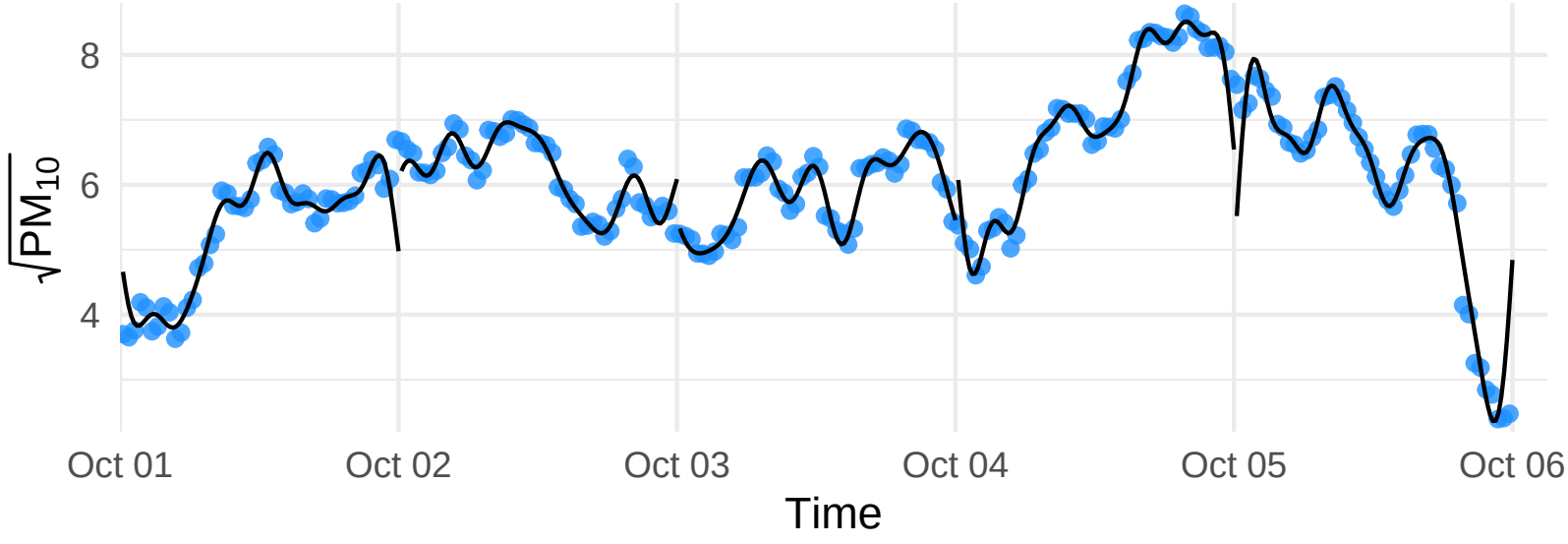
Estimating the conditional distribution in functional regression problems

Thomas Kuenzer

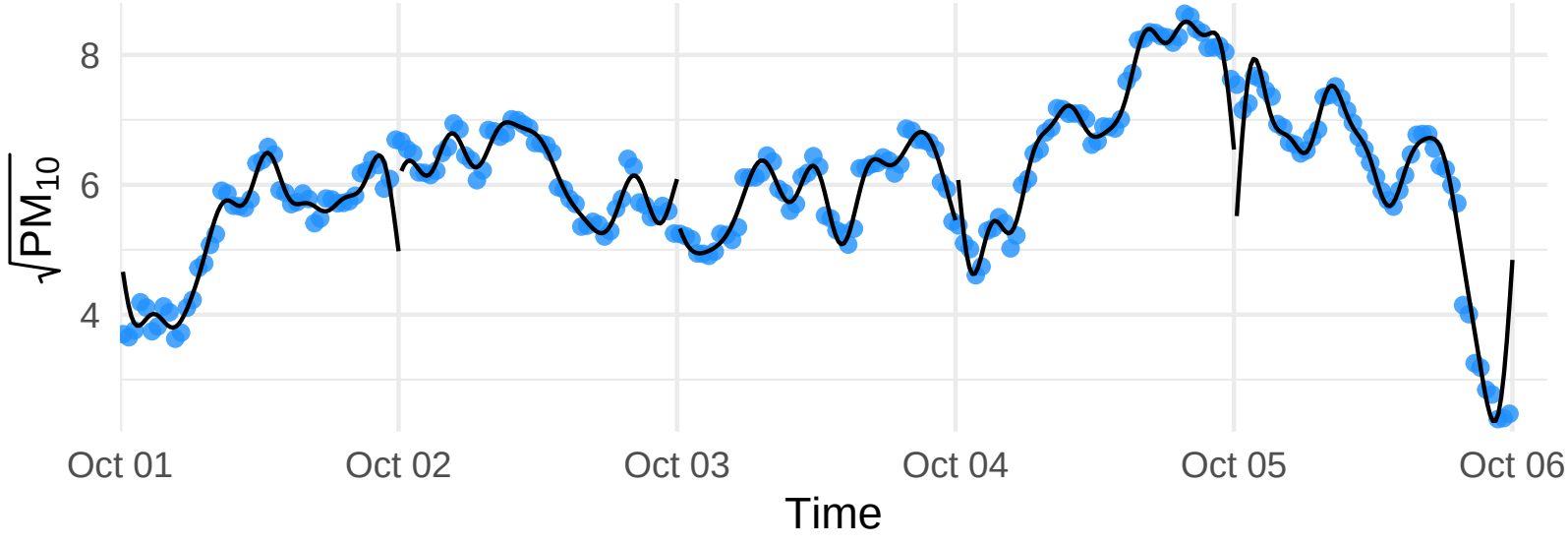
25th Young Statisticians Meeting, Vorau, 16 October 2021.

> kuenzer@tugraz.at

What is functional data?

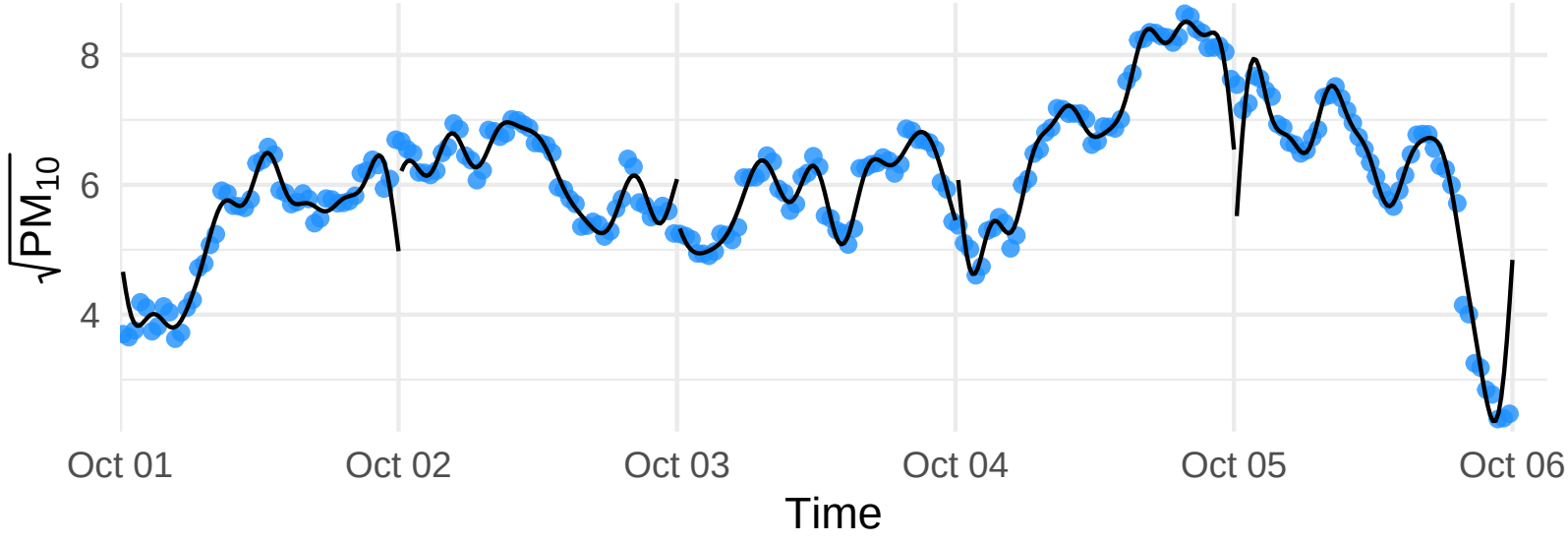


What is functional data?



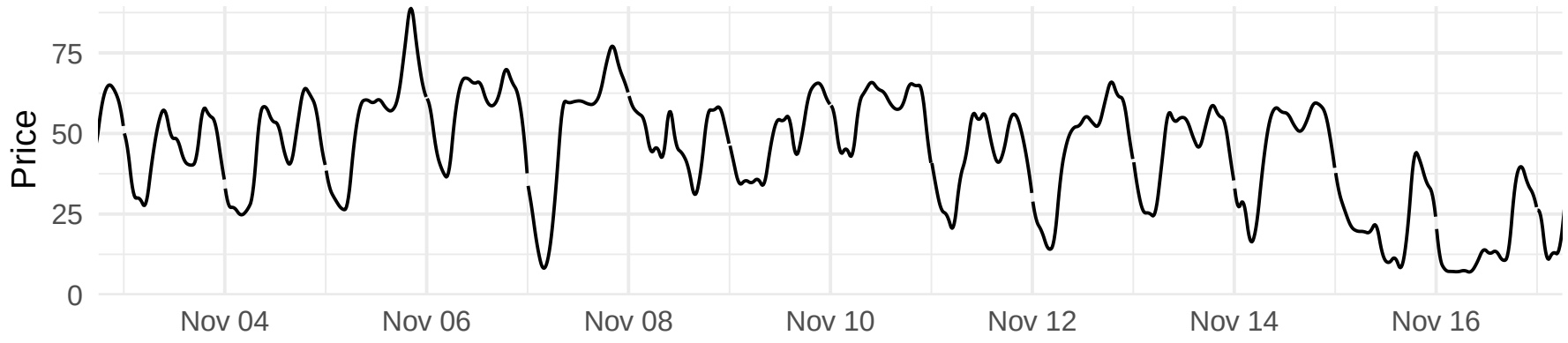
- Infinite-dimensional space instead of $\mathbb{R}^n$
- Use smoothing of "dense" observations as proxy for unobservable functional object

What is functional data?



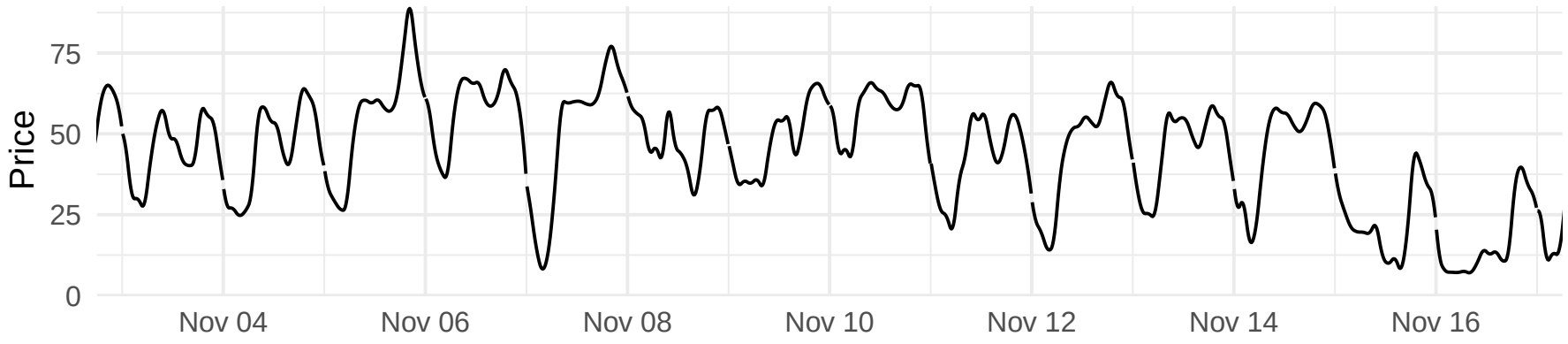
- Infinite-dimensional space instead of $\mathbb{R}^n$
- Use smoothing of "dense" observations as proxy for unobservable functional object

Function-on-function regression



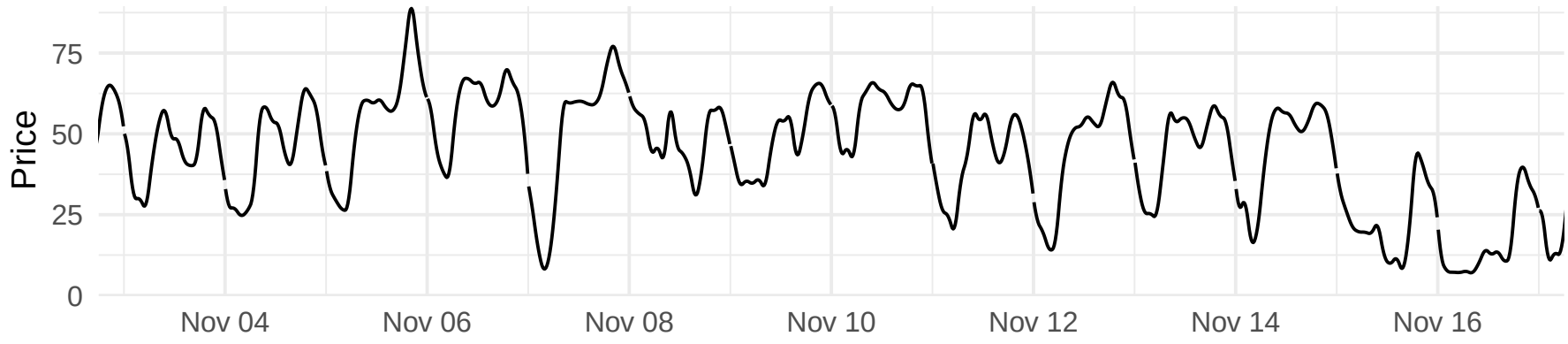
- Observe curve X_k
- Predict response curve Y_k

Function-on-function regression



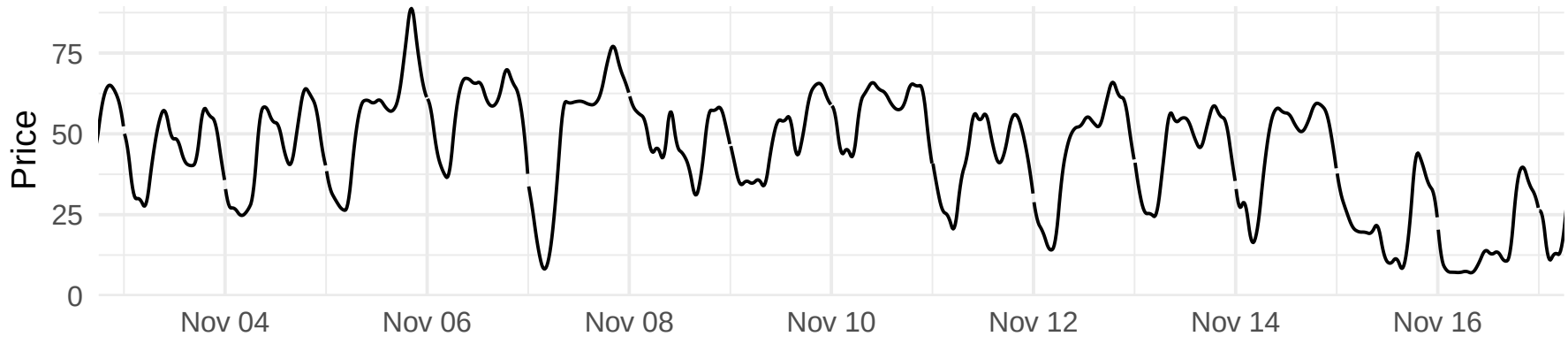
- Observe curve X_k
- Predict response curve Y_k $\Rightarrow$ in **what aspect**?

Function-on-function regression



- Observe curve X_k
- Predict response curve Y_k $\Rightarrow$ in **what aspect**?
- $E[Y_k|X_k]$ often uninformative $\Rightarrow P(Y_k \in A|X_k)$

Level sets

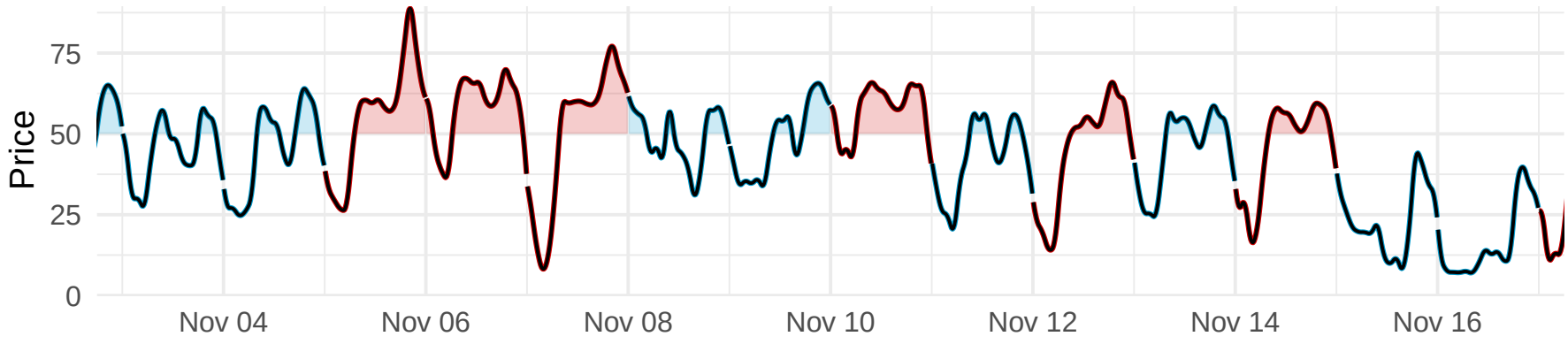


Let

$$A_{\alpha,z} := \left\{ y \in H_2 \colon \lambda(t \colon y(t) > \alpha) \leq z \right\}$$

for some $z \in [0, 1]$ and $\alpha \in \mathbb{R}$.
The *level set* $A_{\alpha,z}$ contains curves that stay a limited amount of time z above the threshold α.

Level sets

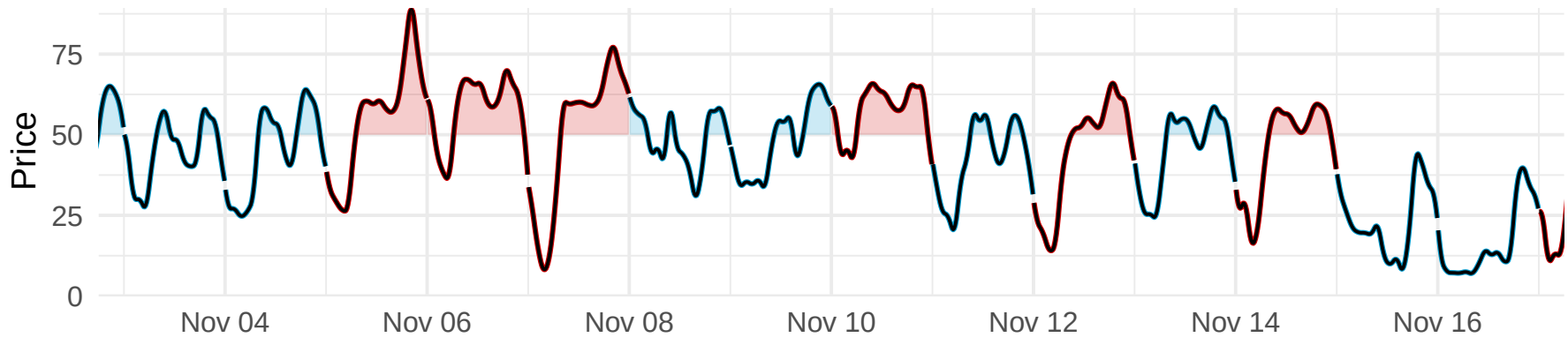


Let

$$A_{\alpha,z} := \left\{ y \in H_2 \colon \lambda(t \colon y(t) > \alpha) \leq z \right\}$$

for some $z \in [0, 1]$ and $\alpha \in \mathbb{R}$.
The *level set* $A_{\alpha,z}$ contains curves that stay a limited amount of time z above the threshold α.

A regression problem



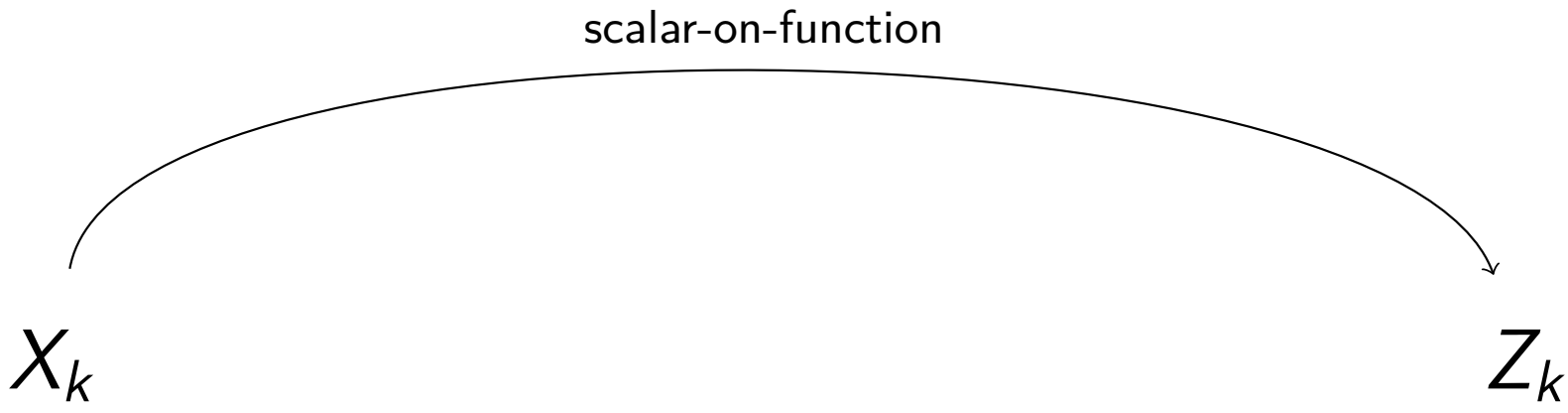
- covariate $X_k \in H_1$.
- response $Y_k \in C[0,1]$.
- $Z_k := \lambda(t\colon\ Y_k(t) > \alpha)\ \in [0,1]$.

$\Rightarrow$ $P(Y_k \in A_{\alpha,z}|X_k) = P(Z_k \leq z|X_k)$.

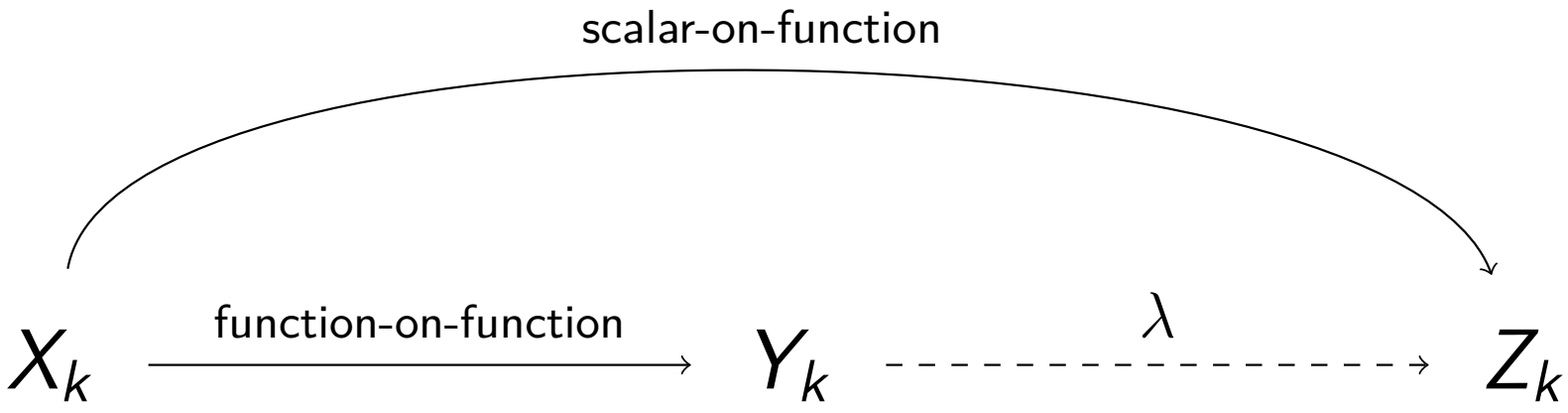
$\Rightarrow$ scalar-on-function regression?

$\Rightarrow$ quantile regression?

A regression problem



A regression problem



A regression problem

$$X_k \xrightarrow{\text{function-on-function}} Y_k \overset{\lambda}{\dashrightarrow} Z_k$$

- Linear function-on-function regression model

$$Y_k = \varrho X_k + \varepsilon_k.$$

- Use the full information of the sample to calculate principal component based estimator $\hat{\varrho}_n$.
- Combine point prediction and model residuals for estimation of $P(Y \in A|X)$

Define the right-truncated estimator

$$\hat{\varrho}_n(x) := \sum_{i=1}^{T_n} \frac{1}{\hat{\lambda}_i} \widehat{C}_{YX}\, \hat{v}_i \otimes \hat{v}_i(x),$$

where $\hat{\lambda}_i$ and $\hat{v}_i$ are the non-increasing eigenvalues and the eigenfunctions of $\widehat{C}_{XX}$.

Define the right-truncated estimator

$$\hat{\varrho}_n(x) := \sum_{i=1}^{T_n} \frac{1}{\hat{\lambda}_i} \widehat{C}_{YX}\, \hat{v}_i \otimes \hat{v}_i(x),$$

where $\hat{\lambda}_i$ and $\hat{v}_i$ are the non-increasing eigenvalues and the eigenfunctions of $\widehat{C}_{XX}$.

Select

$$T_n = \max\left\{j \geq 1 \colon \hat{\lambda}_j \geq m_n^{-1}\right\},$$
$$\text{with} \quad m_n = o\left(n^{\alpha/2}\right) \to \infty.$$

- Calculate model residuals

$$\hat{\varepsilon}_k := Y_k - \hat{\varrho}_n X_k.$$

- Estimate $P(Y \in A|X)$ using the empirical distribution

$$\widehat{P}(Y \in A|X) = \frac{1}{n} \sum_{k=1}^{n} \mathcal{I}(\hat{\varrho}_n X + \hat{\varepsilon}_k \in A).$$

- Also possible: use Gaussian distribution with $\tilde{\varepsilon}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \widehat{C}_{\hat{\varepsilon}})$

$$\widetilde{P}(Y \in A|X) = \frac{1}{B} \sum_{i=1}^{B} \mathcal{I}(\hat{\varrho}_n X + \tilde{\varepsilon}_i \in A)$$

- $(X_k)_{k\in\mathbb{Z}}$: L^4-m-approximable in sep. Hilbert space H_1.
- $(\varepsilon_k)_{k\in\mathbb{Z}}$: mean zero, i.i.d. sequence with $E\|\varepsilon_k\|^4 < \infty$. independent from $(X_j)_{j\leq k}$ for all $k \in \mathbb{Z}$.

$$|\varepsilon_k(t) - \varepsilon_k(s)| \leq M_k|t-s|^\alpha \quad \forall\, t,s \in [0,1].$$

- $\varrho\colon H_1 \to H_2$ is a bounded linear operator.
- A is continuity set of the response, i.e. $P(Y \in \partial A) = 0$.

$\Rightarrow$ Empirical distribution $\widehat{P}(Y \in A|X)$ is consistent.

- $(X_k)_{k\in\mathbb{Z}}$: L^4-m-approximable in sep. Hilbert space H_1.
- $(\varepsilon_k)_{k\in\mathbb{Z}}$: mean zero, i.i.d. sequence with $E\|\varepsilon_k\|^4 < \infty$. independent from $(X_j)_{j\leq k}$ for all $k \in \mathbb{Z}$.

$$|\varepsilon_k(t) - \varepsilon_k(s)| \leq M_k|t-s|^\alpha \quad \forall\, t,s \in [0,1].$$

- $\varrho\colon H_1 \to H_2$ is a bounded linear operator.
- A is continuity set of the response, i.e. $P(Y \in \partial A) = 0$.

$\Rightarrow$ Empirical distribution $\widehat{P}(Y \in A|X)$ is consistent.

Consistency

- $(X_k)_{k\in\mathbb{Z}}$: L^4-m-approximable in sep. Hilbert space H_1.
- $(\varepsilon_k)_{k\in\mathbb{Z}}$: mean zero, i.i.d. sequence with $E\|\varepsilon_k\|^4 < \infty$. independent from $(X_j)_{j\leq k}$ for all $k \in \mathbb{Z}$.

$$|\varepsilon_k(t) - \varepsilon_k(s)| \leq M_k|t-s|^\alpha \quad \forall\, t,s \in [0,1].$$

- ϱ: $H_1 \to H_2$ is a bounded linear operator.
- A is continuity set of the response, i.e. $P(Y \in \partial A) = 0$.

$\Rightarrow$ Empirical distribution $\widehat{P}(Y \in A|X)$ is consistent.

- $(X_k)_{k\in\mathbb{Z}}$: L^4-m-approximable in sep. Hilbert space H_1.
- $(\varepsilon_k)_{k\in\mathbb{Z}}$: mean zero, i.i.d. sequence with $E\|\varepsilon_k\|^4 < \infty$. independent from $(X_j)_{j\leq k}$ for all $k \in \mathbb{Z}$.

$$|\varepsilon_k(t) - \varepsilon_k(s)| \leq M_k|t-s|^\alpha \quad \forall\, t,s \in [0,1].$$

- ϱ: $H_1 \to H_2$ is a bounded linear operator.
- A is continuity set of the response, i.e. $P(Y \in \partial A) = 0$.

$\Rightarrow$ Empirical distribution $\widehat{P}(Y \in A|X)$ is consistent.

- $(X_k)_{k\in\mathbb{Z}}$: L^4-m-approximable in sep. Hilbert space H_1.
- $(\varepsilon_k)_{k\in\mathbb{Z}}$: mean zero, i.i.d. sequence with $E\|\varepsilon_k\|^4 < \infty$. independent from $(X_j)_{j\leq k}$ for all $k \in \mathbb{Z}$.

$$|\varepsilon_k(t) - \varepsilon_k(s)| \leq M_k|t-s|^\alpha \quad \forall\, t,s \in [0,1].$$

- ϱ: $H_1 \to H_2$ is a bounded linear operator.
- A is continuity set of the response, i.e. $P(Y \in \partial A) = 0$.

$\Rightarrow$ Empirical distribution $\widehat{P}(Y \in A|X)$ is consistent.

- If additionally ϱ satisfies

$$|\varrho x(t) - \varrho x(s)| \leq M_\varrho \|x\||t-s|^\alpha \quad \forall\, t,s \in [0,1].$$

$\Rightarrow$ Gaussian distribution $\widetilde{P}(Y \in A|X)$ is consistent.

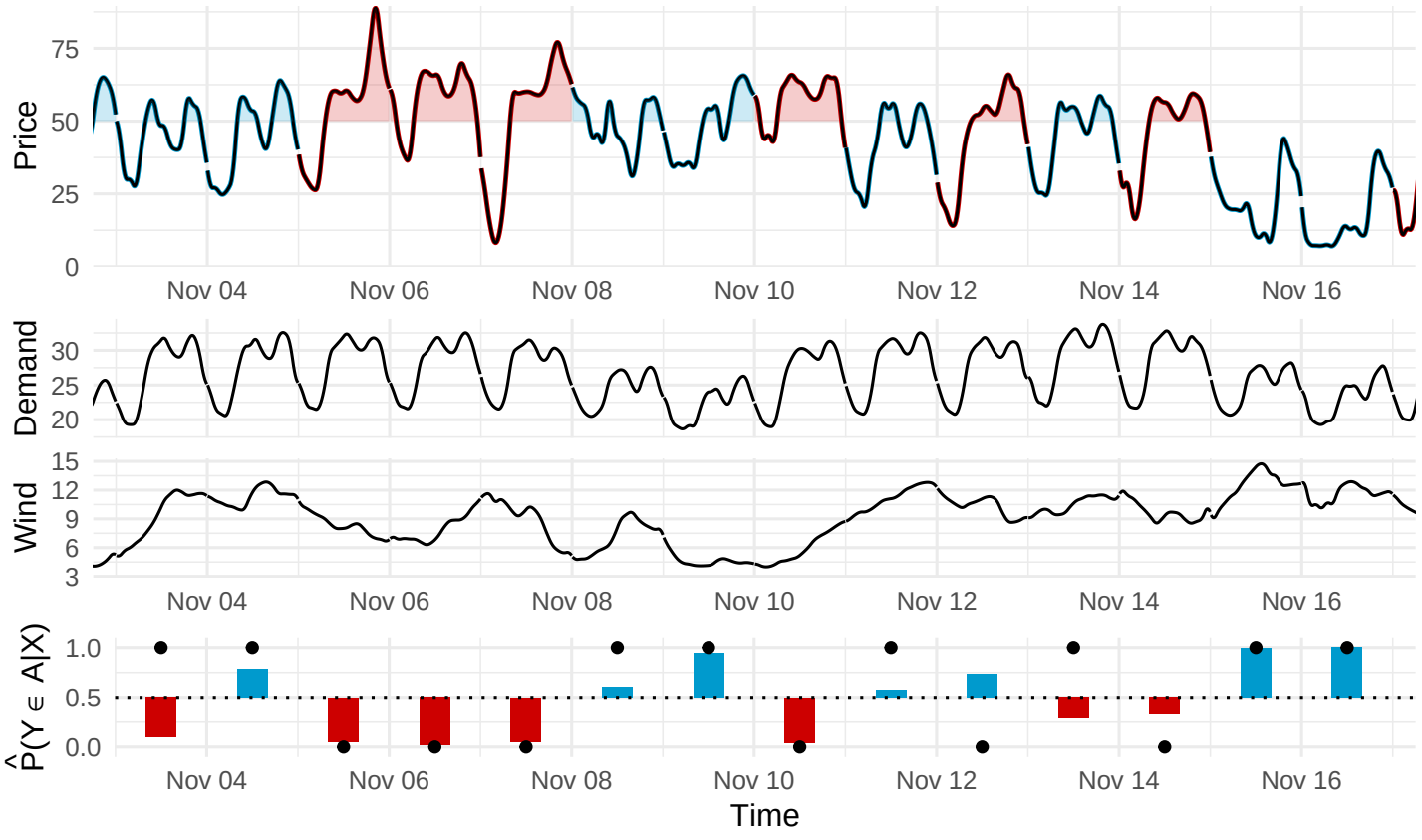
Sets of interest in function spaces:

- contrast sets
- bands
- level sets, extremal sets, excursion sets

Applications:

- conditional probability estimation for a large class of sets
- prediction bands for functional time series
- functional quantile regression

Application



- FARX(7) model. Estimate $P(Y \in A|X)$ using $\widehat{P}$.
- Competing methods: Nadaraya–Watson, GLMs
- Assess performance using cross-entropy on test set

	α	30	40	50	60	70
$z = 0$	**empir**	**0.03**	**0.10**	**0.16**	0.23	**0.16**
	GLM	0.11	0.23	0.30	**0.22**	0.24
	N–W	0.05	0.20	0.22	0.29	0.26
$z = \frac{1}{3}$	**empir**	**0.05**	**0.12**	**0.20**	**0.15**	**0.09**
	GLM	0.25	0.26	0.23	0.23	0.39
	N–W	0.10	0.23	0.26	0.24	0.14
$z = \frac{2}{3}$	**empir**	**0.09**	**0.14**	**0.22**	**0.12**	0.03
	GLM	0.12	0.27	0.25	0.20	0.12
	N–W	0.16	0.23	0.28	0.19	**0.02**

Cross-entropy of the estimated conditional probability $\widehat{P}(\lambda(Y > \alpha) \leq z|X)$ evaluated on the test set.

Conclusion

- **Simple and modular** method
- Consistency under **mild assumptions**
- **Many applications**:
  quantile regression, prediction bands, etc.
- **Better performance** than more specialized and sophisticated methods

Thank you for your attention

Estimating the conditional distribution in functional regression problems

Siegfried Hörmann [1], Thomas Kuenzer [1], Gregory Rice [2]

Preprint on https://arxiv.org/abs/2105.01412

[1] Institute of Statistics, Graz University of Technology, Graz, Austria.

[2] Department of Statistics and Actuarial Science, University of Waterloo, Canada.