Design-marginal calibration of gaussian process predictive distributions: bayesian and conformal approaches

Aurélien Pion [1] [2]
Joint work with PhD Supervisor: Emmanuel Vazquez[2]

[1] TRANSVALOR S.A.
[2] Laboratoire des Signaux et Systèmes,
L2S – CNRS – CentraleSupélec – Université Paris-Saclay

March 31, 2026

Introduction

Prediction setting

➜ Unknown **deterministic** function $f : \mathbb{R}^d \to \mathbb{R}$ observed through pointwise evaluations

$$\mathcal{D}_n = \{(X_i, Z_i),\ Z_i = f(X_i),\ i = 1, \ldots, n\}$$

➜ The design points are sampled from a **design measure** μ on the input space

$$X_i \sim \mu$$

➜ A prediction method returns, at each x, a **predictive CDF**

$$\hat{F}_n(\cdot \mid x)$$

Introduction

What does this predictive CDF mean?

➜ It is a **cumulative distribution function** in the variable z, that is, in the output value

$$z \longmapsto \hat{F}_n(z \mid x)$$

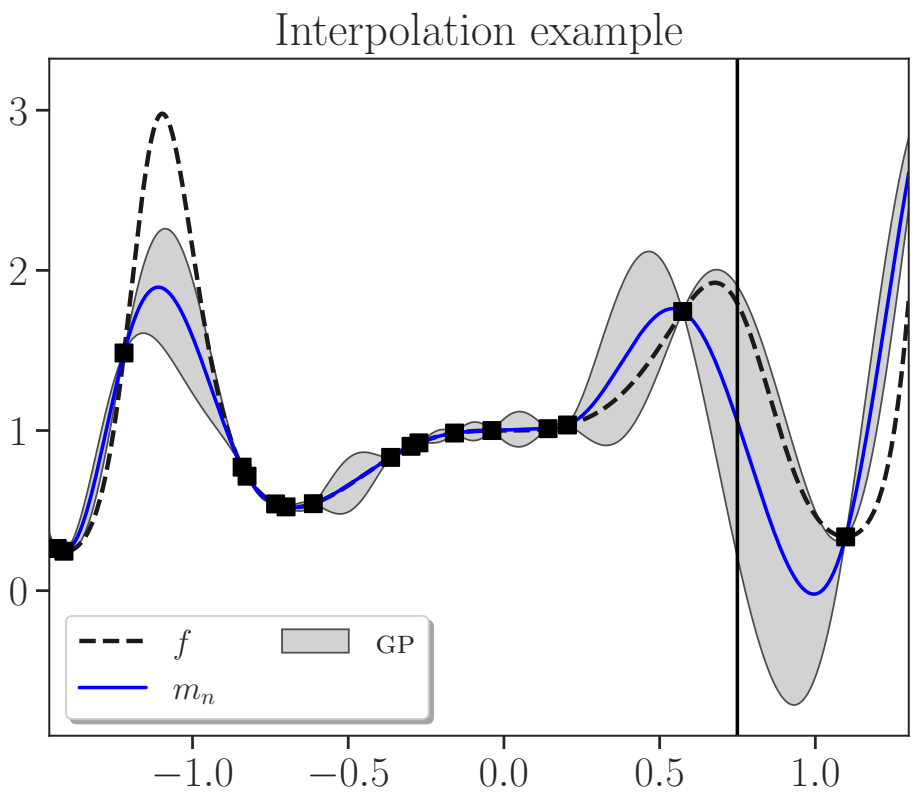
➜ It summarizes our uncertainty about the unknown value $f(x)$

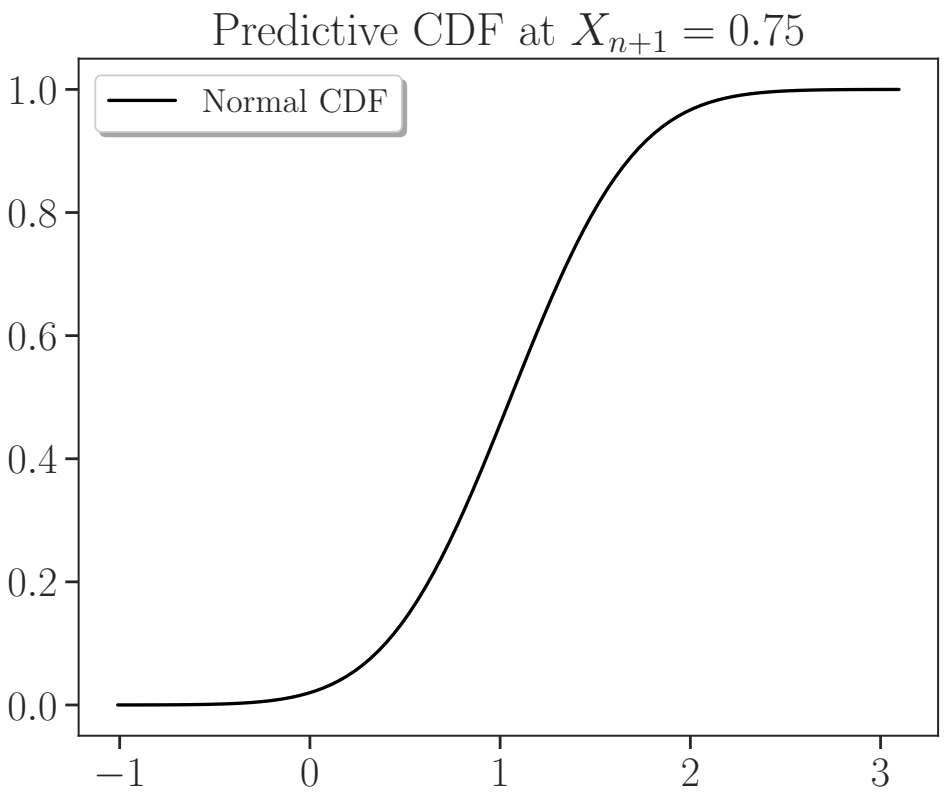
➜ A classical route is Gaussian process (GP) approximation: at each x, it yields a Gaussian predictive CDF

$$\hat{F}_n(\cdot \mid x) = \mathcal{N}\left(m_n(x), \sigma_n^2(x)\right)$$

($m_n(x)$ is the posterior mean, $\sigma_n^2(x)$ the posterior variance)

Example: GP predictions and predictive CDF at $x = 0.75$





Introduction

➜ **How should we validate $\hat{F}_n$?**

➜ Is it accurate as a predictor, and reliable as a quantification of uncertainty?

➜ First requirement: good **point prediction**

➜ For instance, using the predictive mean $\hat{m}_n(x)$ and a test set

$$\mathcal{D}_m^\star = \{(X_j^\star, Z_j^\star)\}_{j=1}^m$$

➜ Classical criterion: **mean squared error**

$$\text{MSE} = \frac{1}{m}\sum_{j=1}^{m}\left(\hat{m}_n(X_j^\star) - Z_j^\star\right)^2$$

➜ **But** MSE does not evaluate the quality of the uncertainty quantification provided by $\hat{F}_n(\cdot \mid x)$

Introduction

Questions

➜ What should it mean for predictive CDFs to be **calibrated** in this deterministic setting?

➜ How can we **assess calibration in practice**, with or without an independent test set?

➜ How can we **improve the calibration** of GP predictive CDFs?

Calibration

Why calibration must involve μ

➜ After observing $\mathcal{D}_n$, the function f is fixed

➜ At fixed x, the value $f(x)$ is deterministic

➜ The only remaining randomness comes from

$$X \sim \mu$$

➜ So calibration must be defined after averaging over inputs drawn from μ

➜ This is **design-marginal** calibration, or **μ-calibration**

➜ We define **two notions of μ-calibration**

➜ Both are **spatial** properties: they concern the distribution of $f(X)$ under $X \sim \mu$

Calibration

μ-coverage calibration

➜ For prediction intervals $\mathcal{C}_{n,1-\alpha}(x)$, define **$\mu$-coverage** by

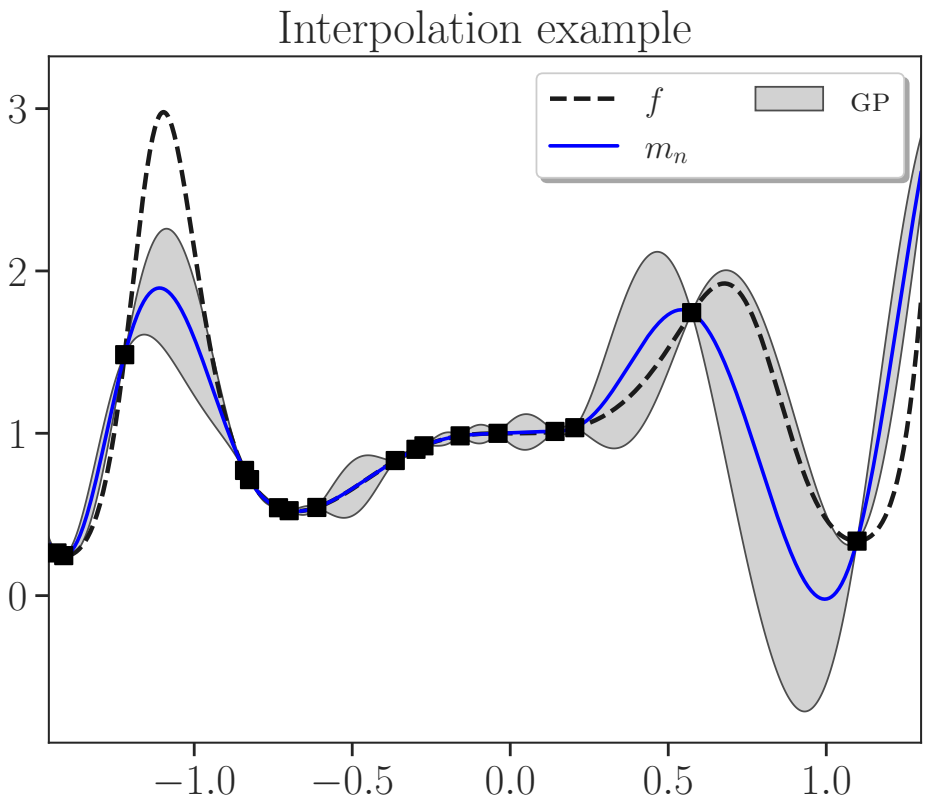
$$\delta_\alpha(\hat{F}_n; \mu) = \mu\big(\{x : \ f(x) \in \mathcal{C}_{n,1-\alpha}(x)\}\big)$$

➜ Calibration at level $1 - \alpha$ means

$$\delta_\alpha(\hat{F}_n; \mu) = 1 - \alpha$$

Calibration

Example: for $\alpha = 0.1$, $\delta_\alpha(\hat{F}_n; \mu) \approx 0.73 < 1 - \alpha = 0.9$



Calibration

μ-probabilistic calibration

➜ For a random variable Z with continuous CDF F, the **probability integral transform (PIT)** [Dawid, 1984, Gneiting and Resin, 2023]

$$F(Z),$$

is uniformly distributed on $[0, 1]$

➜ In probabilistic forecasting, this idea is used to formalize calibration from repeated forecast–observation pairs [Gneiting and Resin, 2023]

➜ In our setting, the relevant random quantity is

$$f(X), \qquad X \sim \mu$$

➜ Given a family of predictive CDFs $\hat{F}_n(\cdot \mid x)$, this leads to the **μ-PIT**

$$U_{\hat{F}_n} = \hat{F}_n(f(X) \mid X)$$

Calibration

μ-probabilistic calibration

➜ We say that the predictive CDFs are **μ-probabilistically calibrated** if

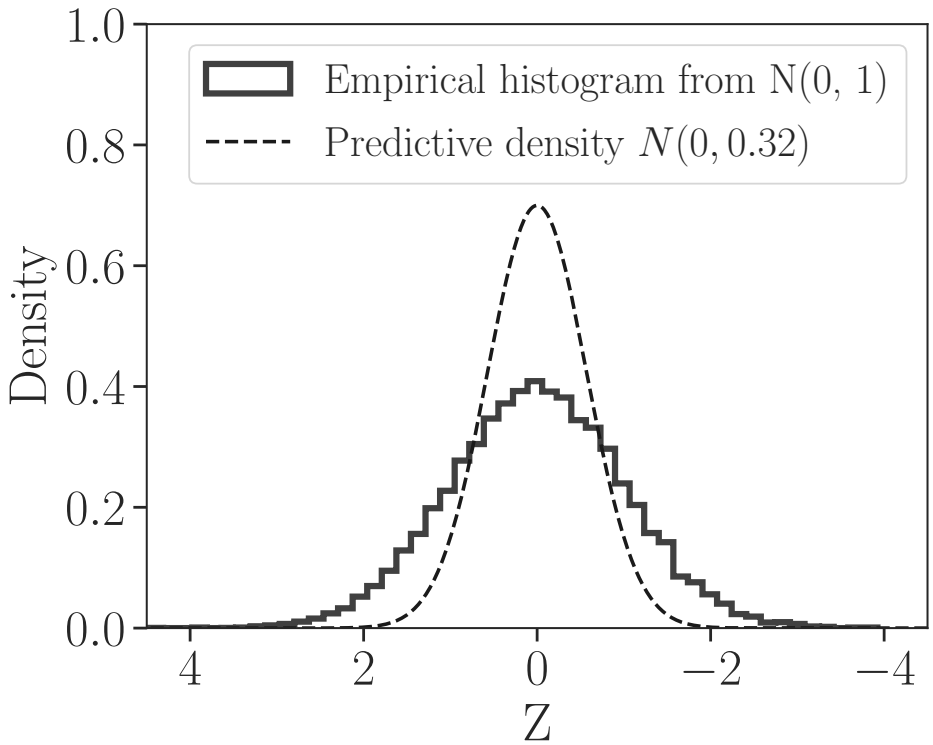
$$U_{\hat{F}_n} \sim \mathcal{U}(0,1)$$

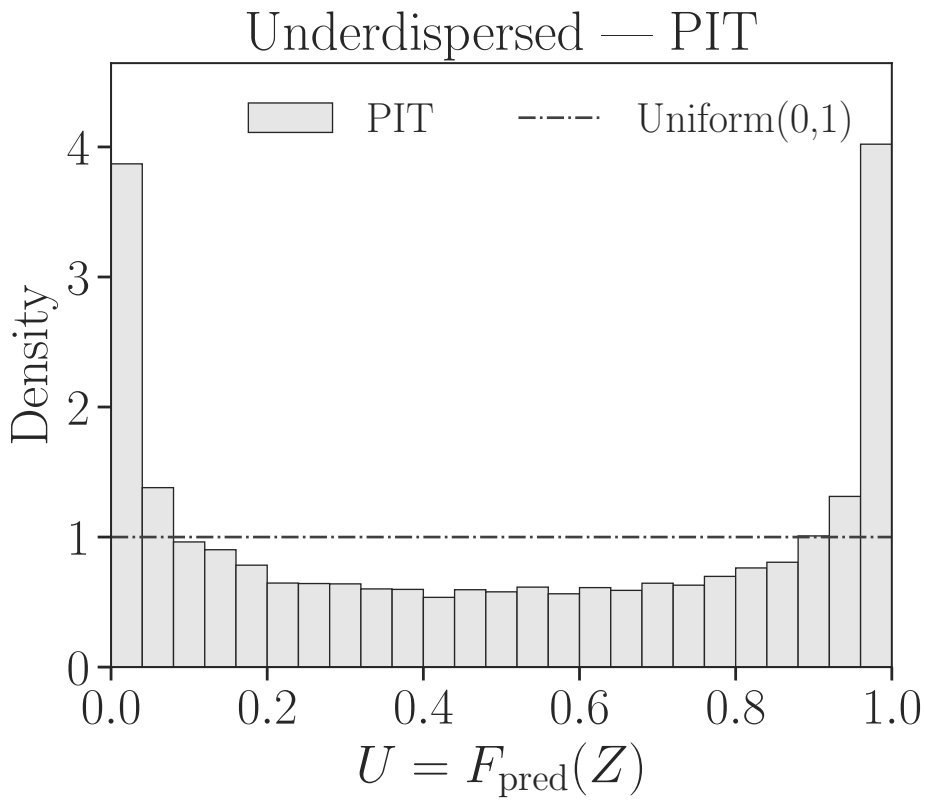
➜ This is the design-marginal analogue of the classical notion of *probabilistic calibration* [Gneiting and Resin, 2023]

➜ It is a global distributional notion of calibration, beyond coverage at a fixed prediction-interval level

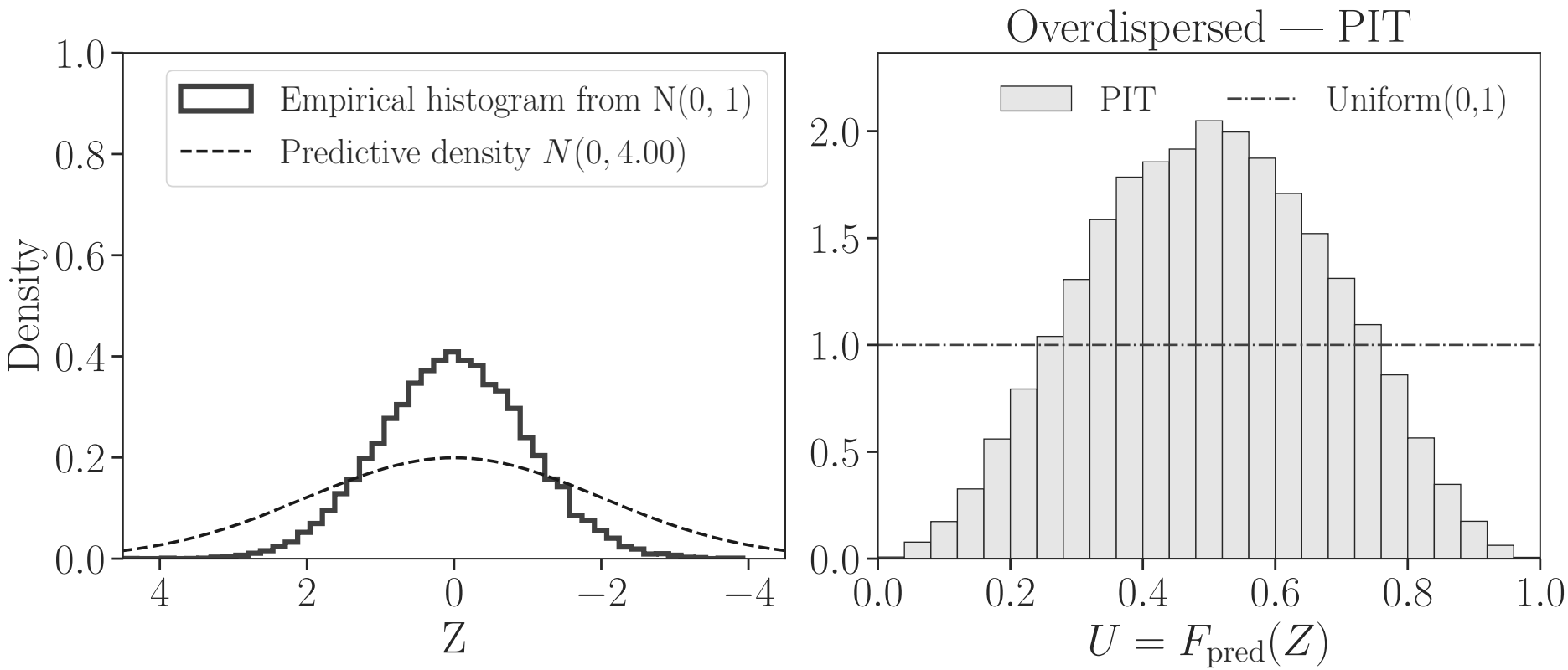
Calibration

PIT diagnostics: under-dispersion (optimistic prediction)





PIT diagnostics: over-dispersion (pessimistic prediction)



Calibration

Ideal situation: an independent test set

➜ Suppose we have test inputs drawn from the same design measure

$$X_j^\star \sim \mu, \qquad j = 1, \dots, m$$

➜ Because μ-calibration is a spatial property, it can be **estimated and tested** on such a test set

➜ Then we can assess:

- point prediction, for instance through MSE
- **scoring rules** such as NLPD, CRPS [Gneiting and Raftery, 2007] or SCRPS [Bolin and Wallin, 2023]; see also [Petit et al., 2023]
- **μ-coverage** at several levels with IAE [Marrel and Iooss, 2024]
- **μ-PIT** and its distance to uniformity [Diebold et al., 1998]
- ...

Calibration

Cross-validation and leave-one-out

➜ No test set in practice

➜ Cross-validation provides **pseudo-test predictive CDFs** from the available data

➜ In leave-one-out (LOO), for each i, we remove (X_i, Z_i) and construct

$$\hat{F}_{n,-i}(\cdot \mid X_i)$$

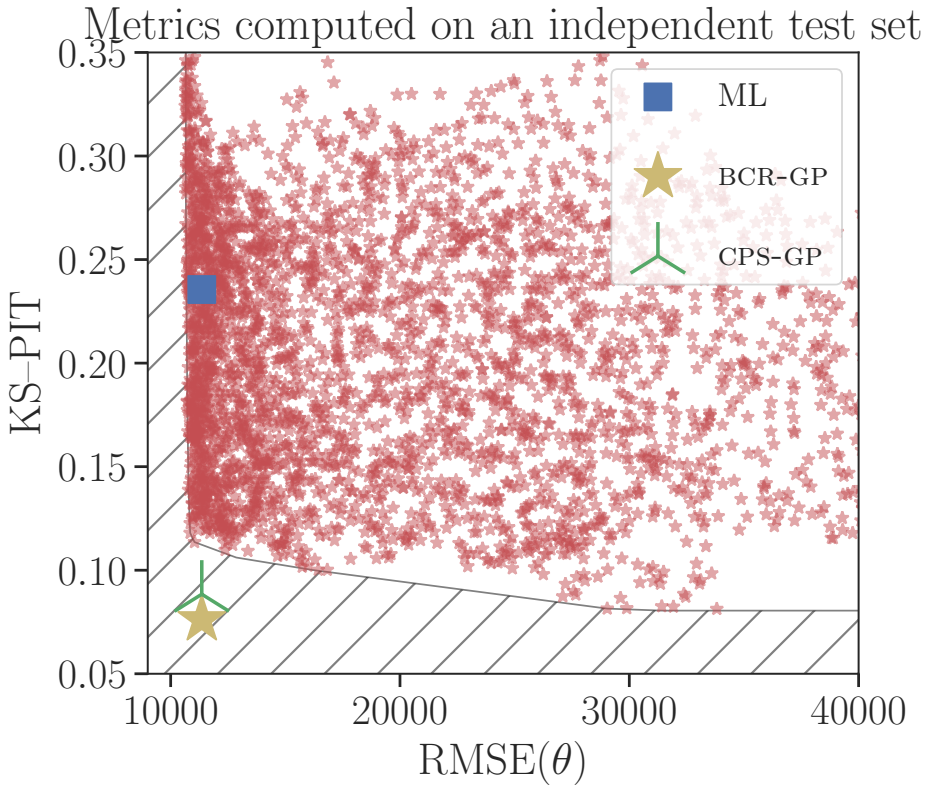
➜ These LOO predictive CDFs play the role of **pseudo-test distributions** at the observed inputs

➜ In practice, these **LOO distributions** are used to: (i) assess a default of calibration, (ii) correct this default using a **post-hoc calibration method**

Post-hoc calibration of GP

Why post-hoc calibration?

➜ KS–PIT (distance of the PIT to uniformity) vs RMSE for GP predictive CDFs



Post-hoc calibration of GP

Two post-hoc calibration approaches

➜ We propose **two post-hoc calibration methods** for GP predictive CDFs

➜ Both are built from **LOO predictive distributions and residuals**

➜ The goal is to improve calibration while keeping useful predictive CDFs

➜ The first one is based on **conformal predictive systems**

➜ The second one is a **Bayesian residual-calibration method**

CPS-GP

From CP to CPS

➜ **Conformal prediction (CP)** [Vovk et al., 2005] is a classical method to produce prediction intervals with *finite-sample guarantees* in a **frequentist** sense (Tutorial: [Da Veiga, 2024])

➜ **Conformal predictive systems (CPS)** extend this idea from intervals to **full predictive CDFs** [Vovk et al., 2019]

➜ In our GP setting, the construction is based on **standardized LOO scores**

Standardized LOO scores

➜ Fix a test input x_{n+1} and a candidate value z, and form the augmented dataset

$$\mathcal{D}_{n+1}^{z} = \mathcal{D}_n \cup \{(x_{n+1}, z)\}$$

➜ Test-point score:

$$R_{n+1}^{z} = \frac{z - m_n(x_{n+1})}{\sigma_n(x_{n+1})}$$

➜ For each observed point, the corresponding LOO score is

$$R_i^{z} = \frac{z_i - m_{n+1,-i}(x_i)}{\sigma_{n+1,-i}(x_i)}, \qquad i = 1, \ldots, n$$

➜ These are the **standardized LOO scores**

The object $\pi(z)$

➜ Compare the test score R^z_{n+1} with the other scores through the randomized rank

$$\pi(z) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}\{R^z_i < R^z_{n+1}\} + \frac{\tau}{n+1} \sum_{i=1}^{n+1} \mathbb{1}\{R^z_i = R^z_{n+1}\}$$

➜ Here τ is a random tie-breaker

➜ It is the **randomized rank** of the test score in the augmented sample

➜ The CPS output is the map $z \mapsto \pi(z)$ acts as a **predictive CDF proxy**

Main idea

➜ We want to model the distribution of the normalized prediction error

$$R_n(X_{n+1}, Z_{n+1}) = \frac{Z_{n+1} - m_n(X_{n+1})}{\sigma_n(X_{n+1})}$$

when $X_{n+1} \sim \mu$

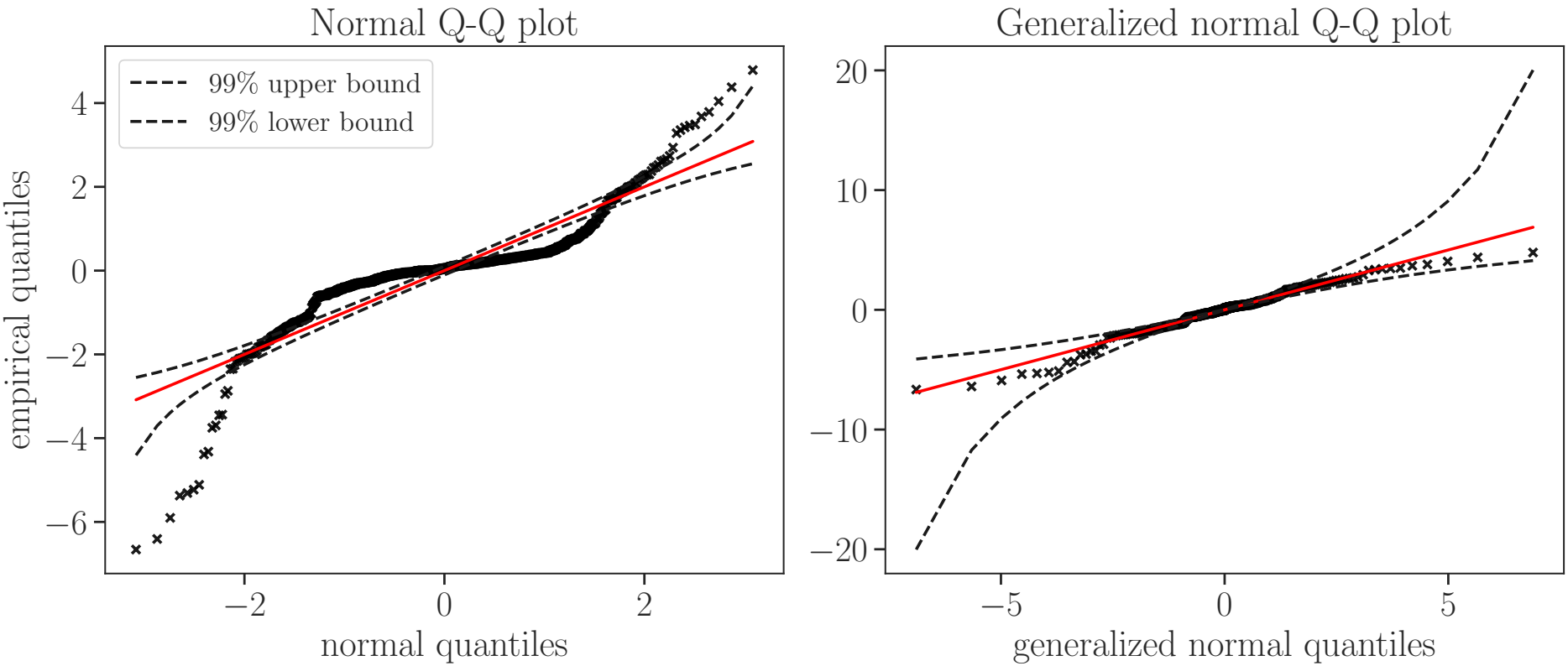
➜ In practice, we approximate this distribution by the standardized **LOO residuals**

Residual model

➜ We then fit a generalized normal distribution

$$\mathcal{GN}(\beta, 0, \lambda)$$

to capture both dispersion and tail behavior



Posterior uncertainty on residual dispersion

➜ From the LOO residuals, we build a posterior distribution

$$p(\beta, \lambda \mid R_{n,-1:n})$$

for the residual-distribution parameters

➜ Each posterior draw (β, λ) yields one candidate residual variance, hence one candidate dispersion for the predictive distribution

$$v(\beta, \lambda) = \mathrm{Var}(\mathcal{GN}(\beta, 0, \lambda))$$

➜ The posterior sample of these variances represents our uncertainty about the true residual dispersion

Choosing a conservative predictive distribution

➜ The posterior gives many plausible residual distributions, but for prediction we need to choose **one** distribution

➜ Under-dispersion is risky: prediction intervals become too narrow

➜ So we choose a distribution that is **slightly more dispersed** than most posterior-plausible ones

➜ Concretely, we choose (β^*, λ^*) so that

$$v^* = \mathrm{Var}(\mathcal{GN}(\beta^*, 0, \lambda^*))$$

is the $(1-\delta)$**-quantile** of the posterior variance distribution

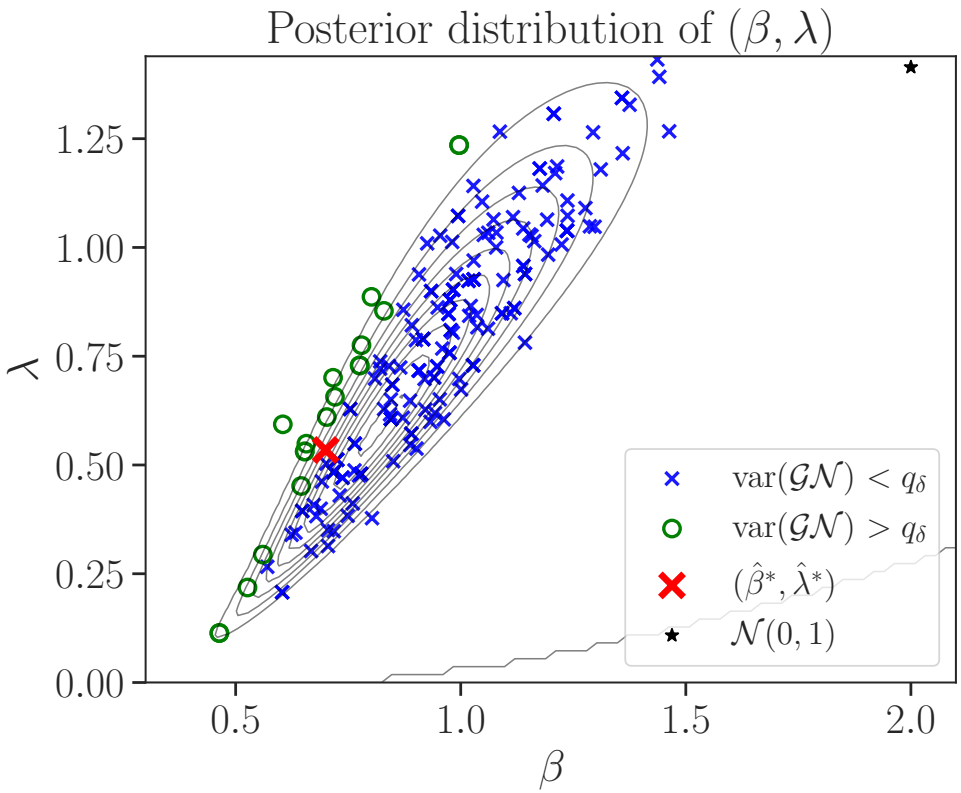
➜ With posterior probability at least $1-\delta$, the true residual variance is no larger than v^*

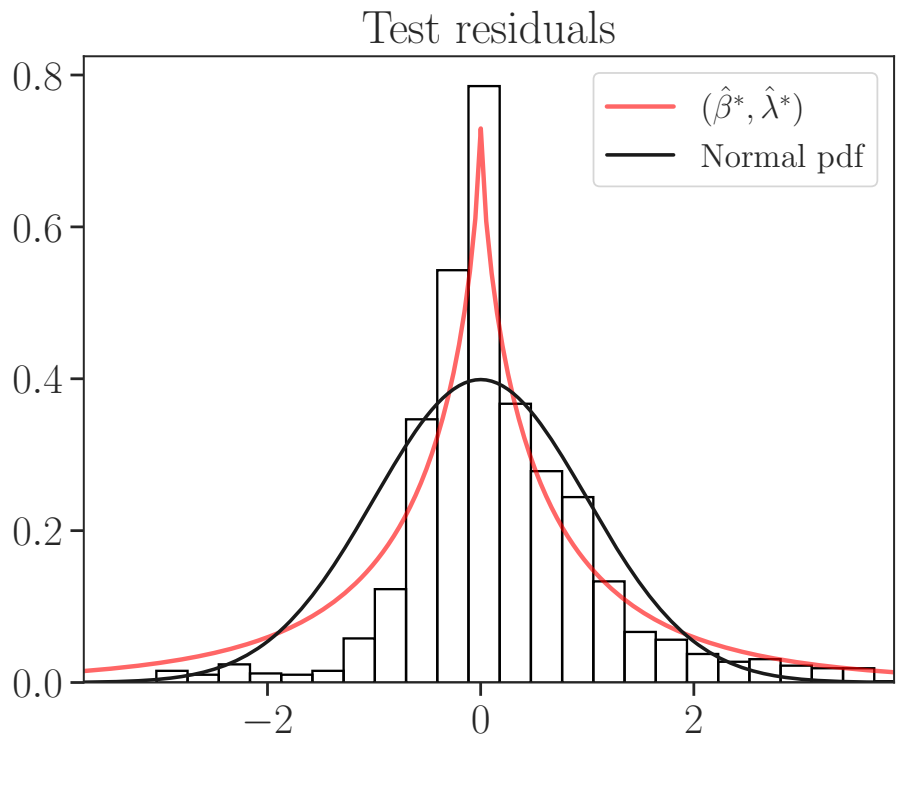
➜ This is a **defensive strategy**: it helps avoid under-dispersed predictive distributions

➜ Smaller δ means more conservative, wider predictions

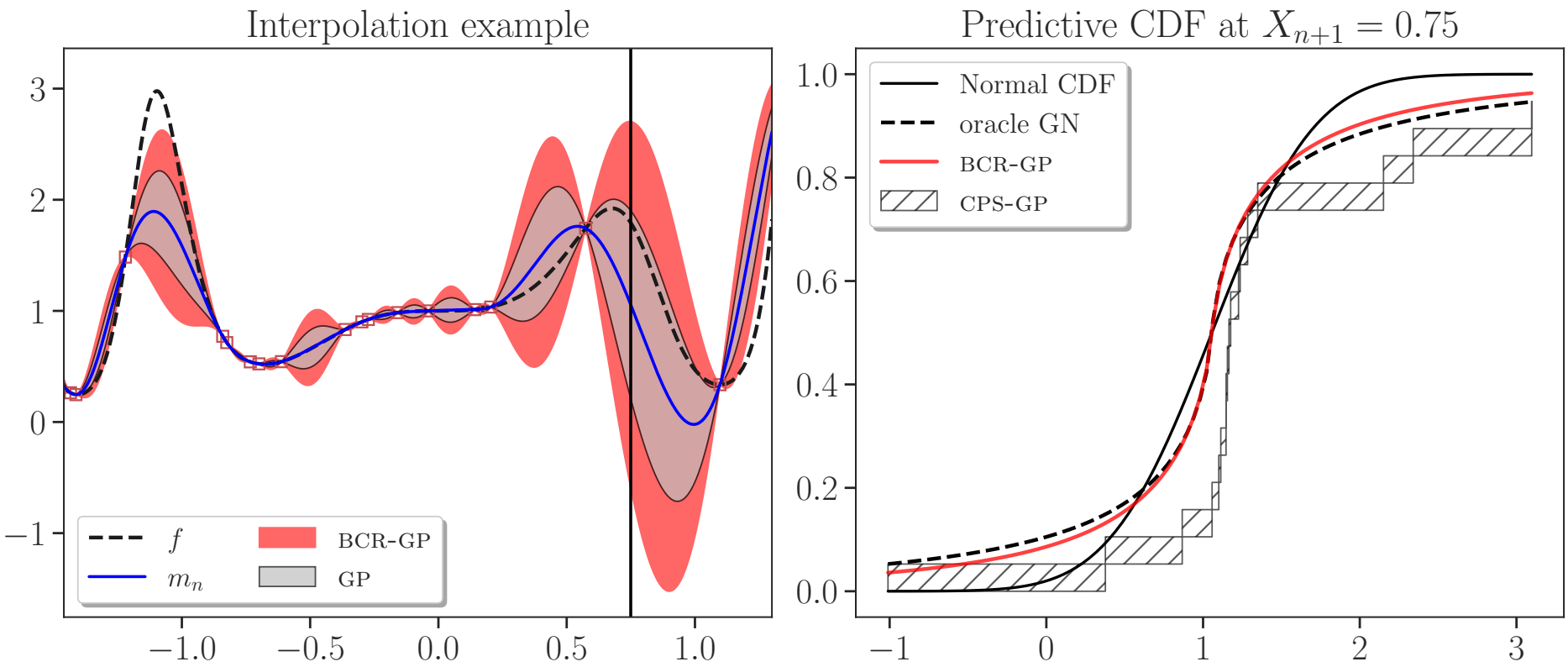
Posterior on calibration parameters

➜ Sample from $p(\beta, \lambda \mid R_{n,-1:n})$

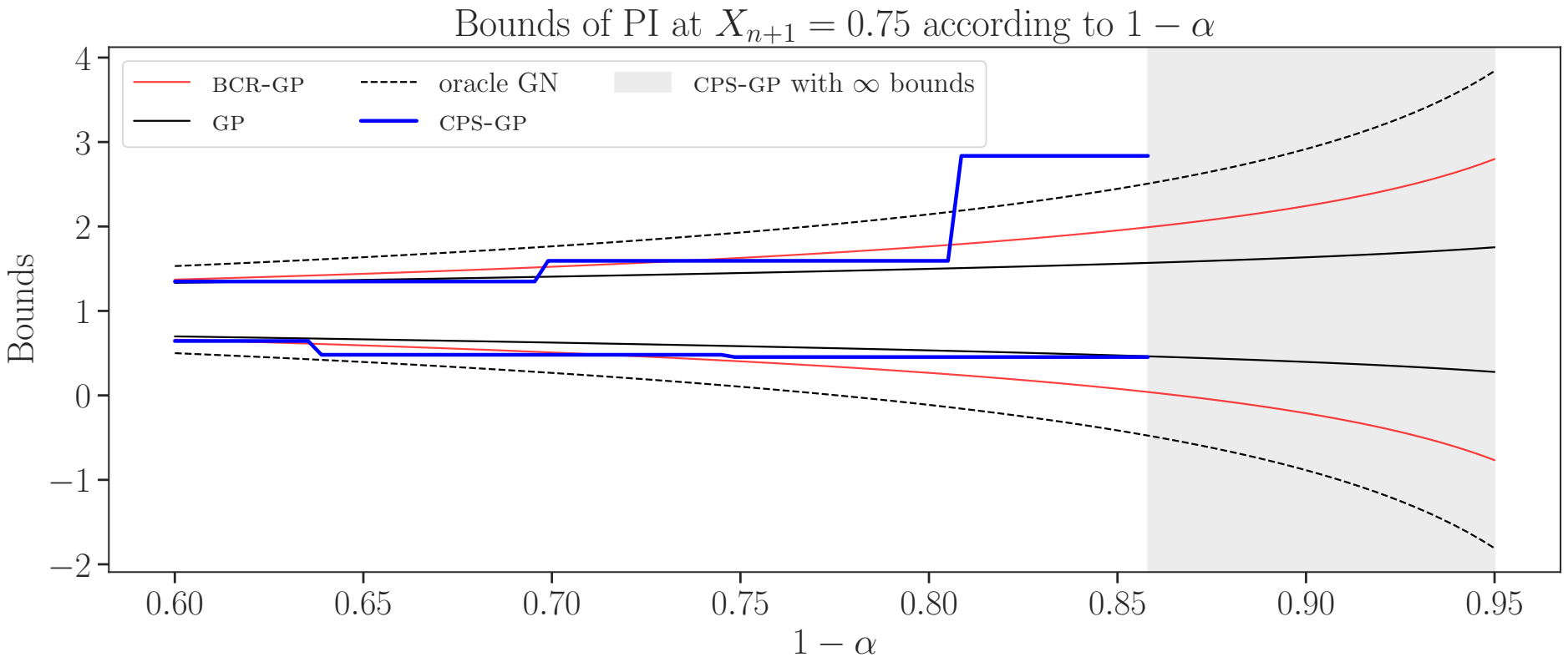




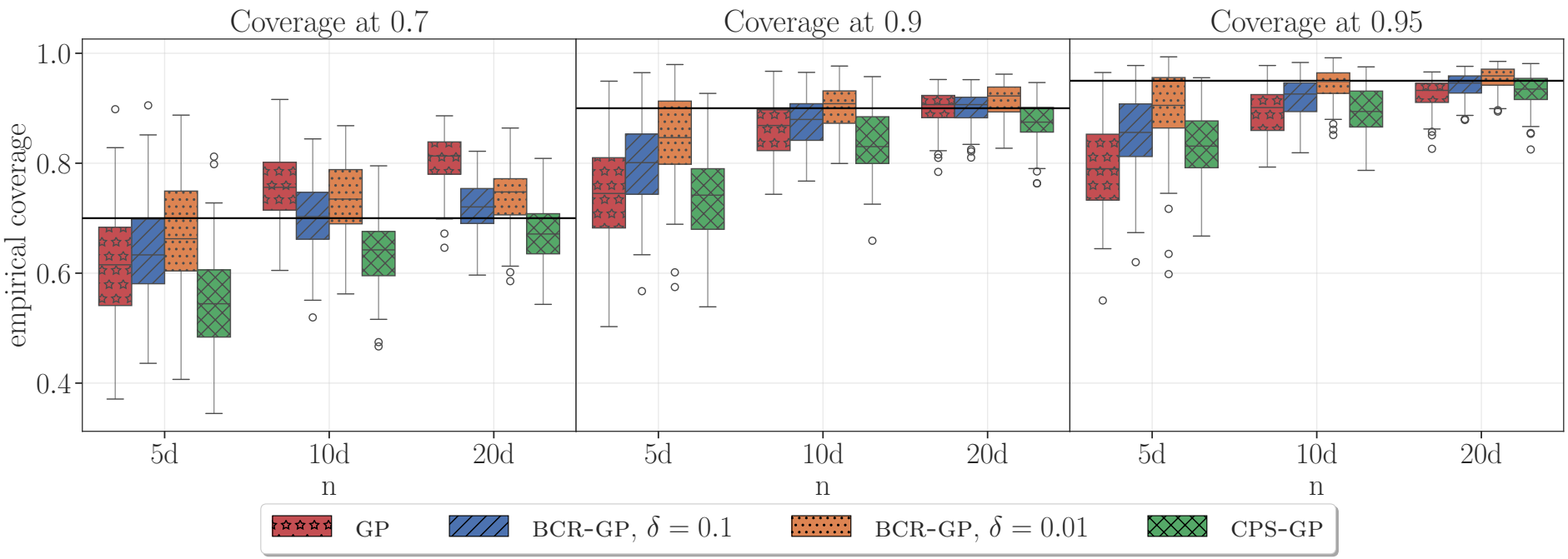
Predictive CDF at $x = 0.75$



Prediction interval at $x = 0.75$



Results


Coverage at 0.7
Coverage at 0.9
Coverage at 0.95
empirical coverage
1.0
0.8
0.6
0.4
5d
10d
20d
n
GP
BCR-GP, $\delta = 0.1$
BCR-GP, $\delta = 0.01$
CPS-GP

Conclusions

➜ For deterministic computer experiments, calibration should be assessed through μ-**calibration**, that is, in a **design-marginal** sense

➜ We propose **two post-hoc calibration methods** for GP predictive CDFs, both based on LOO quantities

➜ CPS-GP provides a conformal route, while BCR-GP yields smooth predictive CDFs and better control of tail behavior

➜ **Calibration notions** also extend to **sequential strategies**, such as Bayesian Optimization, through the notion of **goal-oriented calibration**

Thanks for your attention !

Pre-Print available online: *Design-marginal calibration of Gaussian process predictive distributions: Bayesian and conformal approaches*.

References I


Bolin, D. and Wallin, J. (2023).
Local scale invariance and robustness of proper scoring rules.
Statistical Science, 38(1):140–159.


Da Veiga, S. (2024).
Tutorial on conformal prediction and related methods - ETICS 2024 Research School.
Lecture.


Dawid, A. P. (1984).
Present position and potential developments: Some personal views: Statistical theory: The prequential approach.
Journal of the Royal Statistical Society. Series A (General), 147(2):278–292.


Diebold, F., Gunther, T., and Tay, A. (1998).
Evaluating density forecasts with applications to financial risk management.
International Economic Review, 39(4):863–883.

References II


Gneiting, T. and Resin, J. (2023).
Regression diagnostics meets forecast evaluation: conditional calibration, reliability diagrams, and coefficient of determination.
Electronic Journal of Statistics, 17(2):3226 – 3286.


Gneiting, T. G. and Raftery, A. E. (2007).
Strictly proper scoring rules, prediction, and estimation.
J. Am. Stat. Assoc., 102:359–378.


Marrel, A. and Iooss, B. (2024).
Probabilistic surrogate modeling by gaussian process: A review on recent insights in estimation and validation.
Reliab. Eng. Syst. Saf., page 110094.


Petit, S. J., Bect, J., Feliot, P., and Vazquez, E. (2023).
Parameter selection in gaussian process interpolation: An empirical study of selection criteria.
SIAM/ASA J. Uncertain. Quantif., 11(4):1308–1328.

References III


Vovk, V., Gammerman, A., and Shafer, G. (2005).
Algorithmic Learning in a Random World.
Springer.

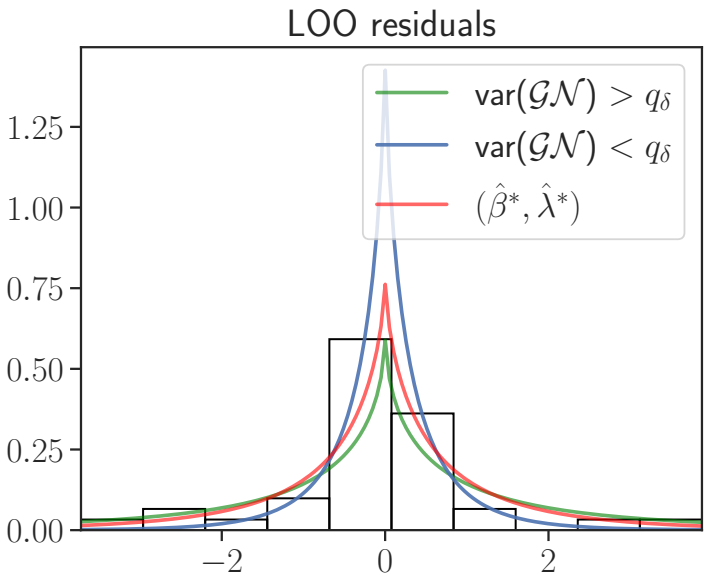

Vovk, V., Shen, J., Manokhin, V., and Xie, M. (2019).
Nonparametric predictive distributions based on conformal prediction.
Machine Learning, 108(3):445–474.

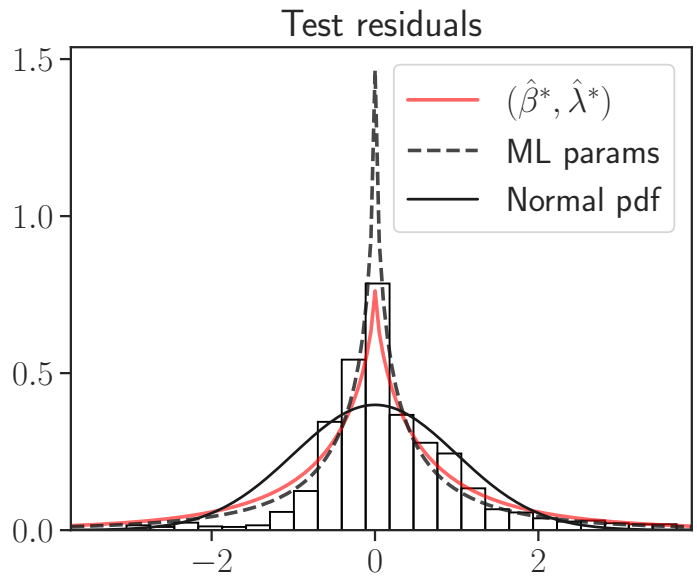
Generalized normal distribution

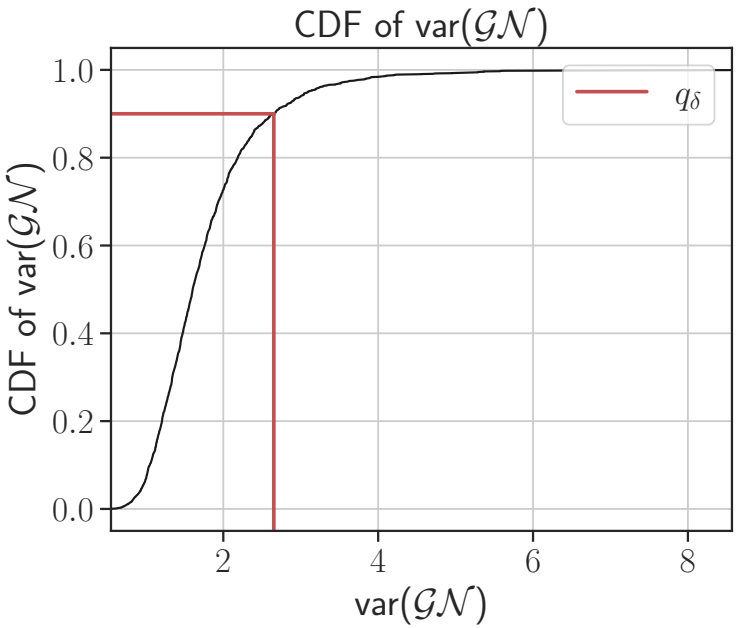
$$f(z) = \frac{\beta}{2\Gamma(1/\beta)\lambda} \exp\left[-\left(\frac{|z-\mu|}{\lambda}\right)^{\beta}\right].$$

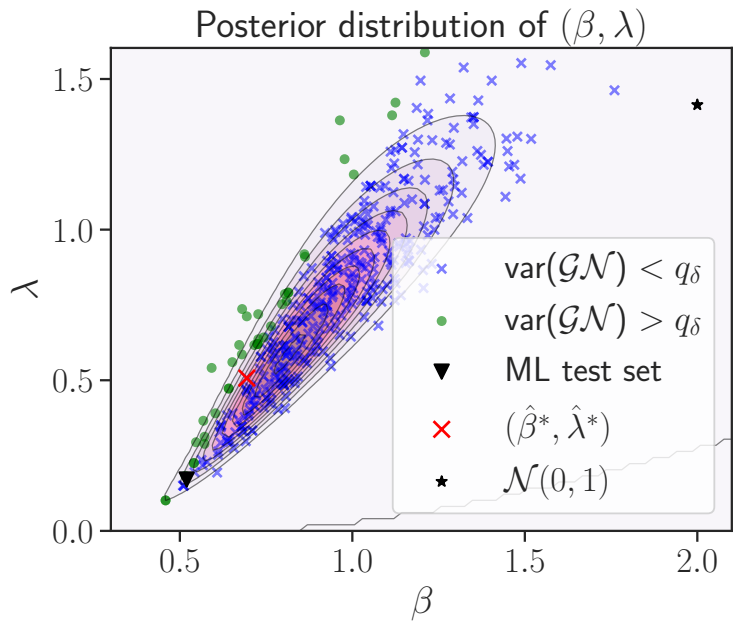
If $\beta < 2$, the distribution has heavier tail and if $\beta > 2$, the distribution has lighter tails.

Posterior distribution



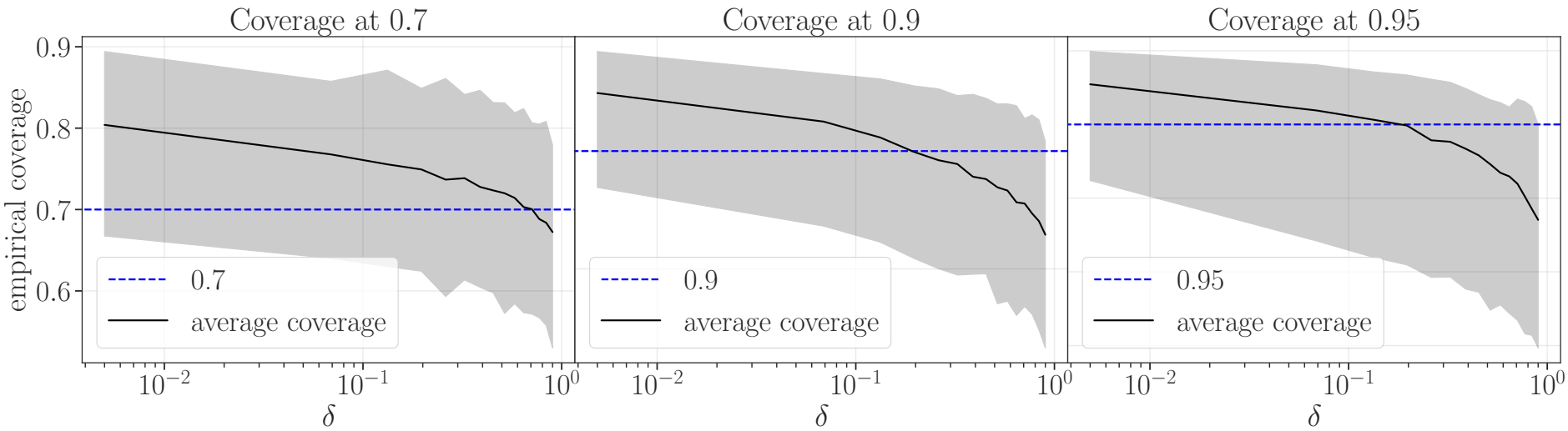






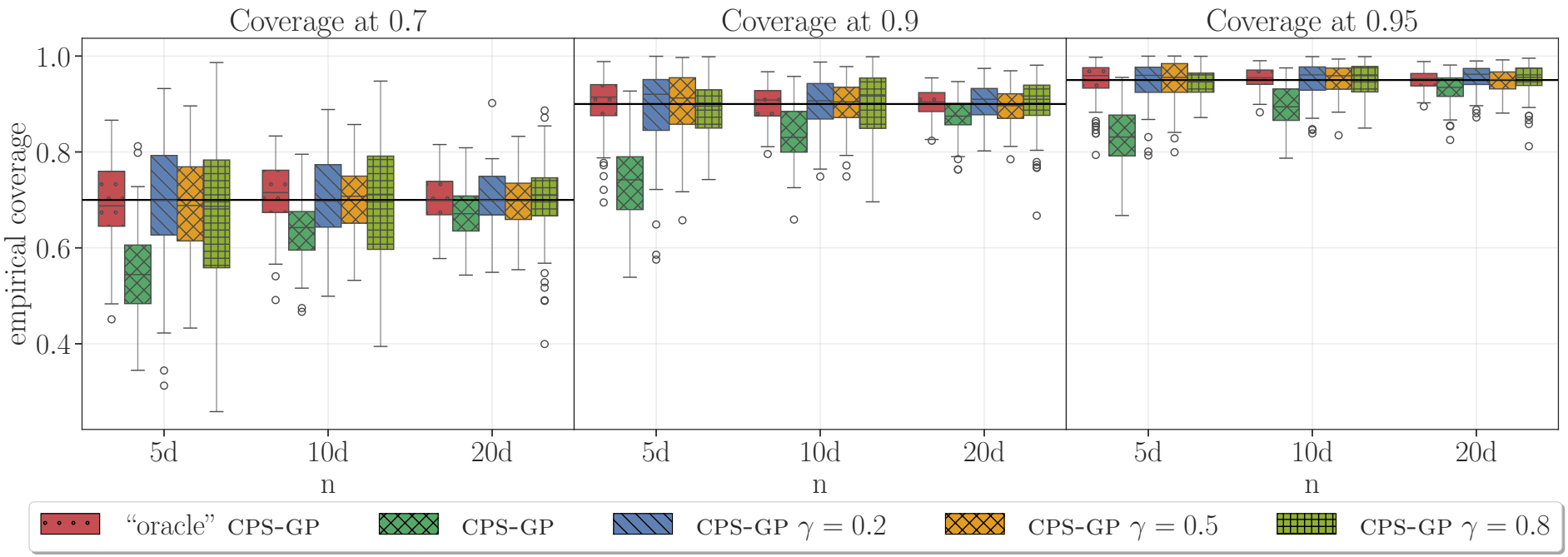
Results

Tolerance level



Results

CPS with different dataset



Some limitations of CPS-GP

➜ The resulting predictive CDF is typically **stepwise**

➜ As a consequence, it does not extrapolate smoothly outside the observed score range

➜ This is inconvenient for **tail probabilities** and related quantities
➜ If GP hyperparameters are selected from the same training set, exact conformal guarantees no longer hold exactly

Results

Comparison of coverage prediction intervals

➜ Coverage comparison