Valid prediction intervals for regression and time series forecasting problems

Willem Waegeman

Department of Data Analysis and Mathematical Modelling, Ghent University, Belgium

February 2022

Research interests

Fundamental research in machine learning:

- **Multi-target prediction:** multi-label classification, multi-task learning, dyadic prediction, matrix completion, etc.
- **Sequence models and time series analysis:** sequence labelling, autoregressive models, feature learning for sequential data, etc.
- **Uncertainty quantification**: regression and classification settings, Bayes-optimal inference, epistemic and aleatoric uncertainty, open-set recognition, etc.

Research interests

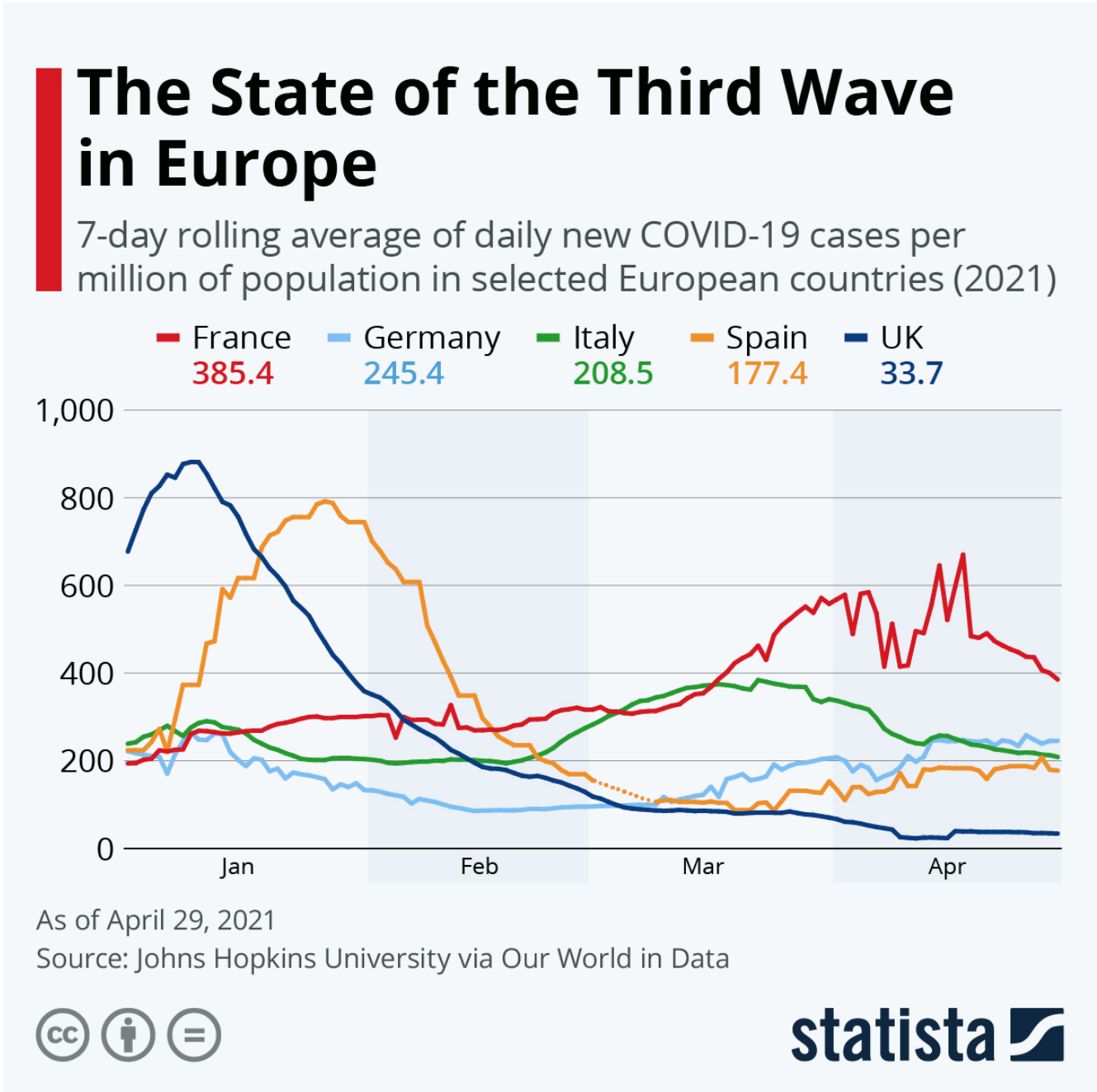
Fundamental research in machine learning:

- **Multi-target prediction:** multi-label classification, multi-task learning, dyadic prediction, matrix completion, etc.
- **Sequence models and time series analysis:** sequence labelling, autoregressive models, feature learning for sequential data, etc.
- **Uncertainty quantification**: regression and classification settings, Bayes-optimal inference, epistemic and aleatoric uncertainty, open-set recognition, etc.

Application domains in which I am active:

- **Analysis of omics data:** genomics, epigenomics, transcriptomics, proteomics, etc.
- **Environmental and medical applications**
- **Climate analysis**

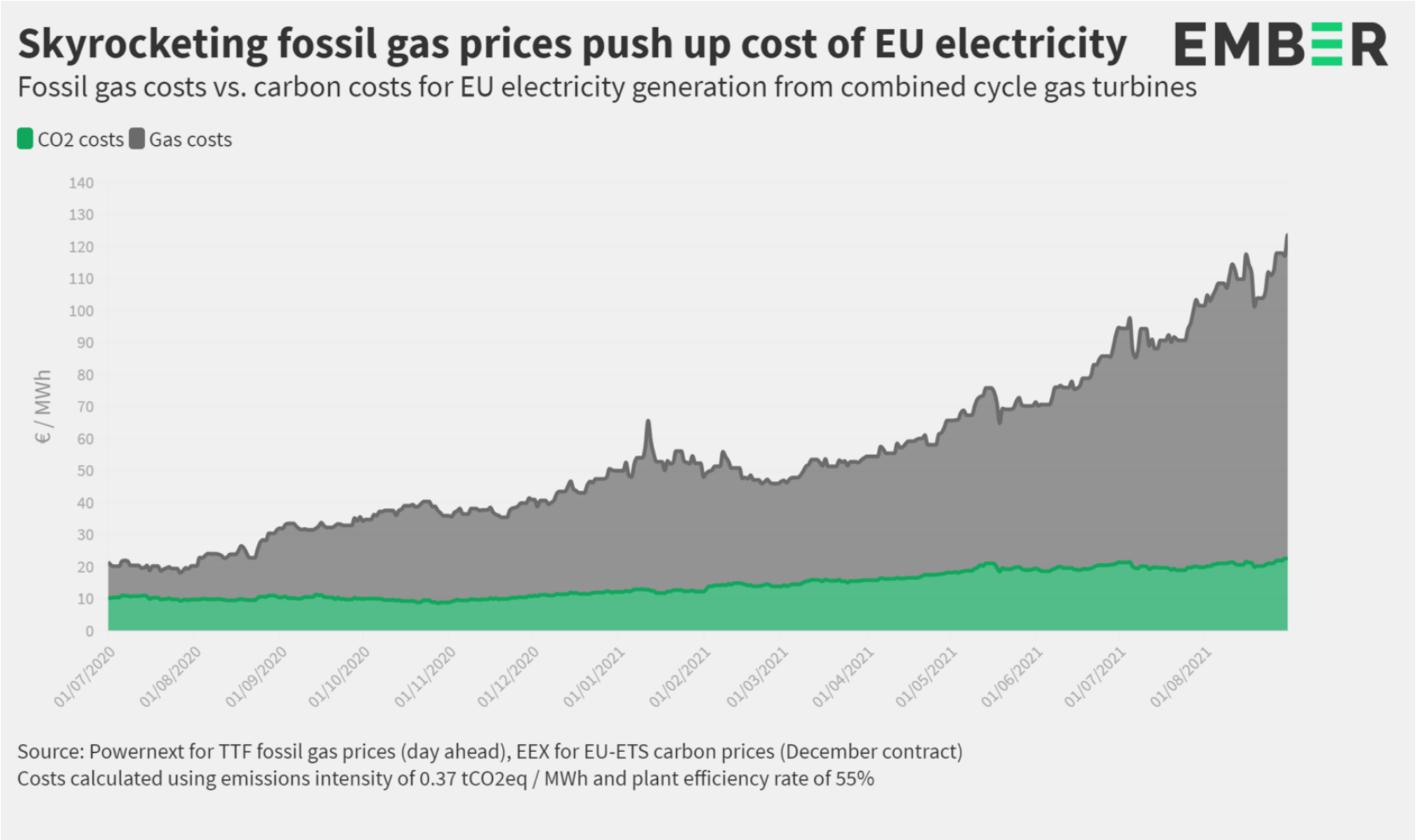
Motivation


The State of the Third Wave in Europe
7-day rolling average of daily new COVID-19 cases per million of population in selected European countries (2021)
France 385.4
Germany 245.4
Italy 208.5
Spain 177.4
UK 33.7
1,000
800
600
400
200
0
Jan
Feb
Mar
Apr
As of April 29, 2021
Source: Johns Hopkins University via Our World in Data
statista

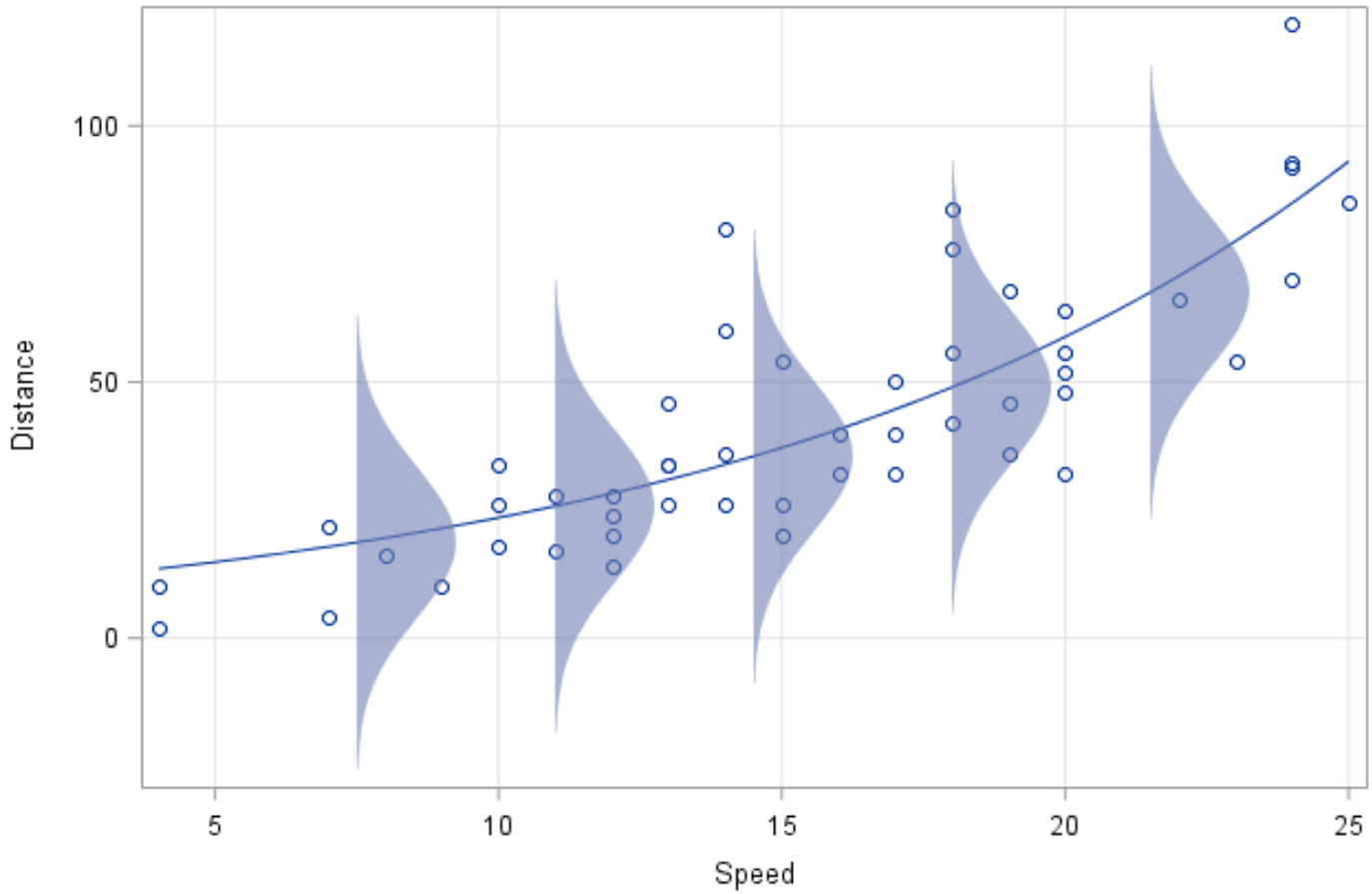
Motivation

Source: https://insideclimatenews.org/news/11082017/europe-rivers-flooding-timing-climate-change-snowmelt-storms/

Motivation


Skyrocketing fossil gas prices push up cost of EU electricity
EMBER
Fossil gas costs vs. carbon costs for EU electricity generation from combined cycle gas turbines
CO2 costs
Gas costs
€ / MWh
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
01/07/2020
01/08/2020
01/09/2020
01/10/2020
01/11/2020
01/12/2020
01/01/2021
01/02/2021
01/03/2021
01/04/2021
01/05/2021
01/06/2021
01/07/2021
01/08/2021
Source: Powernext for TTF fossil gas prices (day ahead), EEX for EU-ETS carbon prices (December contract)
Costs calculated using emissions intensity of 0.37 tCO2eq / MWh and plant efficiency rate of 55%

Central question for this talk:
How can we construct **distribution-free** prediction intervals for **regression** and **time series forecasting** problems?



Recent paper:
Nicolas Dewolf, Bernard De Baets and Willem Waegeman, Valid prediction intervals for regression problems, See Arxiv 2021.

Formal problem definition **regression** setting:

- **Point predictor** $\rho : \mathcal{X} \to \mathbb{R}$
- **Interval predictor** $\Gamma : \mathcal{X} \to [\mathbb{R}]$, where $[\mathbb{R}]$ denotes the set of all closed subintervals of $\mathbb{R}$.
- Training and test data **i.i.d. sampled** according to an unknown distribution P over $\mathcal{X} \times \mathbb{R}$

Formal problem definition **regression** setting:

- **Point predictor** $\rho : \mathcal{X} \to \mathbb{R}$
- **Interval predictor** $\Gamma : \mathcal{X} \to [\mathbb{R}]$, where $[\mathbb{R}]$ denotes the set of all closed subintervals of $\mathbb{R}$.
- Training and test data **i.i.d. sampled** according to an unknown distribution P over $\mathcal{X} \times \mathbb{R}$

Evaluation criteria for test data:

- **Marginal validity** for a desired coverage level $1 - \alpha$:

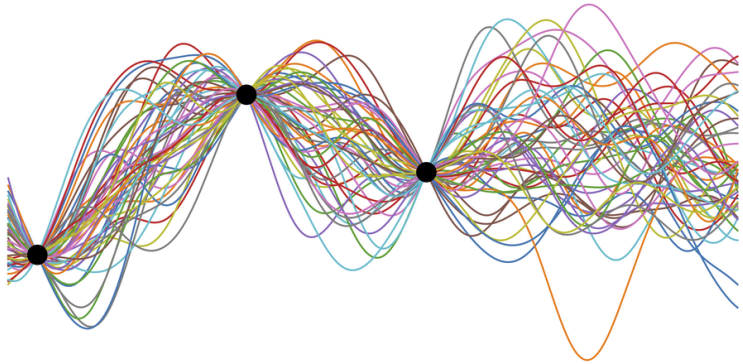
$$\mathcal{C}(\Gamma, P) := \mathsf{E}_{(\mathbf{x},y)\sim P}\left[\mathbb{I}(y \in \Gamma(\mathbf{x}))\right] = \mathsf{Prob}\{y \in \Gamma(\mathbf{x})\} \geq 1 - \alpha$$

- **Calibration:** when the inequality becomes an equality
- **Average width:**

$$\mathcal{W}(\Gamma, P) := \mathsf{E}_{(\mathbf{x},y)\sim P}\ |u(\mathbf{x}) - l(\mathbf{x})|$$

u, l denote the upper and lower bound of the interval produced by Γ

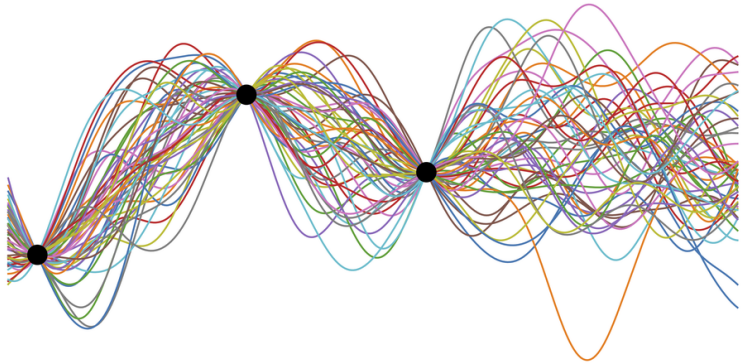
Bayesian methods



Update the **prior belief** $p(\theta)$ based on the data set using Bayes's rule:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathbf{y}, \theta \mid \mathbf{X})}{p(\mathbf{y} \mid \mathbf{X})} = \frac{p(\mathbf{y} \mid \mathbf{X}, \theta)p(\theta)}{\int p(\mathbf{y} \mid \mathbf{X}, \theta)p(\theta)d\theta}$$

Bayesian methods



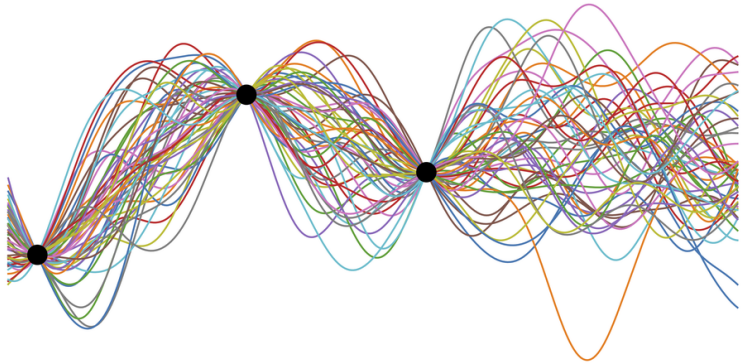
Update the **prior belief** $p(\theta)$ based on the data set using Bayes's rule:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathbf{y}, \theta \mid \mathbf{X})}{p(\mathbf{y} \mid \mathbf{X})} = \frac{p(\mathbf{y} \mid \mathbf{X}, \theta)p(\theta)}{\int p(\mathbf{y} \mid \mathbf{X}, \theta)p(\theta)d\theta}$$

The **posterior predictive distribution** is calculated by marginalizing over the parameters:

$$p(y^* \mid \mathbf{x}^*, \mathcal{D}) = \int p(y^* \mid \mathbf{x}^*, \theta)p(\theta \mid \mathcal{D})d\theta$$

Bayesian methods



Update the **prior belief** $p(\theta)$ based on the data set using Bayes's rule:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathbf{y}, \theta \mid \mathbf{X})}{p(\mathbf{y} \mid \mathbf{X})} = \frac{p(\mathbf{y} \mid \mathbf{X}, \theta)p(\theta)}{\int p(\mathbf{y} \mid \mathbf{X}, \theta)p(\theta)d\theta}$$

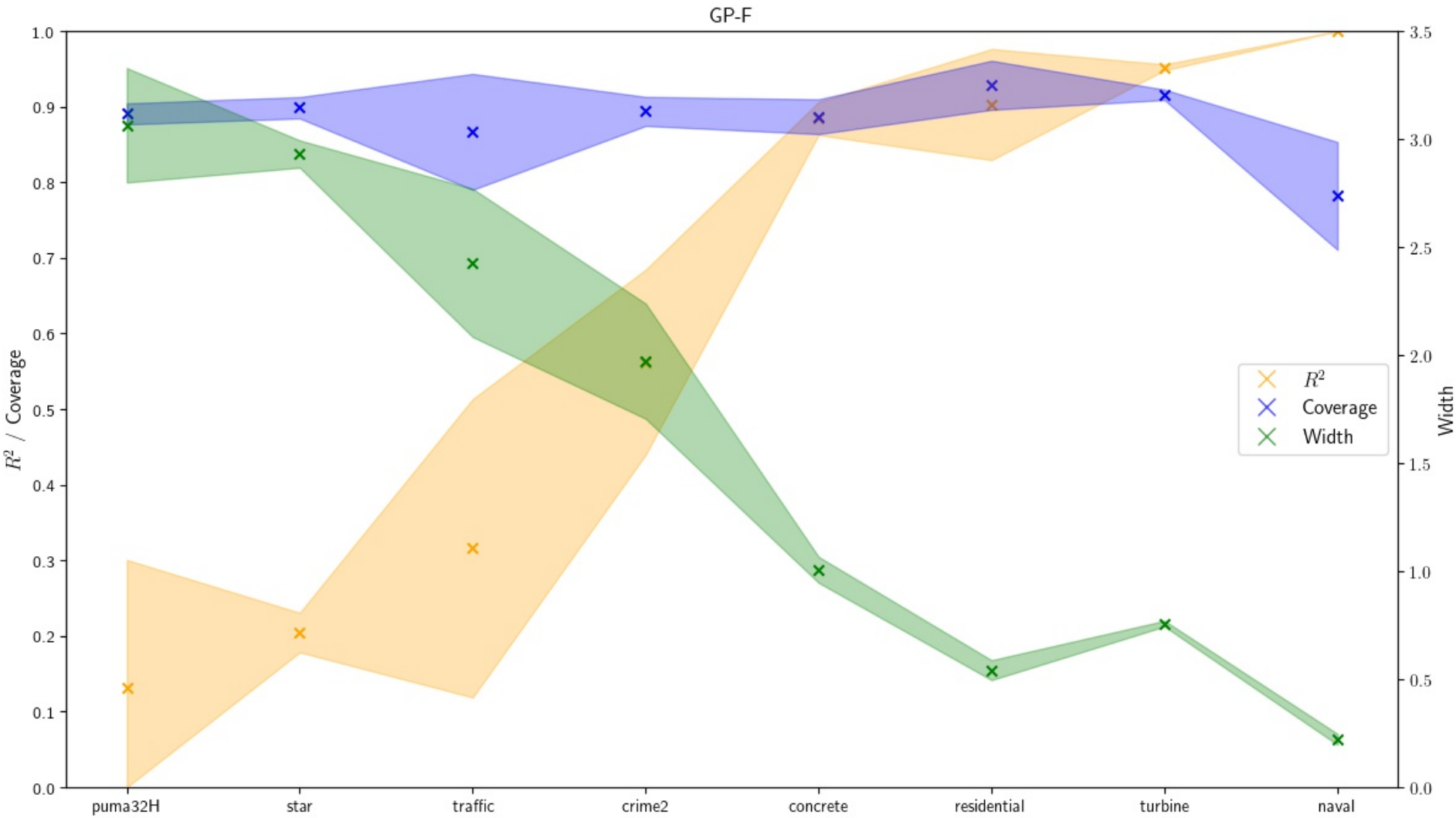
The **posterior predictive distribution** is calculated by marginalizing over the parameters:

$$p(y^* \mid \mathbf{x}^*, \mathcal{D}) = \int p(y^* \mid \mathbf{x}^*, \theta)p(\theta \mid \mathcal{D})d\theta$$

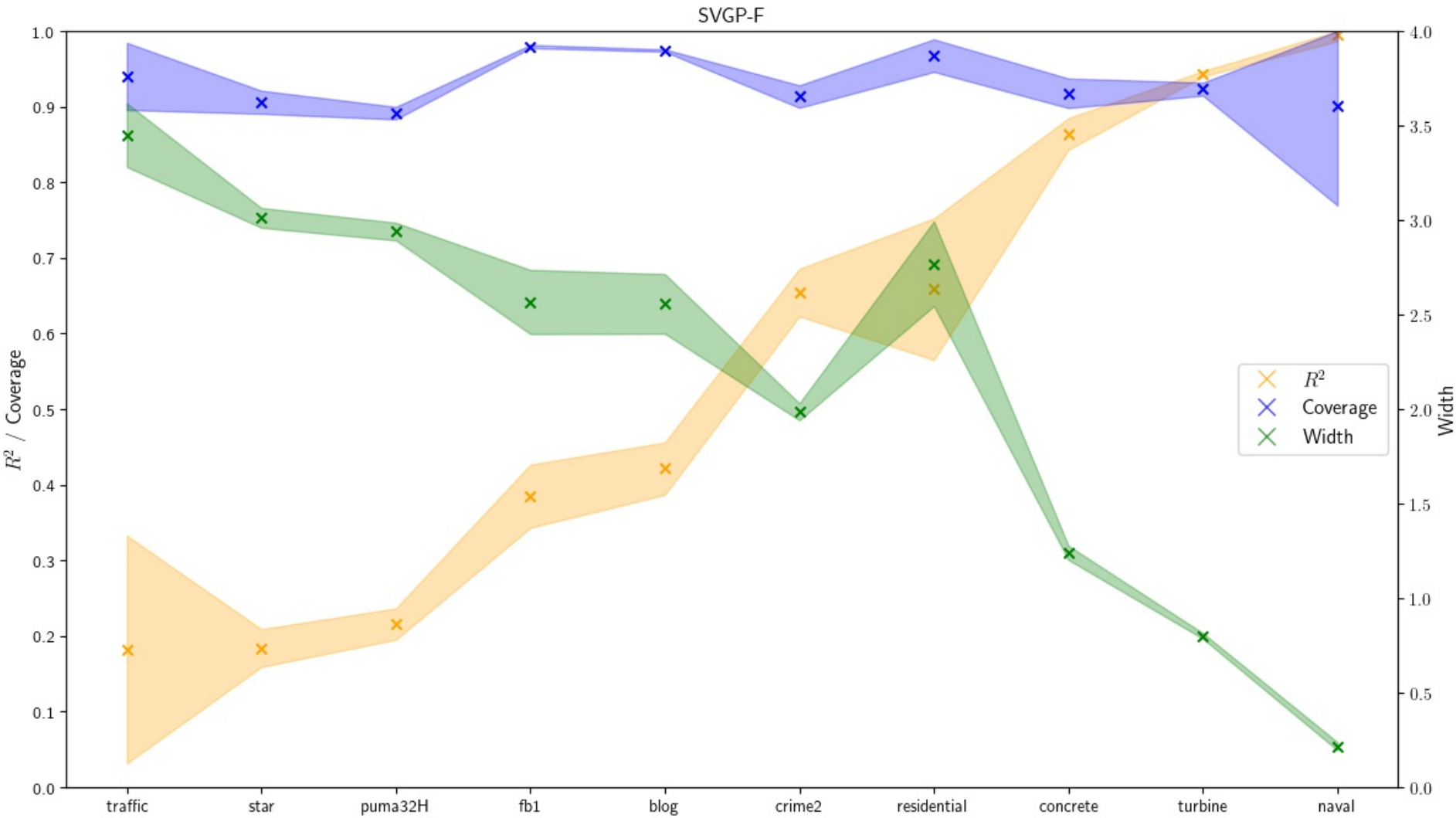
Popular instantiations:

- **Gaussian processes:** strict assumptions, exact inference possible
- **Bayesian neural nets:** less assumptions, approximate inference needed

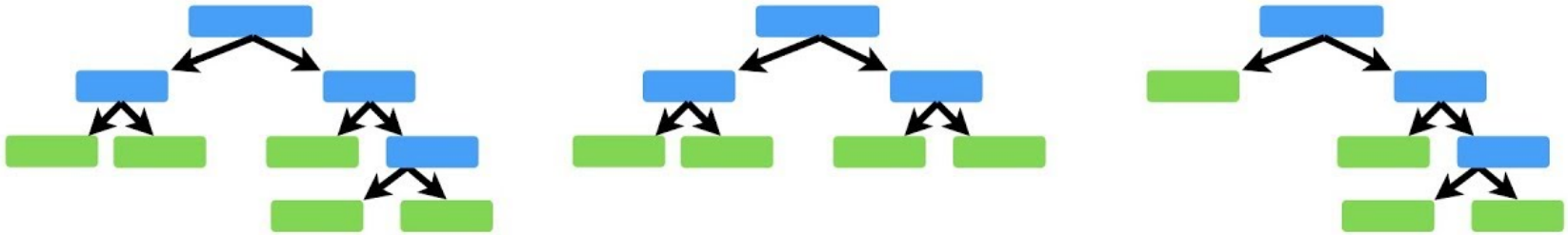
Results on 8/10 datasets (Exact Gaussian processes)



Results on 10 datasets (Approximate GPs)



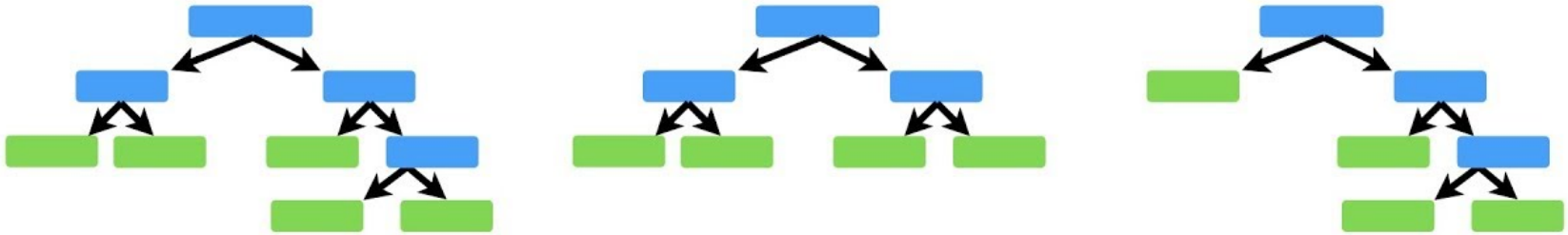
Ensemble methods



Specific methods:

- **Random forests:** intervals derived from out-of-bag estimates, cross-validation or ensemble quantile regression
- **Dropout networks:** naive approach or intervals that capture both epistemic and aleatoric uncertainty using specific loss functions
- **Deep ensembles:** rather focus on network initialization instead of bagging

Ensemble methods



Specific methods:

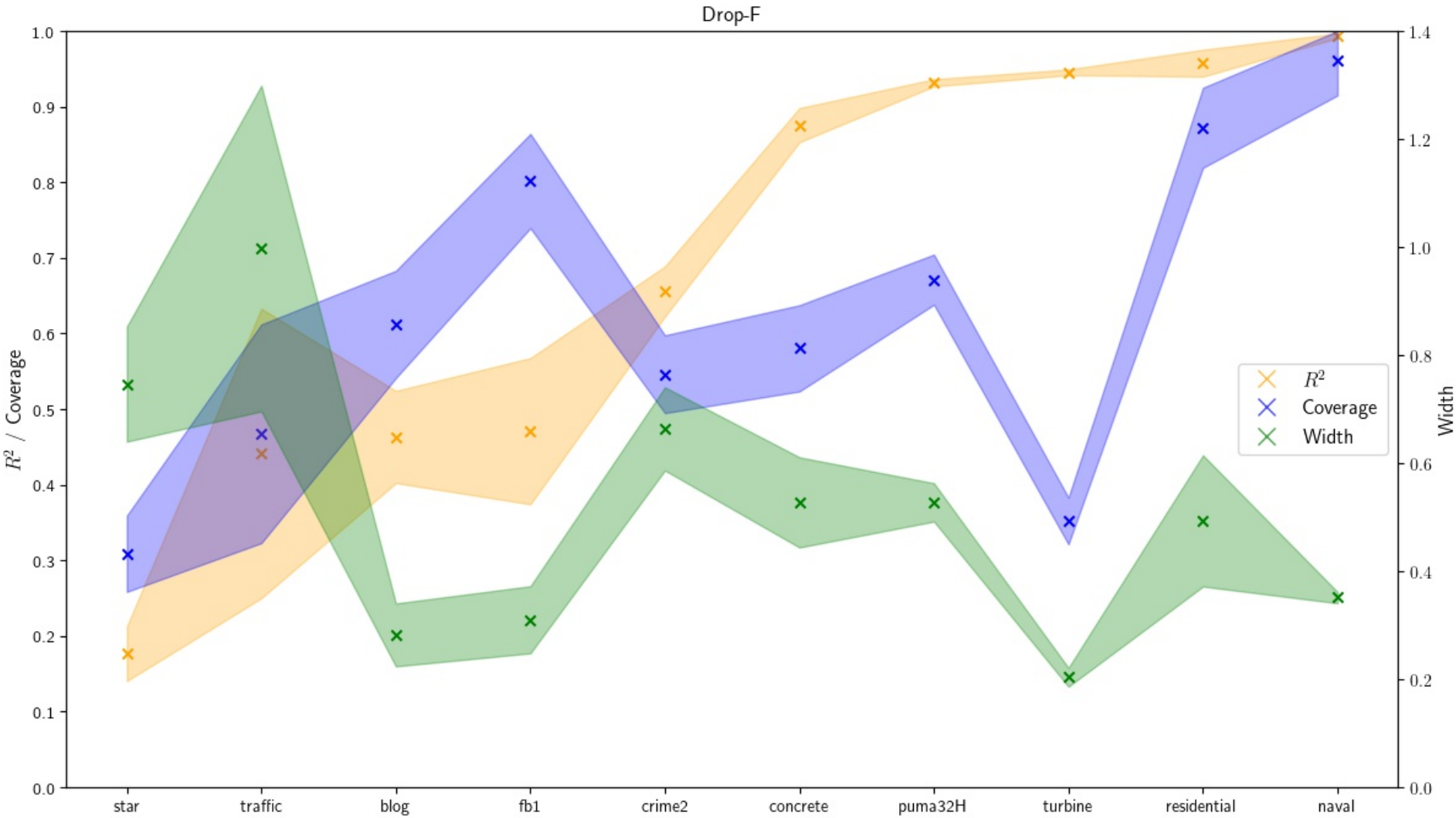
- **Random forests:** intervals derived from out-of-bag estimates, cross-validation or ensemble quantile regression
- **Dropout networks:** naive approach or intervals that capture both epistemic and aleatoric uncertainty using specific loss functions
- **Deep ensembles:** rather focus on network initialization instead of bagging

Naive approach: Gaussian prediction interval from

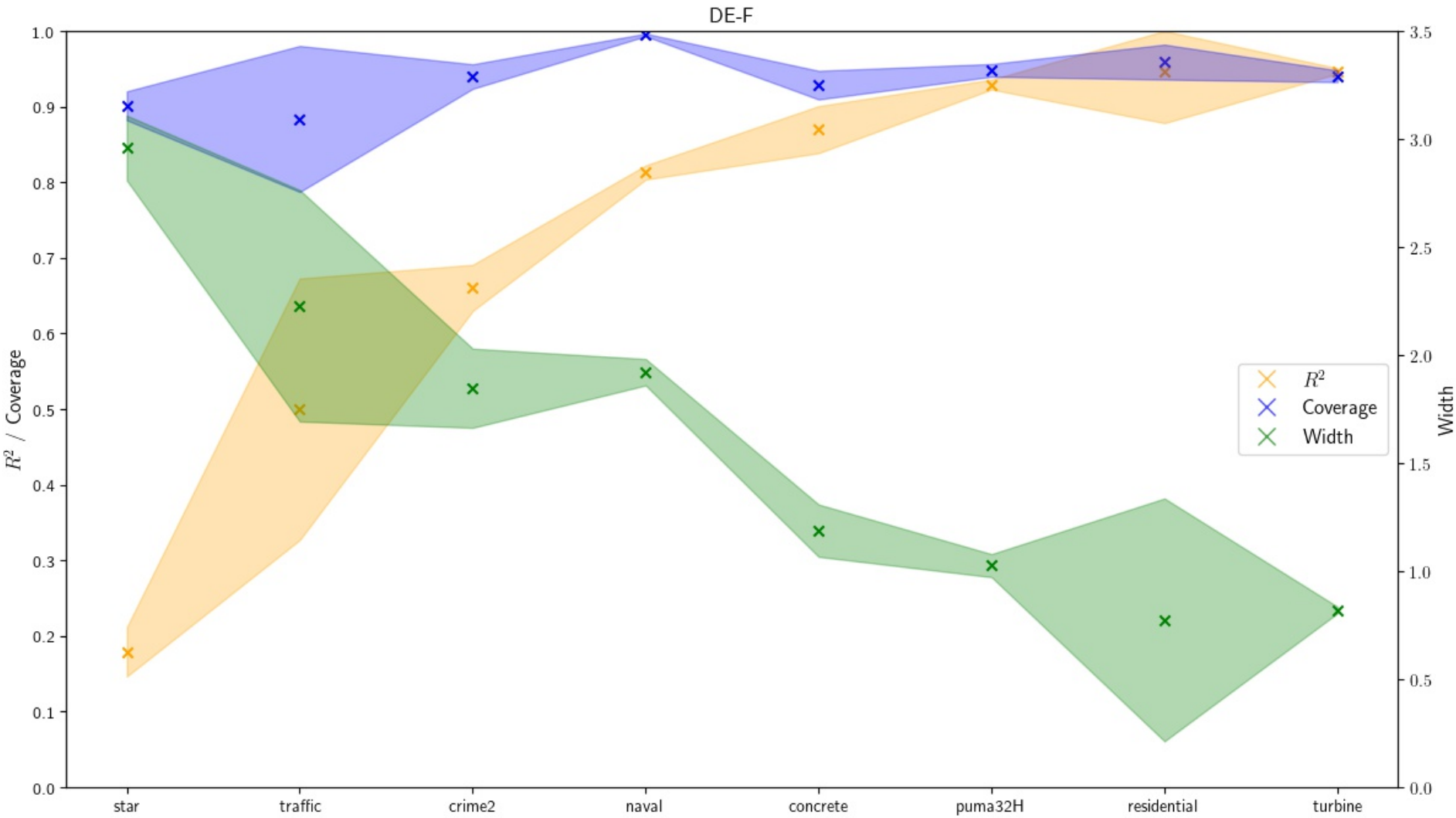
$$\Gamma^{\alpha}_{\text{ensemble}}(\mathbf{x}^*) := \left[\mathsf{E}[\rho \,|\, \mathbf{x}^*] - z^{\alpha}\sqrt{\mathsf{Var}[\rho \,|\, \mathbf{x}^*]}, \mathsf{E}[\rho \,|\, \mathbf{x}^*] + z^{\alpha}\sqrt{\mathsf{Var}[\rho \,|\, \mathbf{x}^*]}\right] ,$$

with z^{α} the two-tailed z-score corresponding to a significance level α

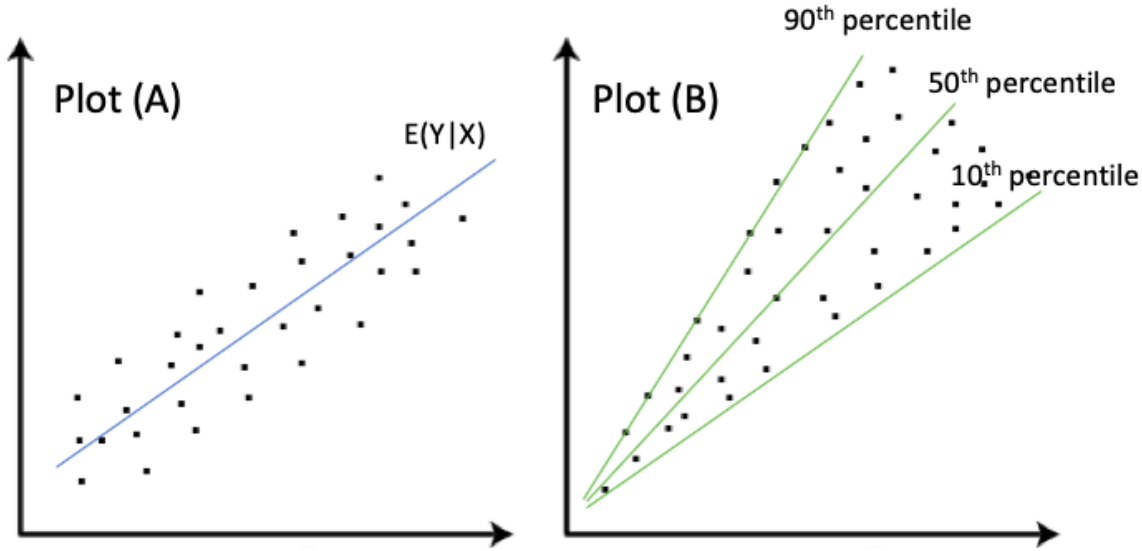
Results on 10 datasets (Dropout networks)



Results on 10 datasets (Deep ensembles)



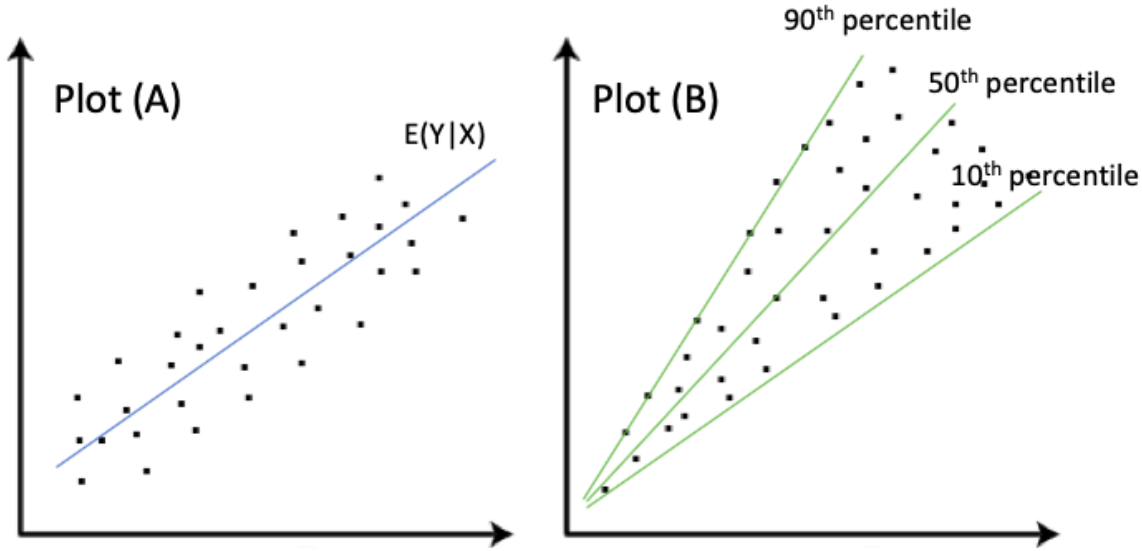
Direct interval estimators



- **Quantile regression:** estimation of quantiles $\hat{q}_\alpha(\mathbf{x})$ using the pinball loss

$$\mathcal{L}_{\text{pinball}}(\mathcal{T}) := \sum_{(\mathbf{x},y)\in\mathcal{T}} \max\Big((1-\alpha)(\hat{q}_\alpha(\mathbf{x})-y), \alpha(y-\hat{q}_\alpha(\mathbf{x}))\Big)$$

Direct interval estimators



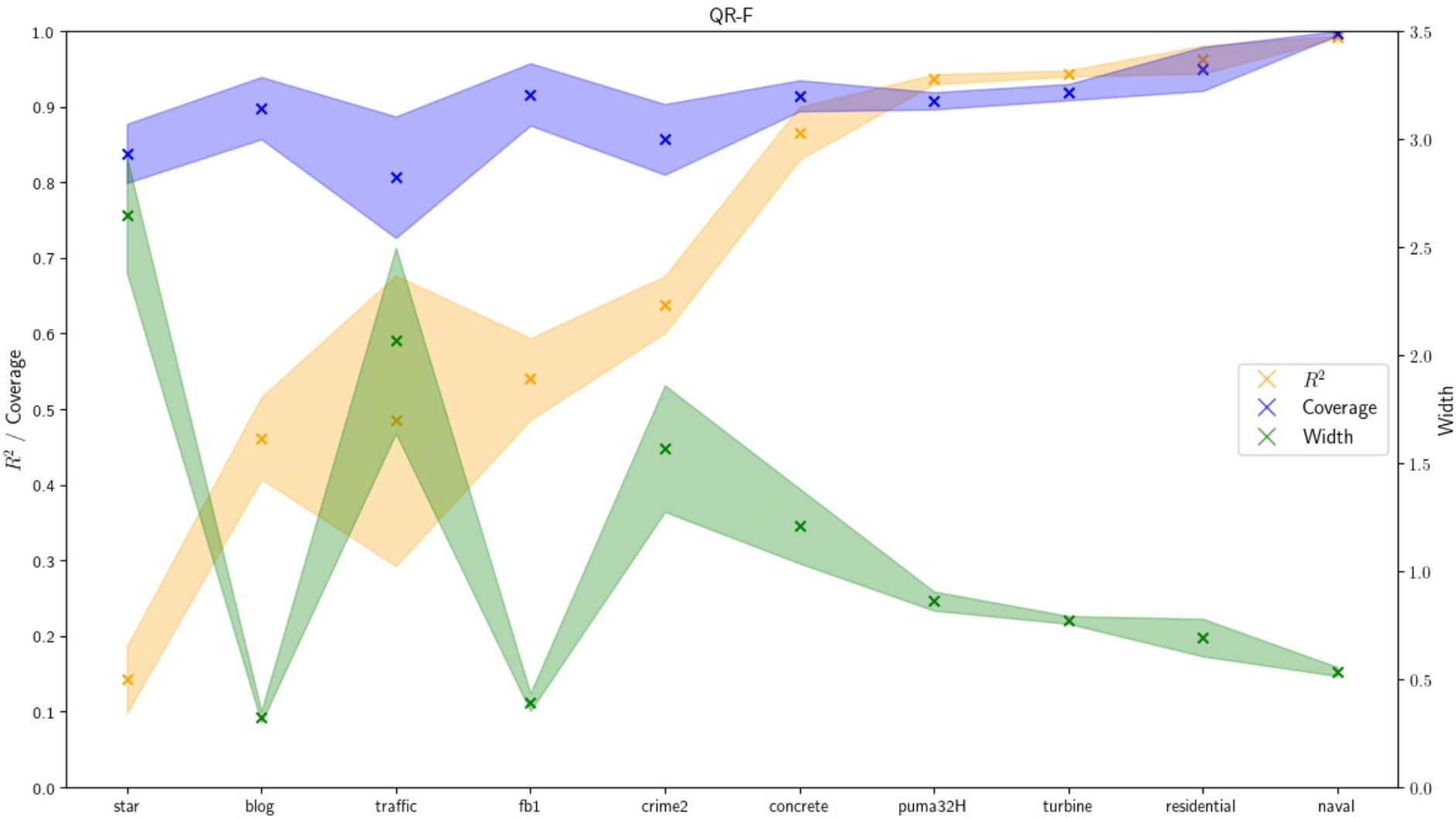
- **Quantile regression:** estimation of quantiles $\hat{q}_\alpha(\mathbf{x})$ using the pinball loss

$$\mathcal{L}_{\text{pinball}}(\mathcal{T}) := \sum_{(\mathbf{x},y)\in\mathcal{T}} \max\Big((1-\alpha)(\hat{q}_\alpha(\mathbf{x})-y), \alpha(y-\hat{q}_\alpha(\mathbf{x}))\Big)$$

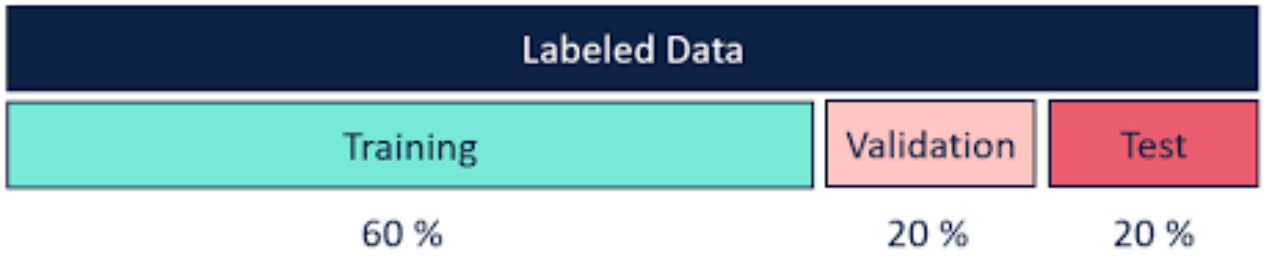
- **High-quality principle:** weighted loss that contains coverage $\mathcal{C}$ and mean interval width MPIW

$$\mathcal{L}_{\text{QD}}(\mathcal{T}) := \text{MPIW}_{\text{capt}} + \lambda \frac{n}{\alpha(1-\alpha)} \max\big(0, (1-\alpha)-\mathcal{C}\big)^2$$

Results on 10 datasets (Quantile regression)

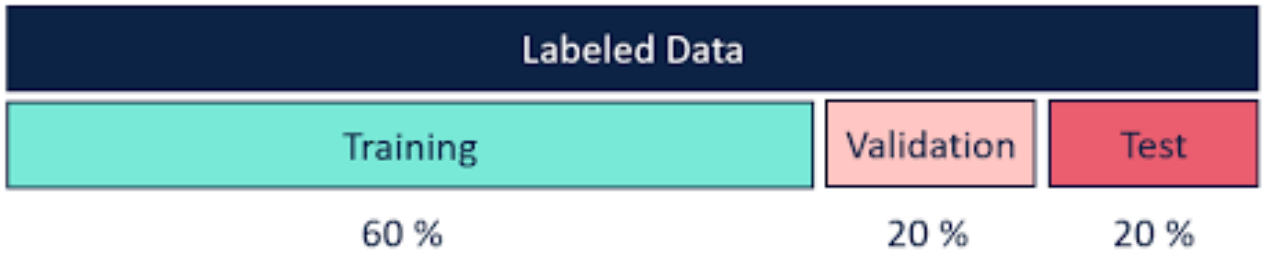


(Inductive) conformal prediction



- Calibrate a prediction interval using **nonconformity scores** computed on a **validation** dataset $\mathcal{V}$

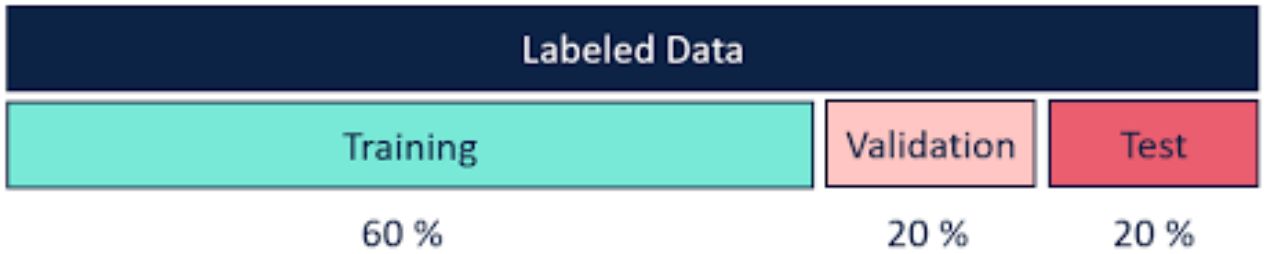
(Inductive) conformal prediction



- Calibrate a prediction interval using **nonconformity scores** computed on a **validation** dataset $\mathcal{V}$
- For a **point predictor**, consider the following nonconformity score:

$$a_i := |\rho(\mathbf{x}_i) - y_i|$$

(Inductive) conformal prediction

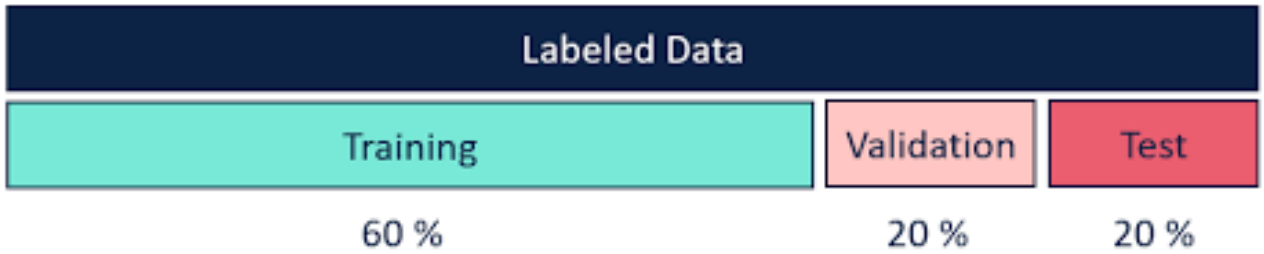


- Calibrate a prediction interval using **nonconformity scores** computed on a **validation** dataset $\mathcal{V}$
- For a **point predictor**, consider the following nonconformity score:

$$a_i := |\rho(\mathbf{x}_i) - y_i|$$

- Compute the $\left((1-\alpha)(1+\frac{1}{|\mathcal{V}|})\right)$-**quantile** of $a_i \in \mathcal{V}$

(Inductive) conformal prediction

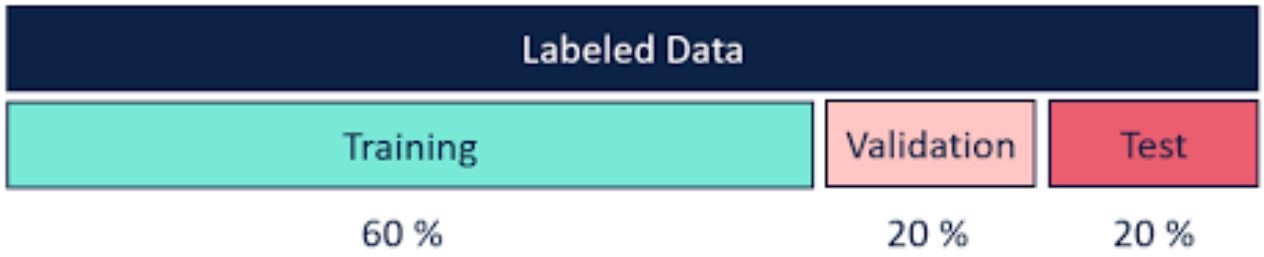


- Calibrate a prediction interval using **nonconformity scores** computed on a **validation** dataset $\mathcal{V}$
- For a **point predictor**, consider the following nonconformity score:

$$a_i := |\rho(\mathbf{x}_i) - y_i|$$

- Compute the $\left((1-\alpha)(1+\frac{1}{|\mathcal{V}|})\right)$-**quantile** of $a_i \in \mathcal{V}$
- For a **test instance** $\mathbf{x}^*$, the interval contains all $y^* \in \mathbb{R}$ for which the nonconformity score is smaller than that quantile

(Inductive) conformal prediction



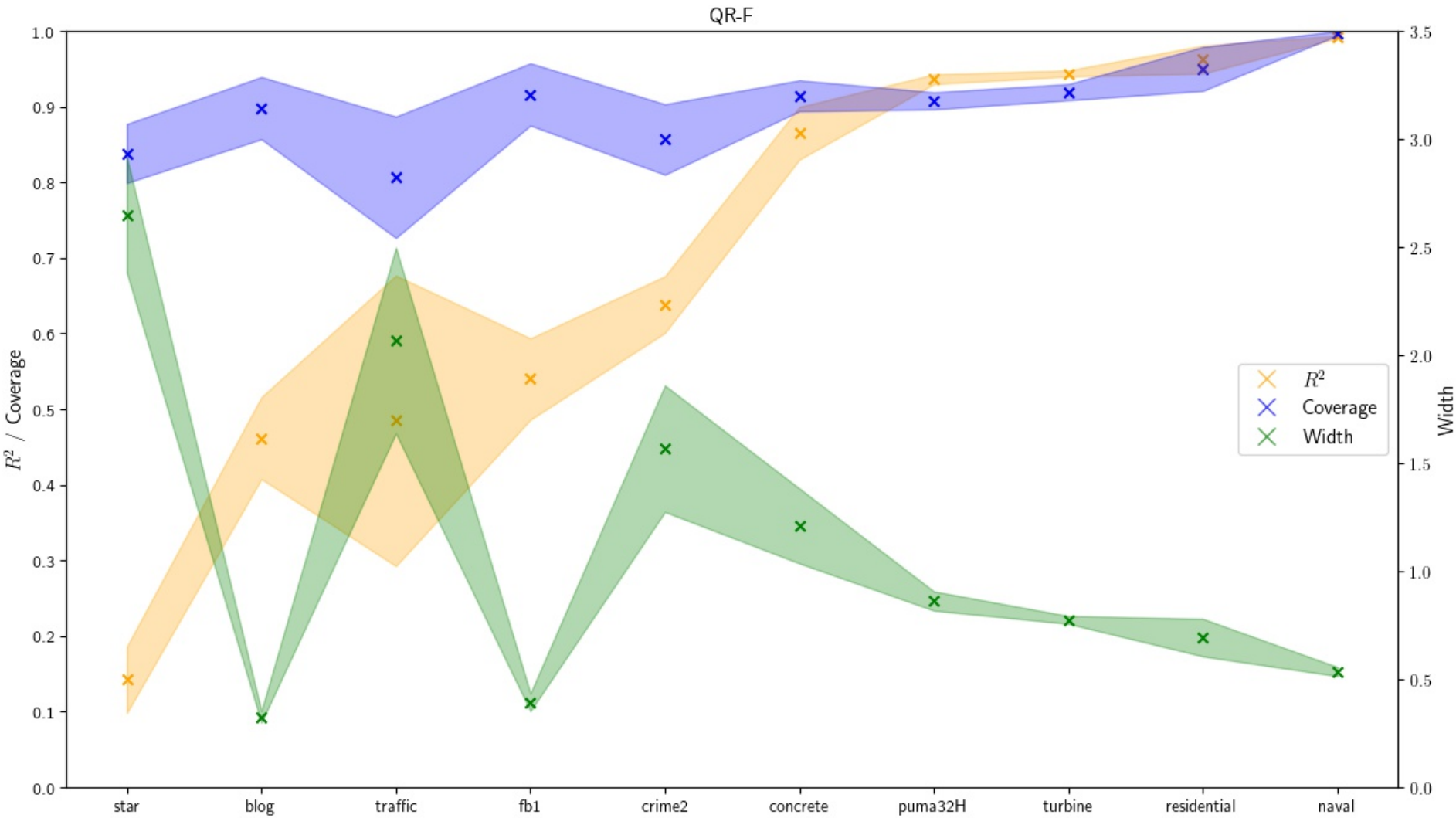
- Calibrate a prediction interval using **nonconformity scores** computed on a **validation** dataset $\mathcal{V}$
- For a **point predictor**, consider the following nonconformity score:

$$a_i := |\rho(\mathbf{x}_i) - y_i|$$

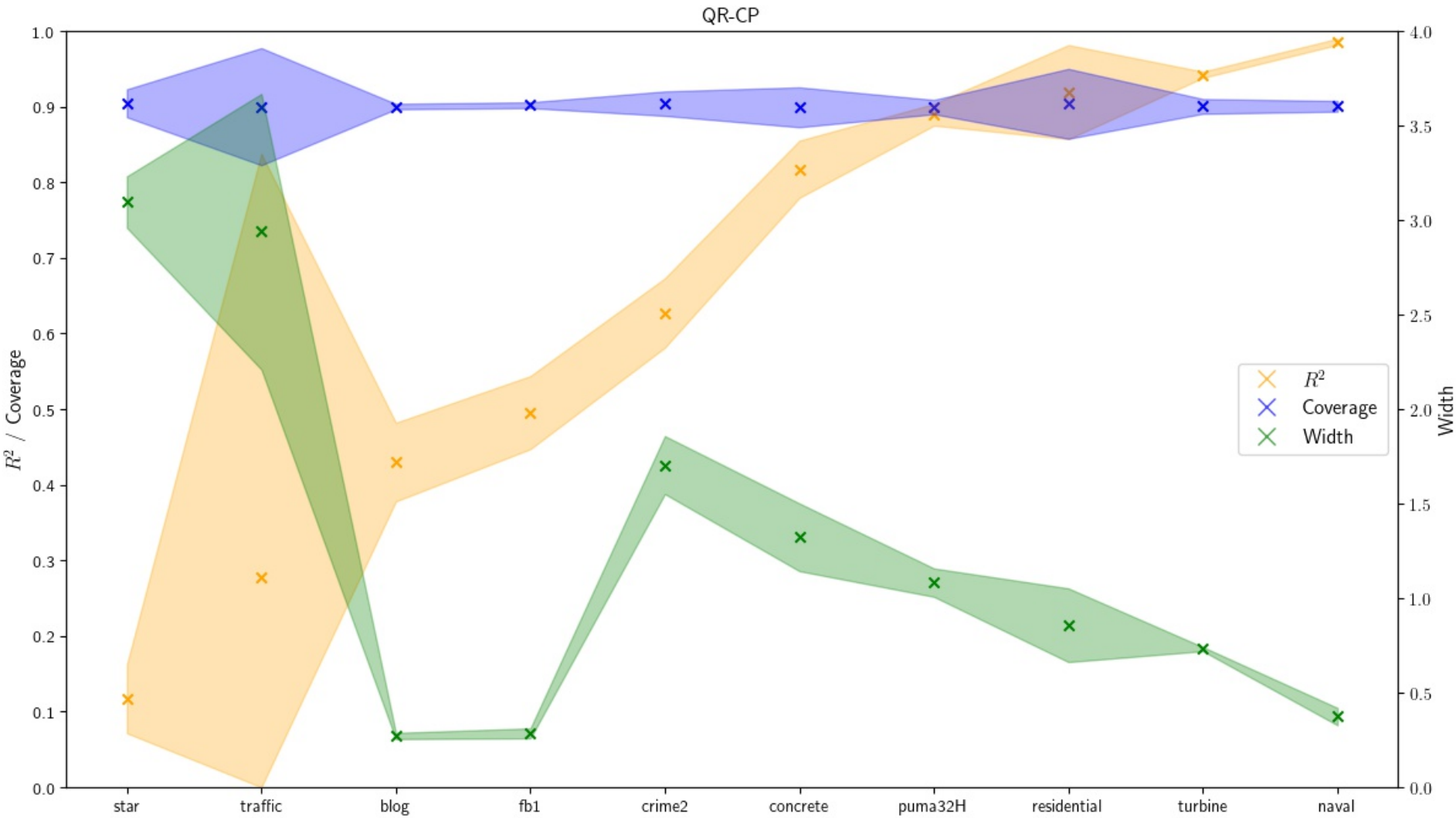
- Compute the $\left((1-\alpha)(1+\frac{1}{|\mathcal{V}|})\right)$-**quantile** of $a_i \in \mathcal{V}$
- For a **test instance** $\mathbf{x}^*$, the interval contains all $y^* \in \mathbb{R}$ for which the nonconformity score is smaller than that quantile

Theorem – Asymptotic validity
If the nonconformity scores in $\mathcal{V} \cup \{\alpha^*\}$ are **exchangeable**, then the interval becomes valid when $|\mathcal{V}| \to \infty$.

Results on 10 datasets (Quantile regression)



Results on 10 datasets (Conformalized quantile regression)



Summary

Method	Marginal validity	Scalability	Domain knowledge	Validation set
Bayesian methods	No (except for exact inference with correct priors)	Only scalable with approximate inference	Yes	No
Ensemble methods	No	Yes (when scalable models are used)	No	No
Direct interval estimation	No	Yes	No	Yes
Conformal prediction	Yes	Yes (for ICP)	No	Yes

Ongoing work: Extensions for time series forecasting

Problem formulation:

- $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3)$ is observed in **sequential** order
- **One-step ahead** forecasting: at time stamp t, predict y_{t+1} using $\mathbf{x}_{t+1}$ and data of previous time stamps

$$y_{t+1} = \rho(\mathbf{x}_{t+1}, \mathbf{x}_t, y_t, \mathbf{x}_{t-1}, y_{t-1}, ..., \mathbf{x}_{t-s}, y_{t-s}) + \epsilon_{t+1}$$

Ongoing work: Extensions for time series forecasting

Problem formulation:

- $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3)$ is observed in **sequential** order
- **One-step ahead** forecasting: at time stamp t, predict y_{t+1} using $\mathbf{x}_{t+1}$ and data of previous time stamps

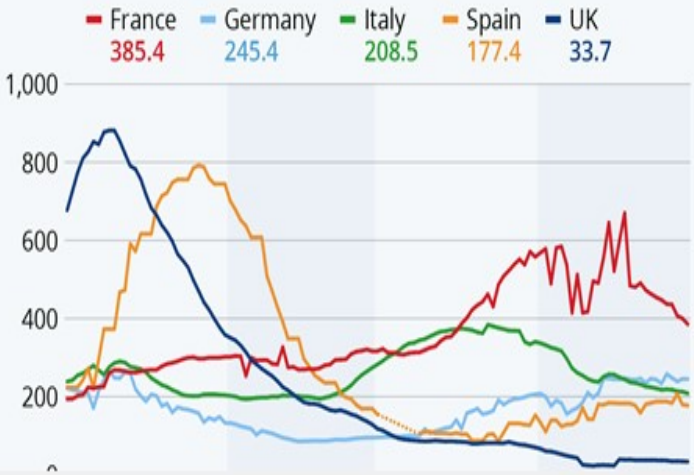
$$y_{t+1} = \rho(\mathbf{x}_{t+1}, \mathbf{x}_t, y_t, \mathbf{x}_{t-1}, y_{t-1}, ..., \mathbf{x}_{t-s}, y_{t-s}) + \epsilon_{t+1}$$

Three challenges:

- **Autocorrelation**: easy to handle
- **Heteroscedasticity**: more difficult to handle
- **Non-stationarity**: very difficult to handle

The State of the Third Wave in Europe

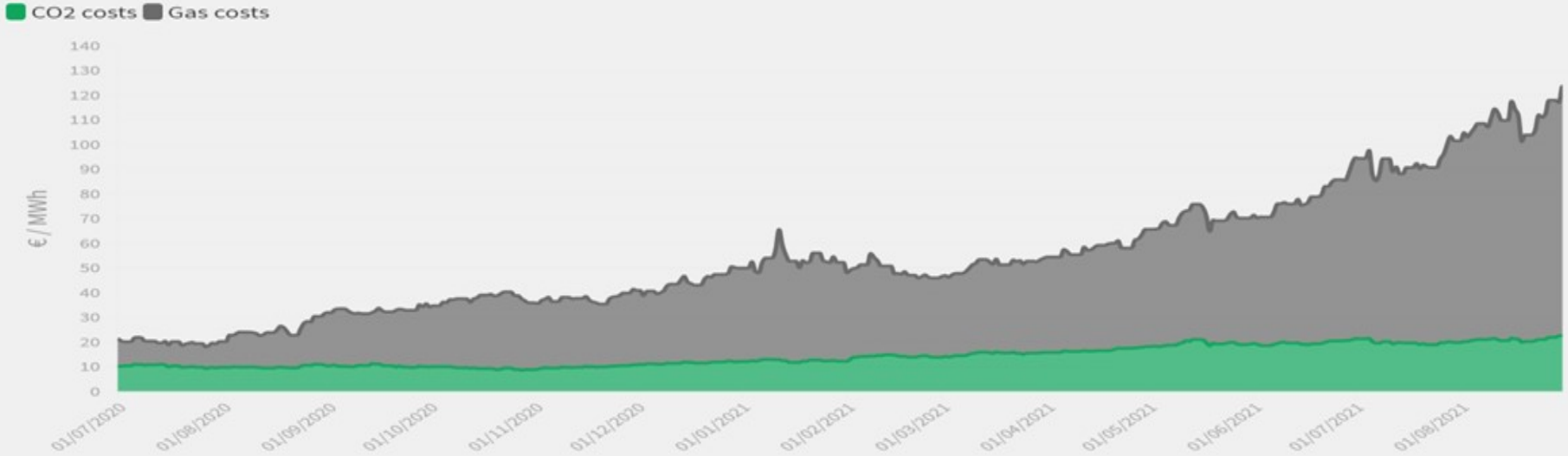
7-day rolling average of daily new COVID-19 cases per million of population in selected European countries (2021)



Skyrocketing fossil gas prices push up cost of EU electricity

EMBER

Fossil gas costs vs. carbon costs for EU electricity generation from combined cycle gas turbines



Source: Powernext for TTF fossil gas prices (day ahead), EEX for EU-ETS carbon prices (December contract)
Costs calculated using emissions intensity of 0.37 tCO2eq / MWh and plant efficiency rate of 55%

Problems with classical methods for time series forecasting

Bayesian methods:

- The residuals ϵ_{t+1} often do not follow a Gaussian distribution

Problems with classical methods for time series forecasting

Bayesian methods:

- The residuals ϵ_{t+1} often do not follow a Gaussian distribution

Ensemble methods:

- Already poor performance for classical regression problems...

Problems with classical methods for time series forecasting

Bayesian methods:

- The residuals ϵ_{t+1} often do not follow a Gaussian distribution

Ensemble methods:

- Already poor performance for classical regression problems...

Direct interval estimation methods:

- Handling autocorrelation via autoregressive features
- Immediate control for heteroscedasticity
- Handling non-stationarity via online learning, e.g. RNNs, LSTMs

Problems with classical methods for time series forecasting

Bayesian methods:

- The residuals ϵ_{t+1} often do not follow a Gaussian distribution

Ensemble methods:

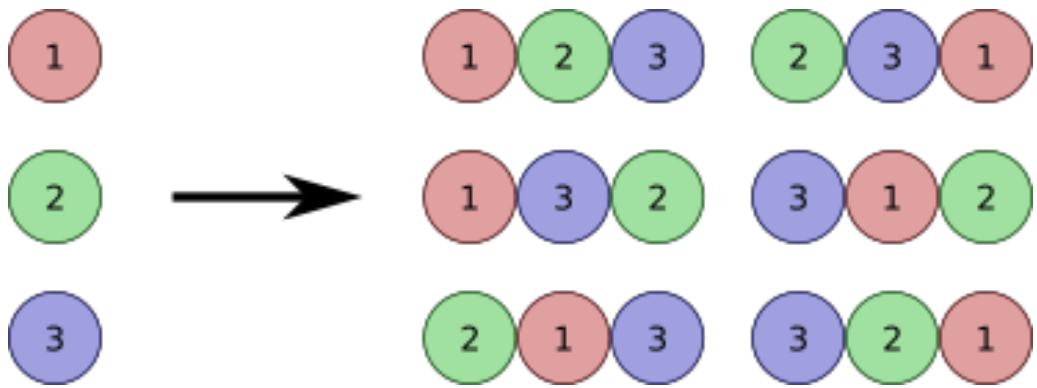
- Already poor performance for classical regression problems...

Direct interval estimation methods:

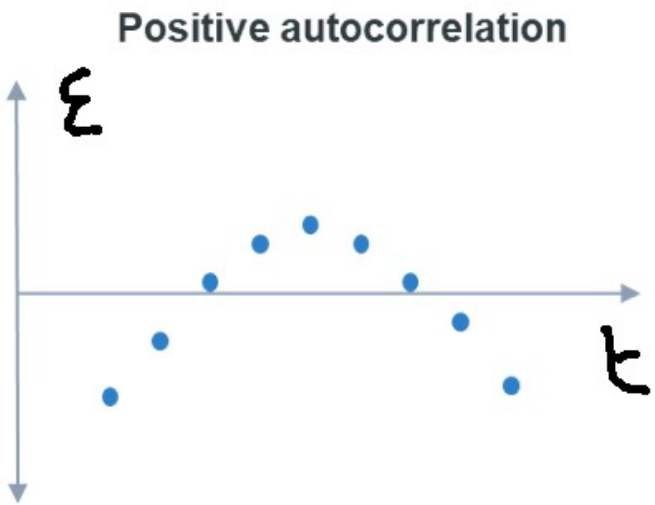
- Handling autocorrelation via autoregressive features
- Immediate control for heteroscedasticity
- Handling non-stationarity via online learning, e.g. RNNs, LSTMs

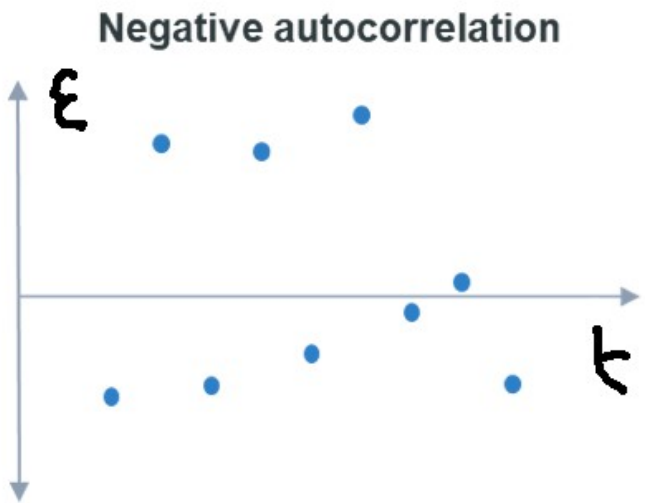
Conformal prediction:

- No theoretical guarantees because the sequence $\epsilon_{t+1}, \epsilon_{t+2}, \epsilon_{t+3}, \ldots$ is not exchangeable

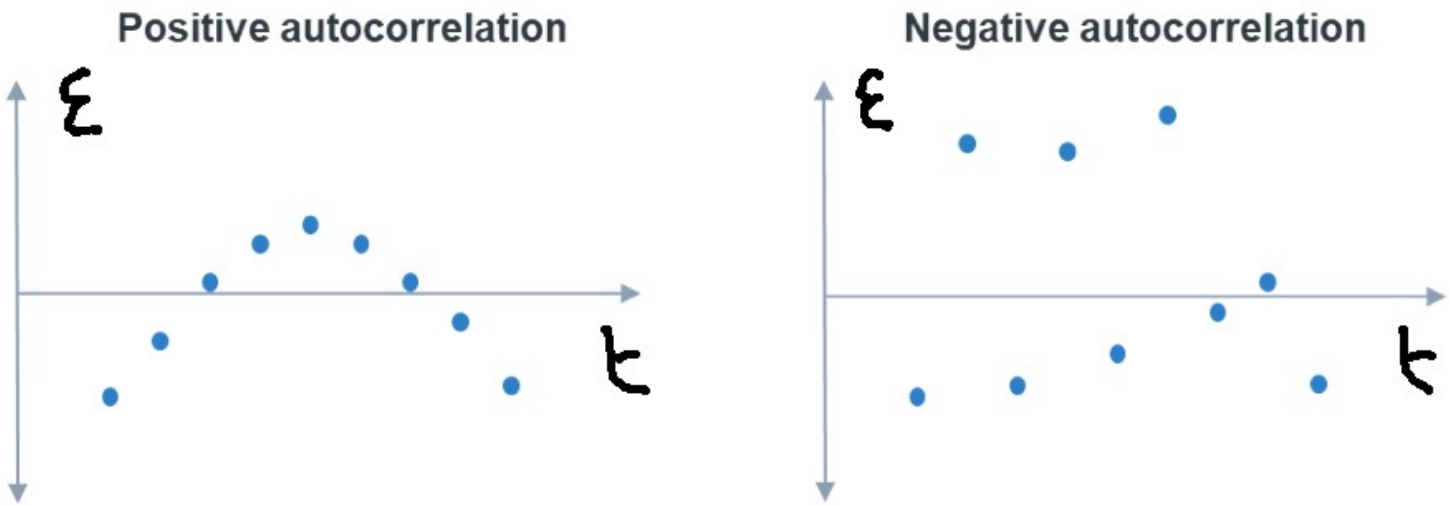


How to handle autocorrelation?





How to handle autocorrelation?

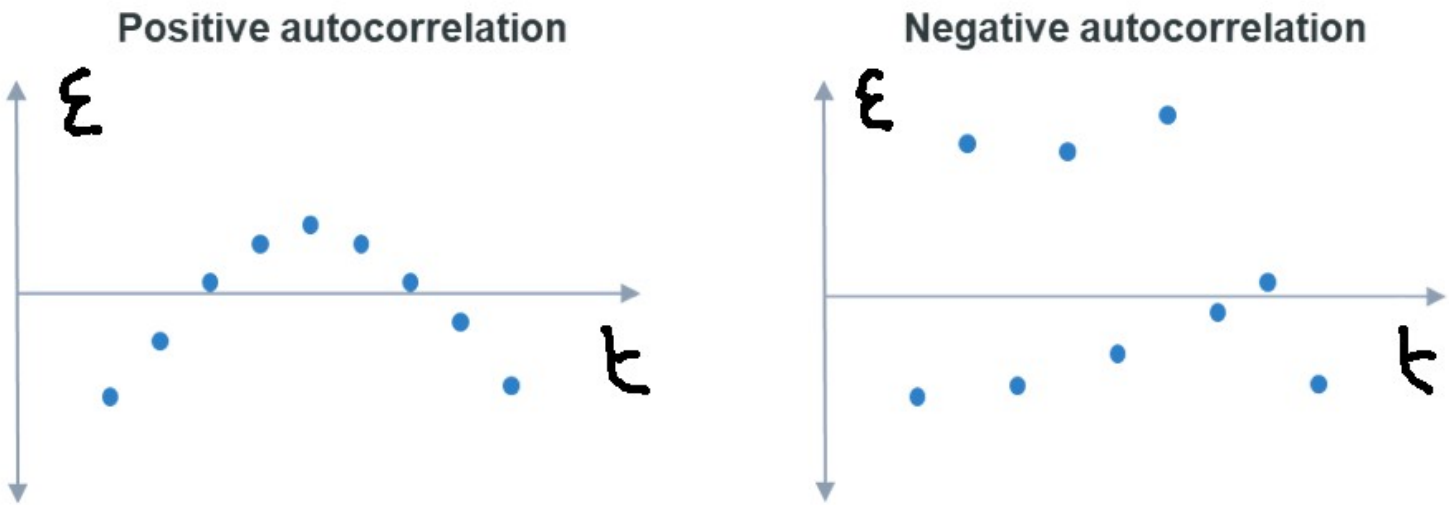


Estimate ρ as an **autoregressive** function

$$y_{t+1} = \hat{\rho}(\mathbf{x}_{t+1}, \mathbf{x}_t, y_t, \mathbf{x}_{t-1}, y_{t-1}, ..., \mathbf{x}_{t-s}, y_{t-s}) + \epsilon_{t+1}$$

such that the residuals ϵ_{t+1},ϵ_{t+2},ϵ_{t+3}, ... are (hopefully) **i.i.d.**

How to handle autocorrelation?



Estimate ρ as an **autoregressive** function

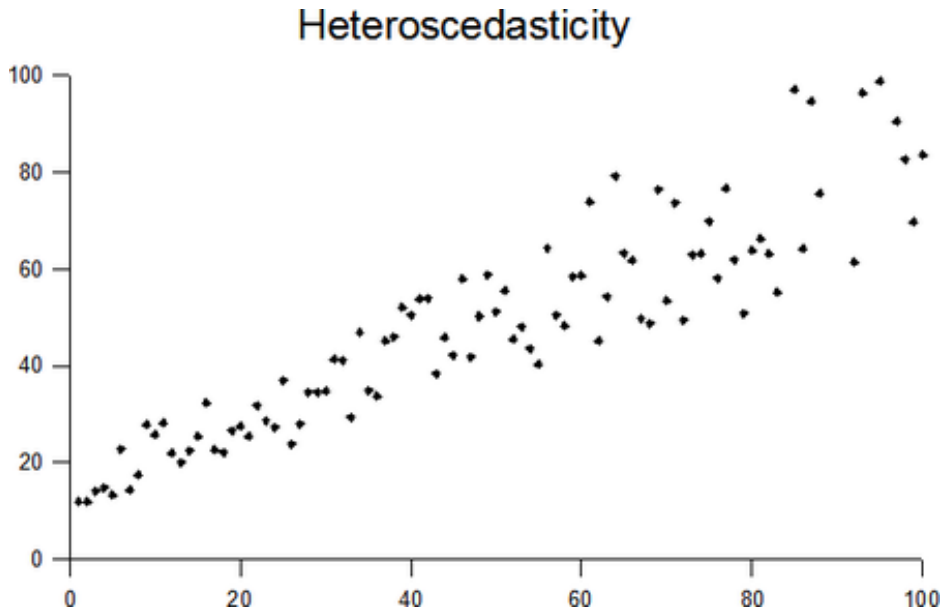
$$y_{t+1} = \hat{\rho}(\mathbf{x}_{t+1}, \mathbf{x}_t, y_t, \mathbf{x}_{t-1}, y_{t-1}, ..., \mathbf{x}_{t-s}, y_{t-s}) + \epsilon_{t+1}$$

such that the residuals ϵ_{t+1},ϵ_{t+2},ϵ_{t+3}, ... are (hopefully) **i.i.d.**

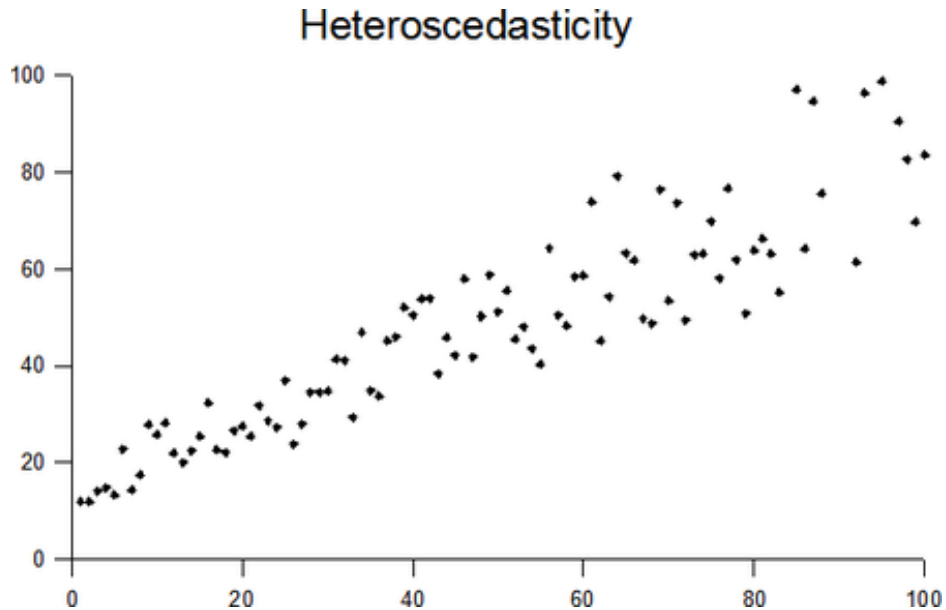
Remark: the above approach will not work when the underlying data is generated according to

$$y_{t+1} = \rho(\mathbf{x}_{t+1}, \mathbf{x}_t, y_t, \mathbf{x}_{t-1}, y_{t-1}, ..., \mathbf{x}_{t-s}, , y_{t-s}, \epsilon_t) + \epsilon_{t+1}$$

How to handle heteroscedasticity?



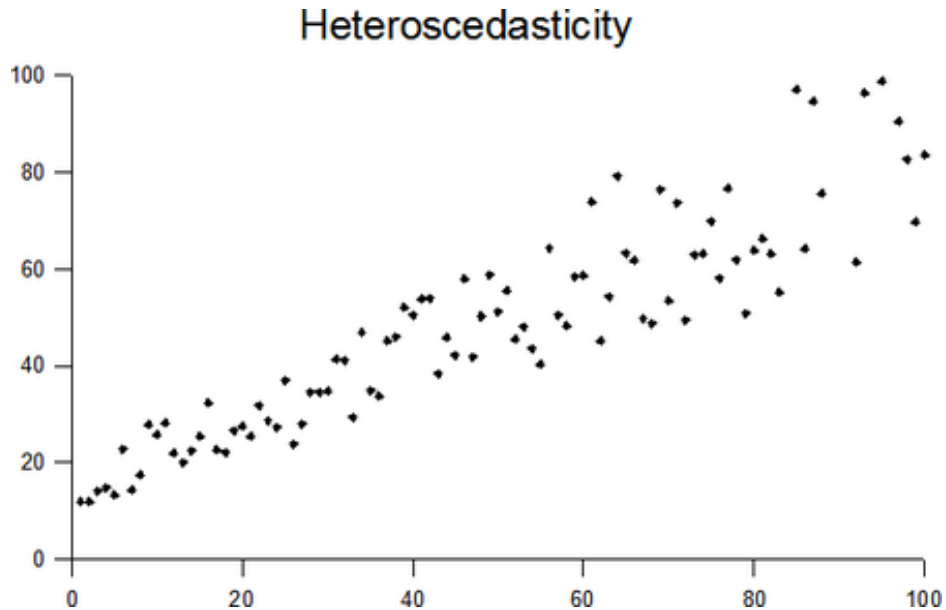
How to handle heteroscedasticity?



Problem for inductive conformal prediction:

- Conditional coverage for $Y = y$ cannot be guaranteed

How to handle heteroscedasticity?



Problem for inductive conformal prediction:

- Conditional coverage for $Y = y$ cannot be guaranteed

Solutions:

- Variance-stabilizing transformations
- Normalized conformal prediction via adjusted nonconformity score

$$a_i = \frac{|\rho(\mathbf{x}_i) - y_i|}{\sigma_i}$$

- Box-Cox transformation:

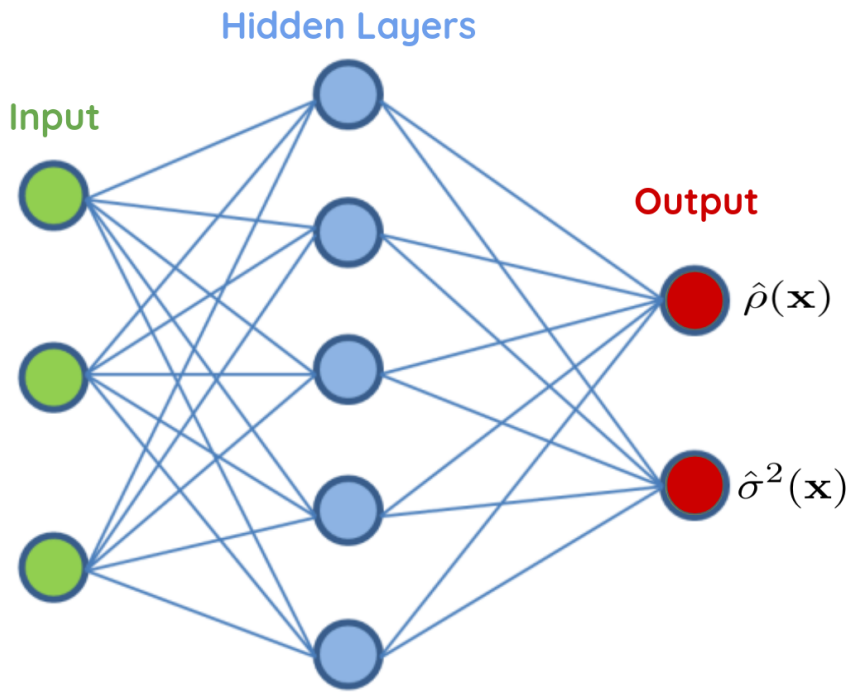
$$y'_t = \begin{cases} \frac{(y_t+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \text{if } \lambda_1 \neq 0 \\ \ln(y_t+\lambda_2) & \text{if } \lambda_1 = 0 \end{cases}$$

- Box-Cox transformation:

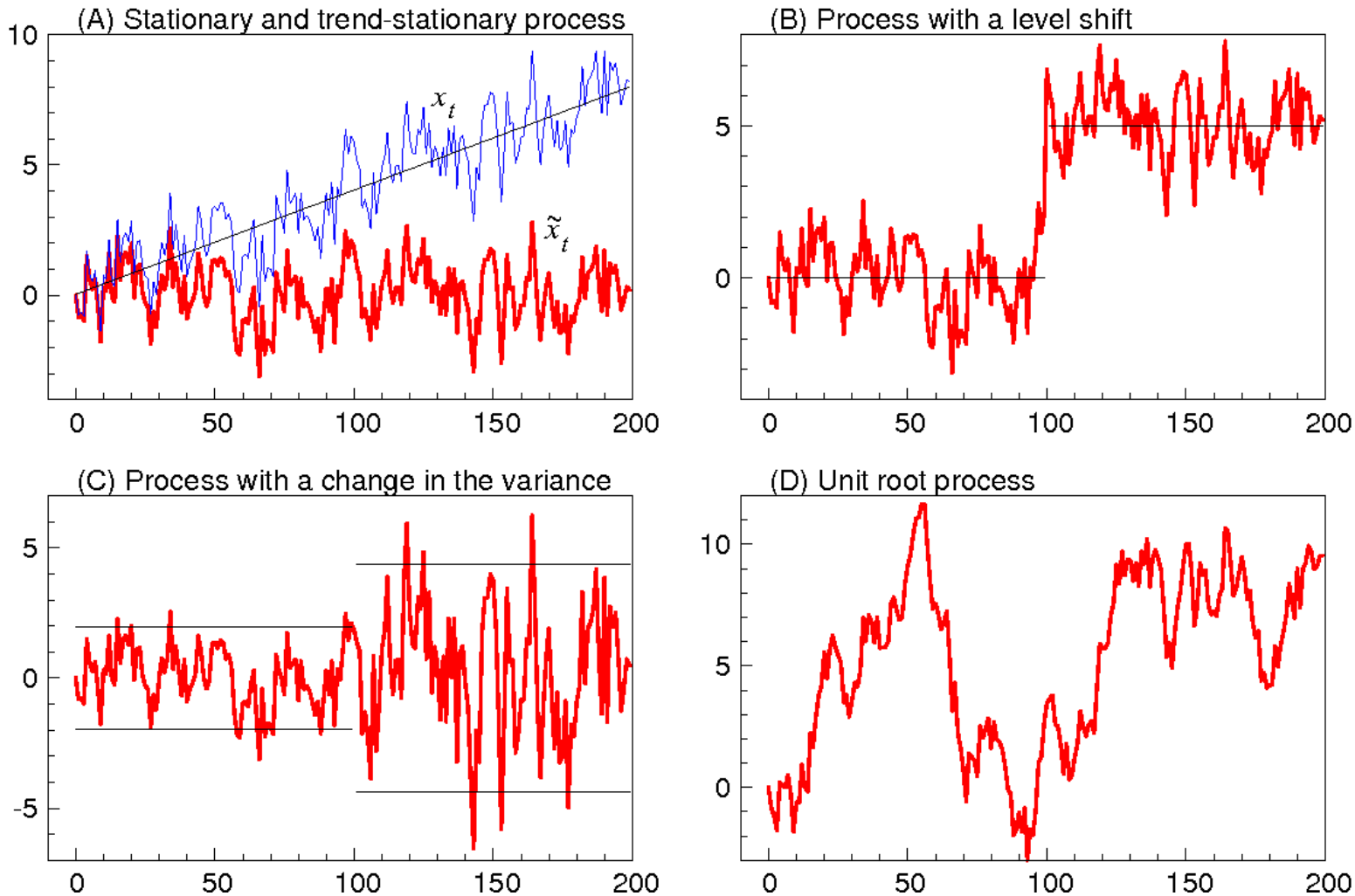
$$y'_t = \begin{cases} \frac{(y_t+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \text{if } \lambda_1 \neq 0 \\ \ln(y_t+\lambda_2) & \text{if } \lambda_1 = 0 \end{cases}$$

- Mean-variance estimation:

$$\mathcal{L}_{\text{Gauss}}(\mathcal{T}) := \sum_{(\mathbf{x},y)\in\mathcal{T}} \frac{(y-\hat{\rho}(\mathbf{x}))^2}{2\hat{\sigma}^2(\mathbf{x})} + \frac{1}{2}\log\hat{\sigma}^2(\mathbf{x})$$



How to handle non-stationarity?



Solutions:

- Transformations that make the process stationary
- Make conformal prediction robust against (small) shifts

Robust conformal prediction (Cauchois et al. Arxiv 2020)

- **Calibration set $\mathcal{C}$:** nonconformity scores sampled from distribution P.

Robust conformal prediction (Cauchois et al. Arxiv 2020)

- **Calibration set $\mathcal{C}$:** nonconformity scores sampled from distribution P.
- **Test set $\mathcal{D}$:** nonconformity scores sampled from some distribution P' such that

$$D_f(P' \| P) \leq \chi$$

for some value χ.

Robust conformal prediction (Cauchois et al. Arxiv 2020)

- **Calibration set $\mathcal{C}$:** nonconformity scores sampled from distribution P.
- **Test set $\mathcal{D}$:** nonconformity scores sampled from some distribution P' such that

$$D_f(P' \| P) \leq \chi$$

  for some value χ.
- Standard CP on $\mathcal{C}$ with **modified significance level** $\overline{\alpha}$ will give valid intervals for $\mathcal{D}$.

Robust conformal prediction (Cauchois et al. Arxiv 2020)

- **Calibration set $\mathcal{C}$:** nonconformity scores sampled from distribution P.
- **Test set $\mathcal{D}$:** nonconformity scores sampled from some distribution P' such that

$$D_f(P' \| P) \leq \chi$$

  for some value χ.
- Standard CP on $\mathcal{C}$ with **modified significance level** $\overline{\alpha}$ will give valid intervals for $\mathcal{D}$.

$$\overline{\alpha} := 1 - g\left((1 + 1/n)g^{-1}(1 - \alpha)\right),$$

$$g(\beta) = \inf\left\{\alpha \in [0,1] \,\middle|\, \beta f\left(\frac{\alpha}{\beta}\right) + (1-\beta) f\left(\frac{1-\alpha}{1-\beta}\right) \leq \chi\right\}$$

$$g^{-1}(\alpha) = \sup\left\{\beta \in [\alpha,1] \,\middle|\, \beta f\left(\frac{\alpha}{\beta}\right) + (1-\beta) f\left(\frac{1-\alpha}{1-\beta}\right) \leq \chi\right\}.$$

Conclusions

Classical regression problem:

- Many methods have been proposed, but very few give statistical guarantees in the distribution-free setting
- Conformal prediction methods have become popular, and in general do what they promise

Conclusions

Classical regression problem:

- Many methods have been proposed, but very few give statistical guarantees in the distribution-free setting
- Conformal prediction methods have become popular, and in general do what they promise

Time series forecasting:

- More challenging due to autocorrelation, heteroscedasticity, and non-stationarity
- No guarantees possible when the distribution changes too quickly
- At this moment unclear which method works best in the distribution-free setting