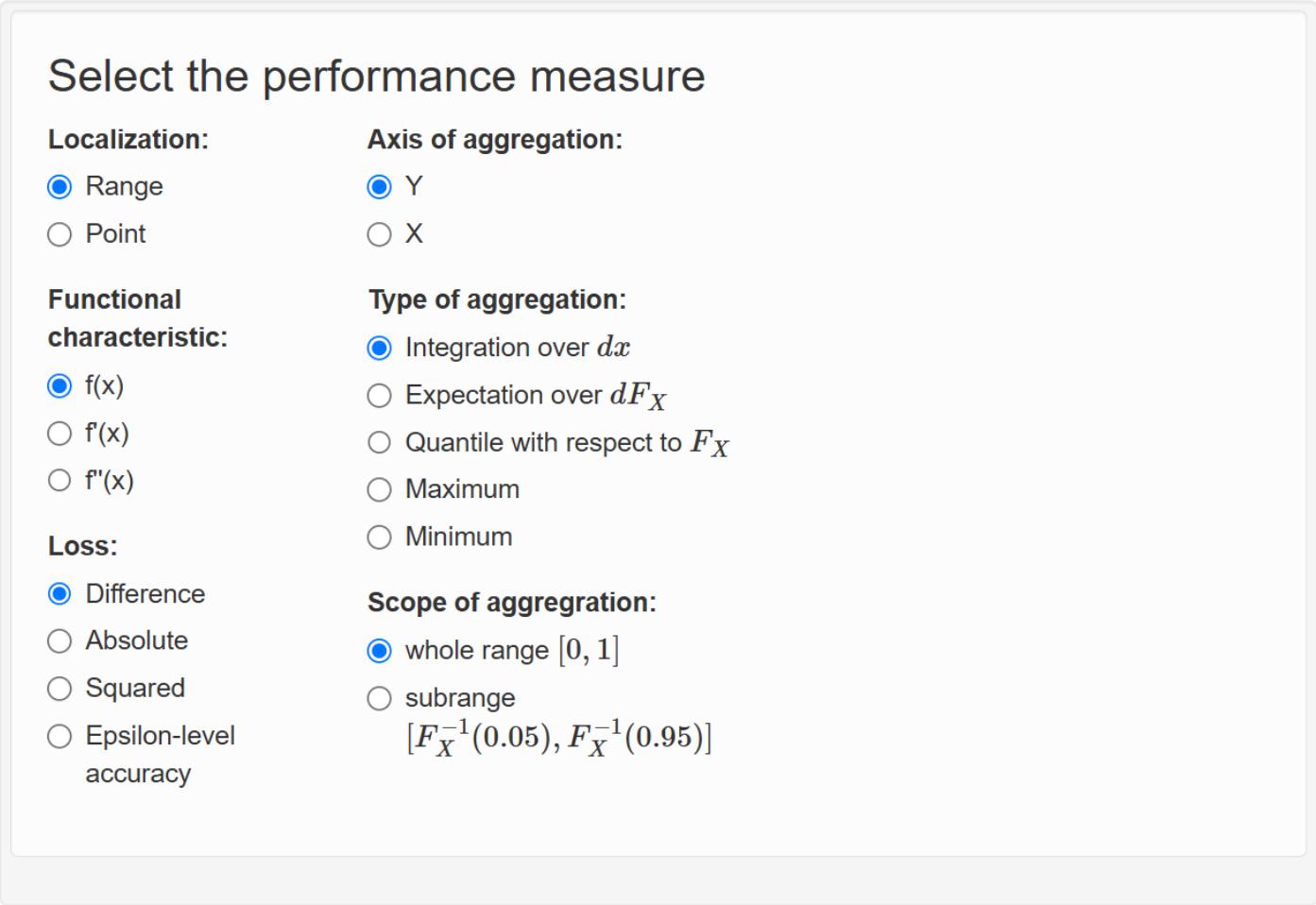
If we consider a range:

Distribution of X



- Axis of aggregation:
  - Y
    - Integration over dx: $\int m(x)\, dx$
    - Integration over dF_X (=expected value): $\int m(x)\, dF_X(x)$
    - Maximum $\max_x m(x)$
    - Minimum $\min_x m(x)$
    - …
  - X
    - Location of maximum/minimum
    - Number of roots

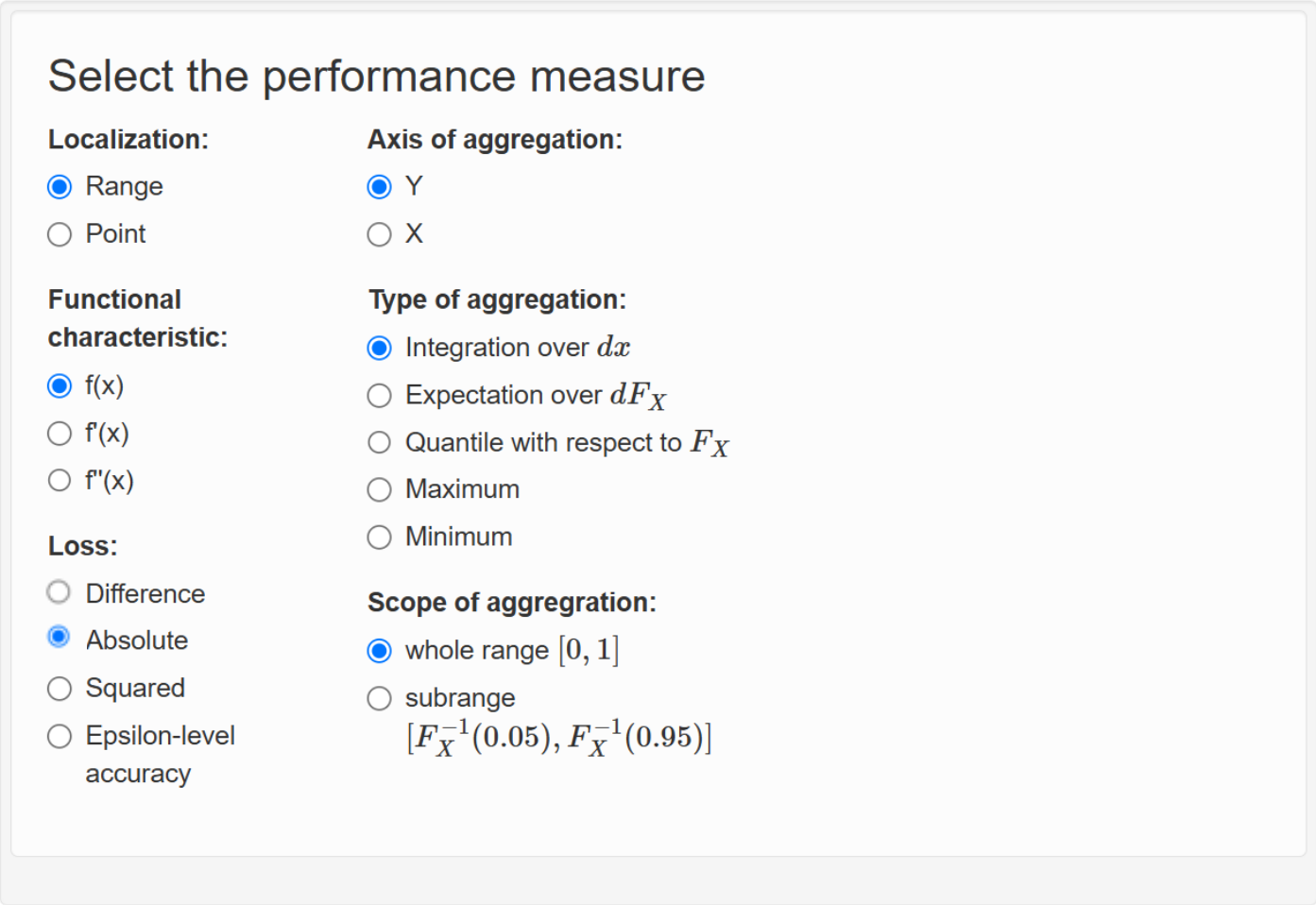
Combining these aspects



$$= \int_0^1 \left(\hat{f}(x) - f(x) \right) dx$$

"mean deviation"

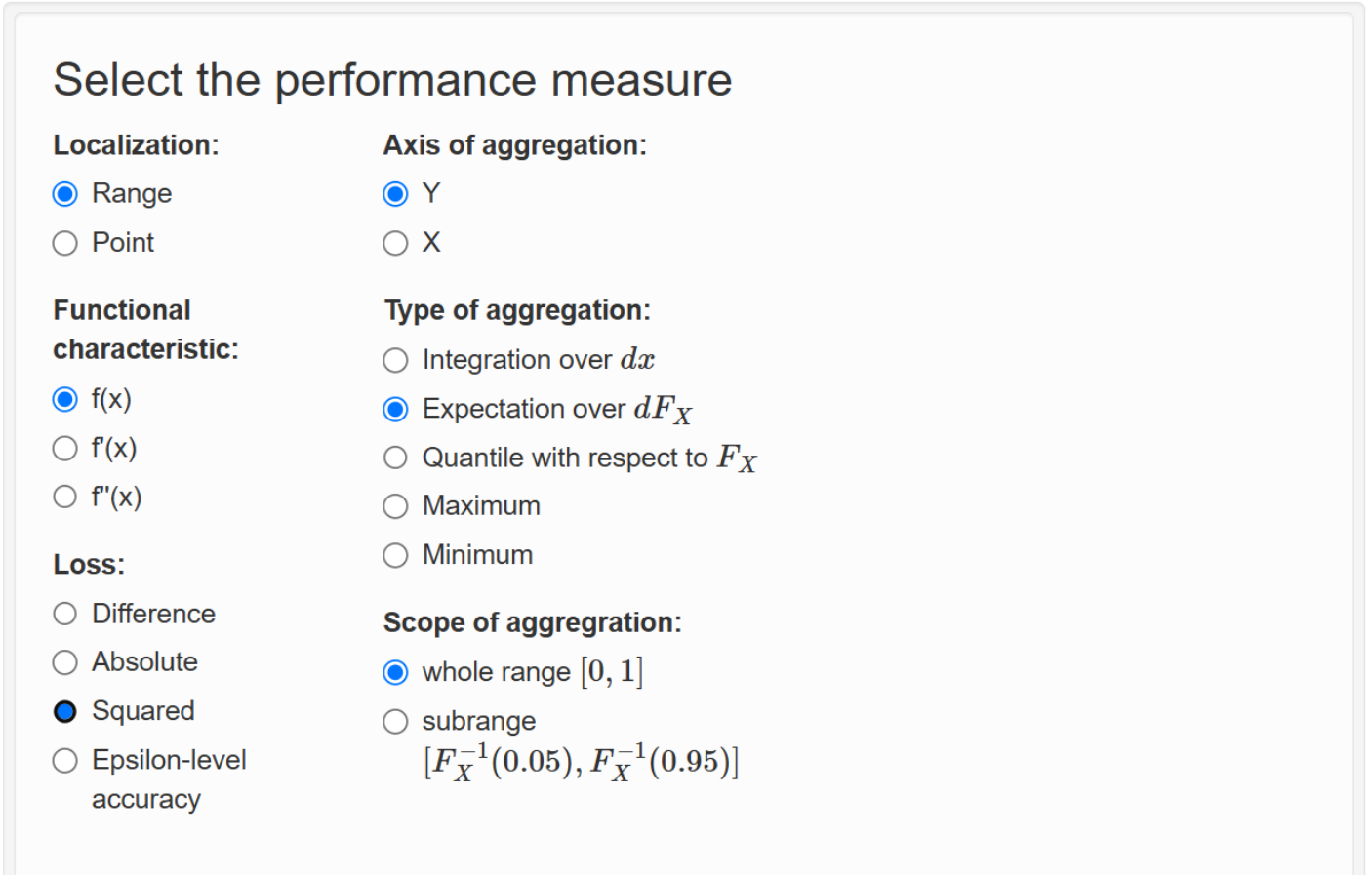
Combining these aspects



$$= \int_0^1 \left|\hat{f}(x) - f(x)\right| dx$$

"mean absolute deviation"

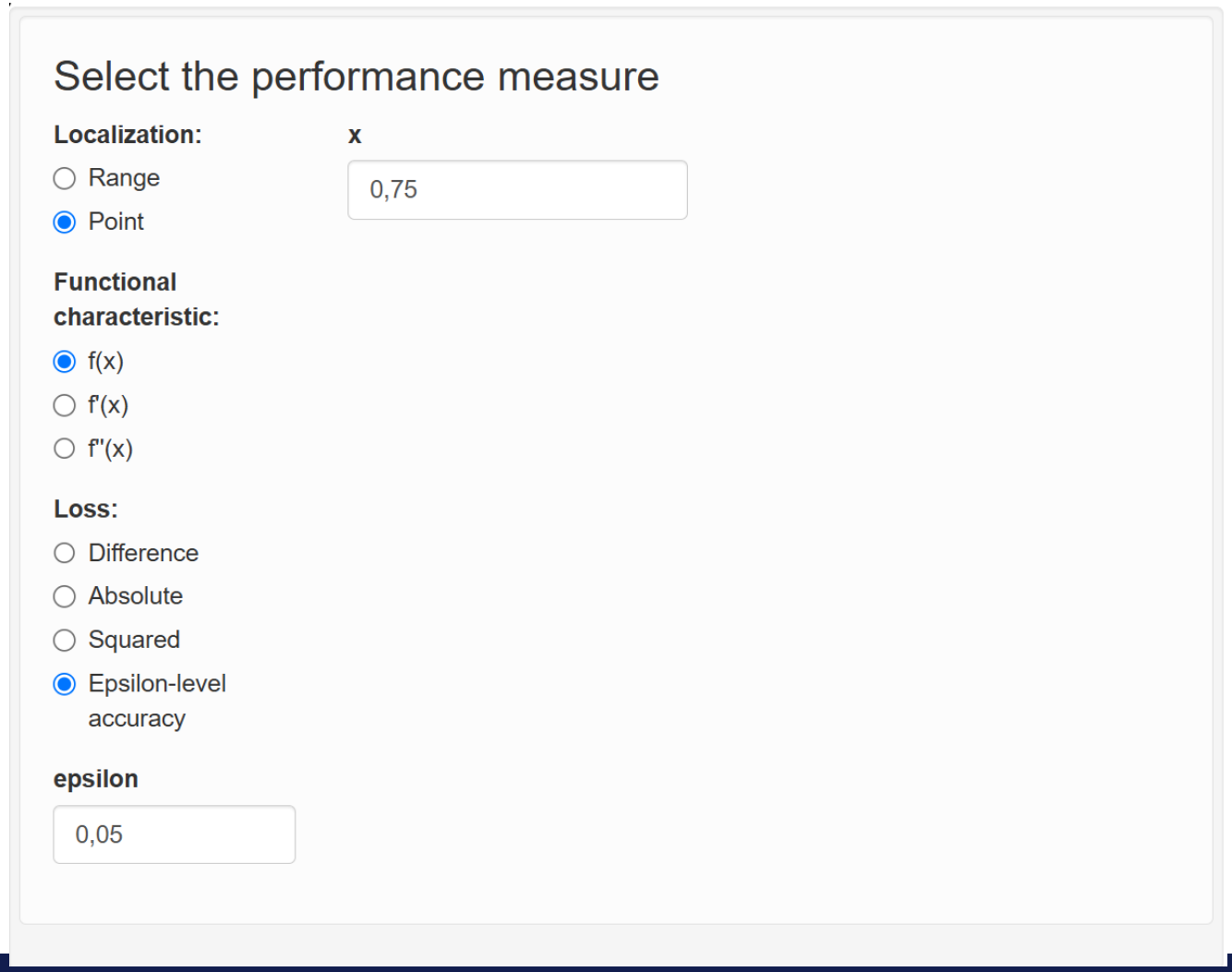
Combining these aspects



$$= \int_0^1 \left(\hat{f}(x) - f(x)\right)^2 dF_X(x)$$

"expected (over F_X) squared deviation"

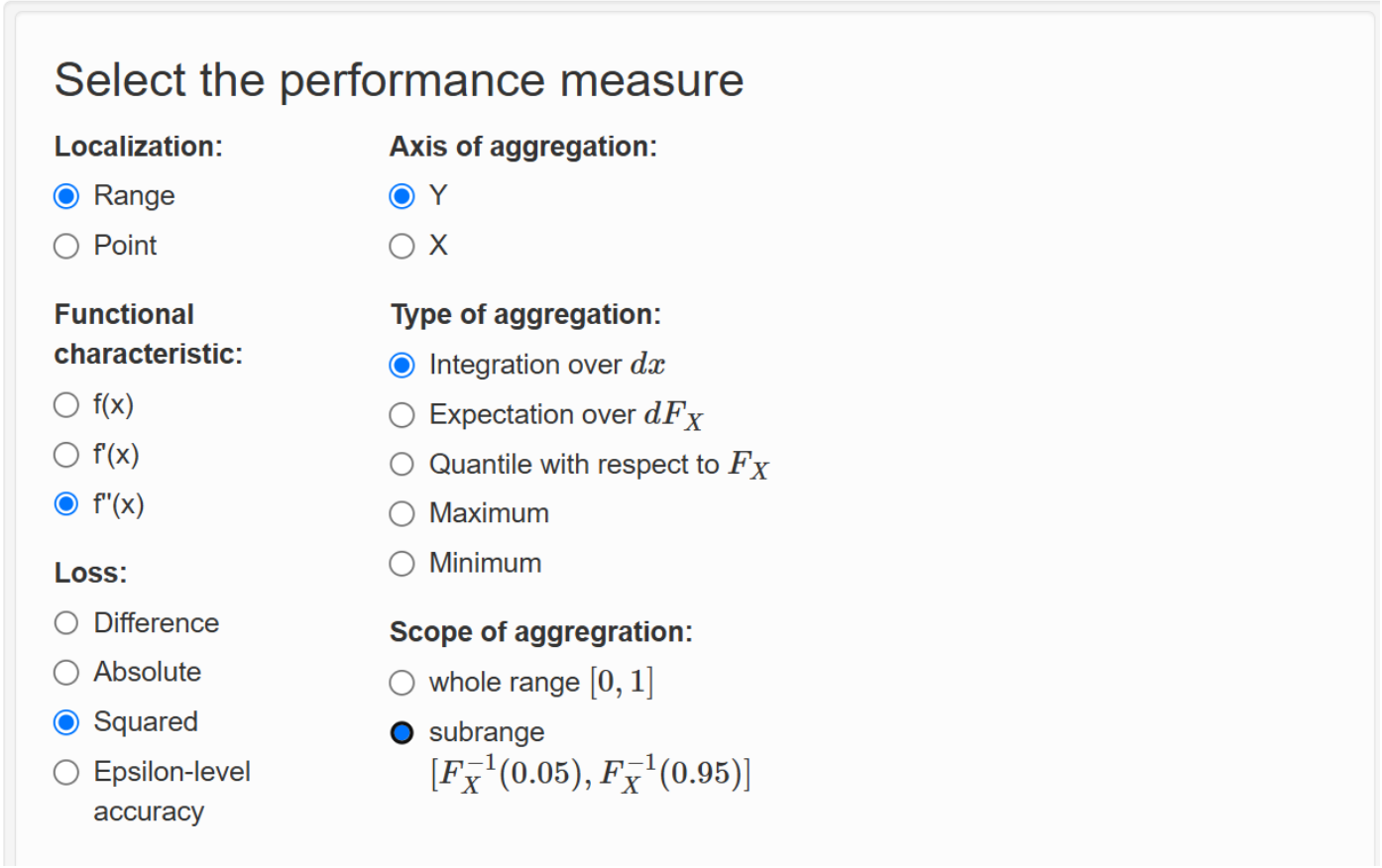
Combining these aspects



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

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

Combining these aspects



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

"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

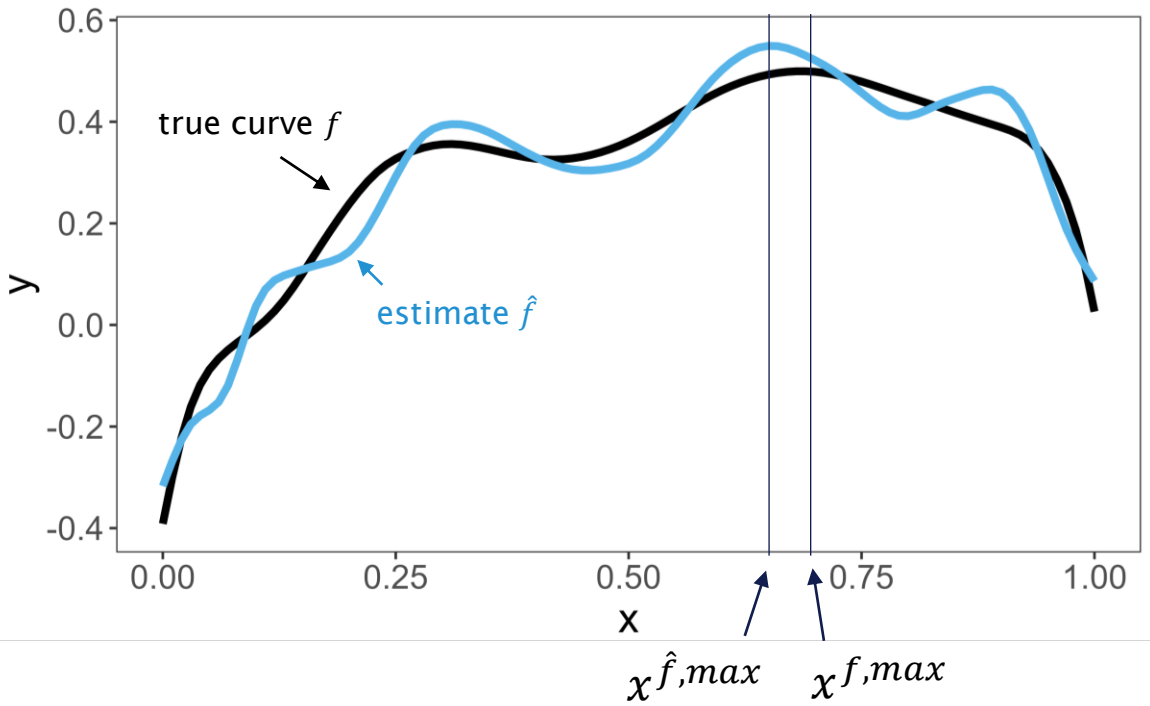
Axis of aggregation:
- ( ) Y
- (●) X

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

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

"Deviation of location of (global) maximum":

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



Does a single measure suffice?

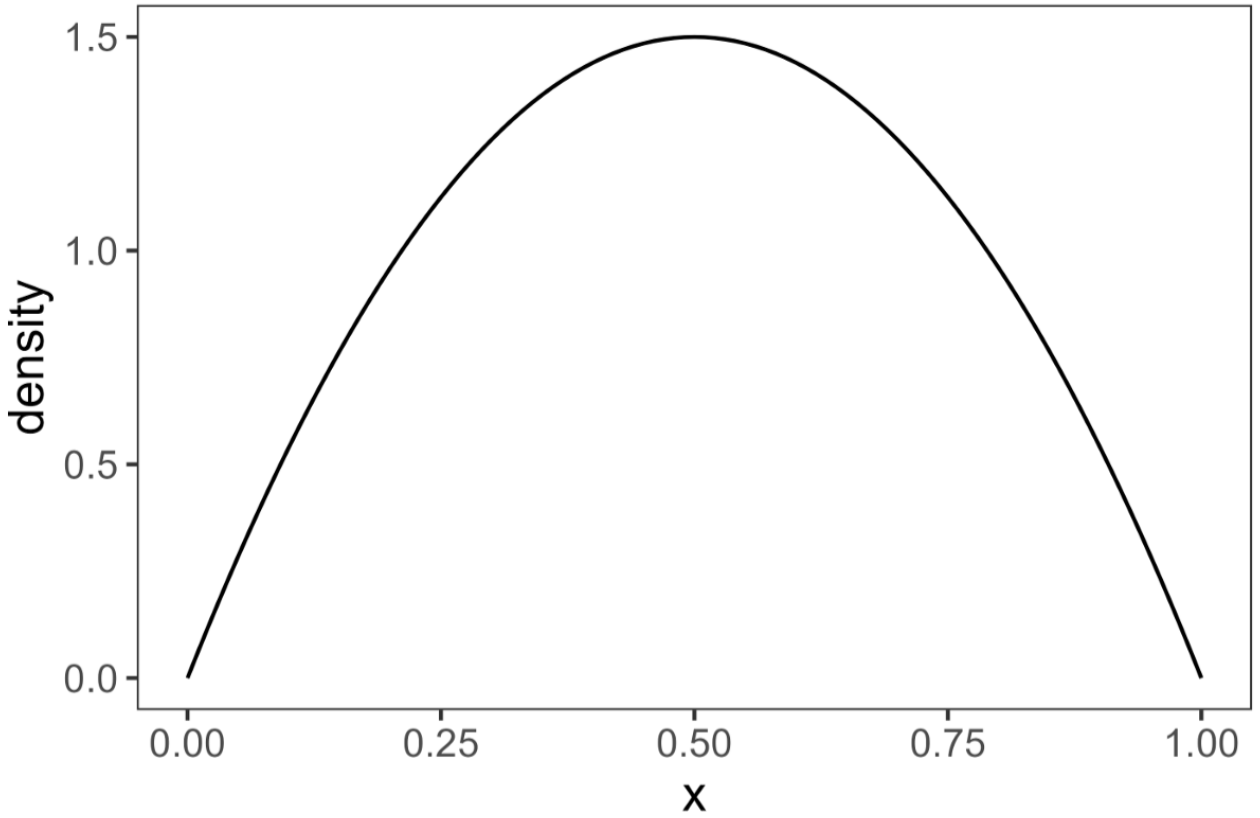
Q: Why should we consider *multiple* performance measures?

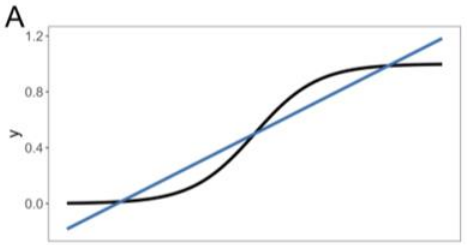
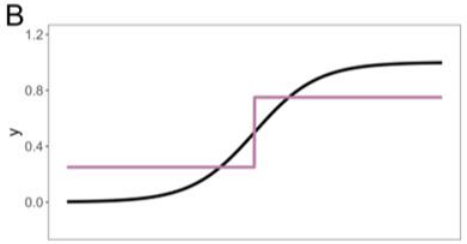
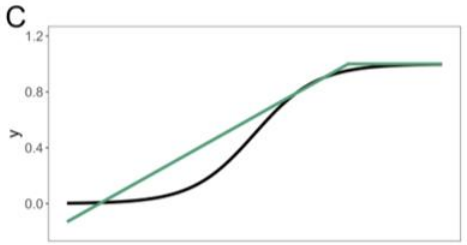
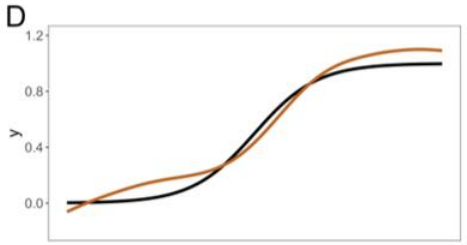
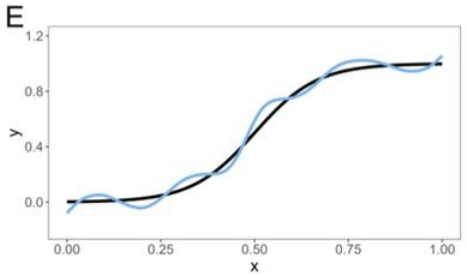
A: Different measures capture distinct properties of estimated curves, and may therefore rank estimates in different ways!

Some examples

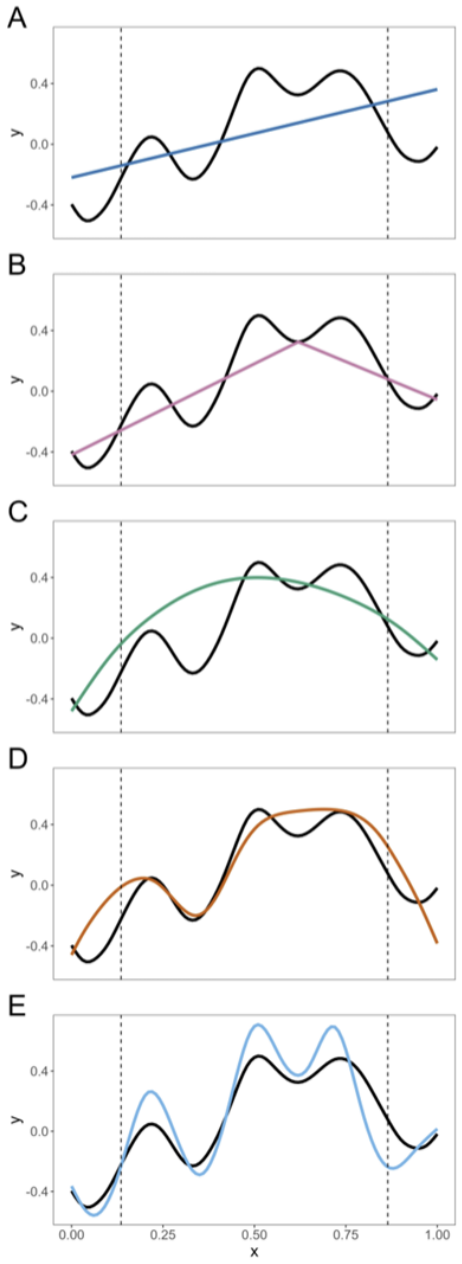
- In these examples, we consider X distributed as Beta(2,2)

- 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...		
	$\int_{\mathcal{X}} \lvert\hat{f}(x) - f(x)\rvert \mathrm{d}x$	$\int_{\mathcal{X}} \lvert\hat{f}'(x) - f'(x)\rvert \mathrm{d}x$	$\int_{\mathcal{X}} \lvert\hat{f}''(x) - f''(x)\rvert \mathrm{d}x$
	4 (0.10)	3 (0.99)	3 (5.94)
	5 (0.18)	4 (0.10)	3 (5.94)
	3 (0.09)	2 (0.75)	3 (5.94)
	2 (0.07)	1 (0.56)	1 (5.31)
	1 (0.04)	5 (1.15)	5 (31.64)

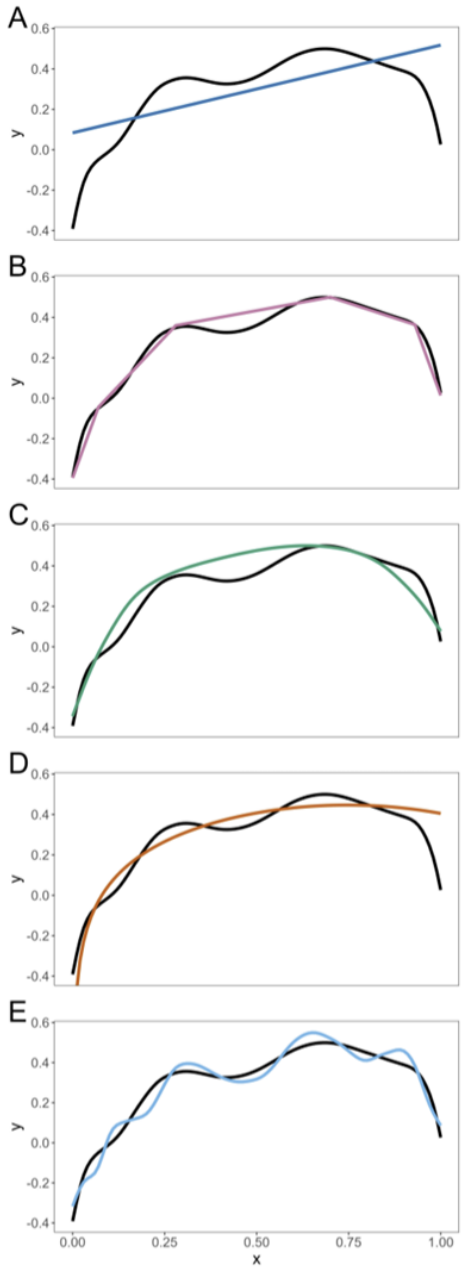
Estimate



Rank according to performance measure...

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

Estimate

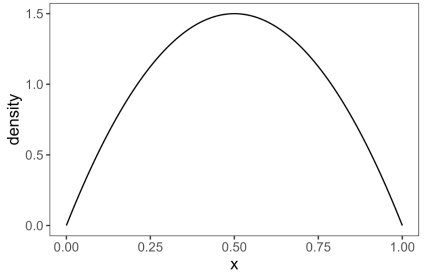


Rank according to performance measure...

$$\int_{\mathcal{X}} \left(\hat{f}(x) - f(x)\right)^2 \mathrm{d}x \qquad \int_{\mathcal{X}} \left(\hat{f}(x) - f(x)\right)^2 \mathrm{d}F_X(x)$$

5 (0.019)	5 (0.010)
1 (0.002)	2 (0.002)
3 (0.005)	4 (0.005)
4 (0.006)	3 (0.002)
2 (0.002)	1 (0.002)

Density of X



Aggregation over simulated data sets

- Our performance measures summarize the quality of the fitted line in 1 simulated data set
- The analyst still has to decide whether…
  - … the expected value of the performance measure,
  - the variance of the performance measure,
  - or another population quantity (e.g., median, p^{th} quantile etc.)

  … is of interest.

Applications and extensions of the categorization

- Univariable models: unadjusted association
- Models where the association of interest is adjusted for a (fixed) set of adjustment variables
- Extension:
  - Multivariable non-linear effects of the form $f(X_1, X_2)$: evaluate performance over a two-dimensional grid on X_1, X_2

Outlook: our next project

- Simulation study for comparing different methods for non-linear modeling in the "univariable" case
  - E.g. fractional polynomials, p-splines…
- Demonstrate how to choose a suitable set of performance measures

How to choose a suitable set of performance measures?

According to our categorization, there are…

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), and a single point for measures with localization = "point"

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

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

→Can we automate this?

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.

Strasak, A. M., Umlauf, N., Pfeiffer, R. M., & Lang, S. (2011). Comparing penalized splines and fractional polynomials for flexible modelling of the effects of continuous predictor variables. Computational Statistics & Data Analysis, 55(4), 1540-1551.