Robust Prediction Interval Estimation for Gaussian Processes by Cross-Validation Method

MASCOT-NUM annual meeting, Clermont-Ferrand, France.

Naoufal Acharki [1,2], Josselin Garnier[1], Antoine Bertoncello[2]

June 6, 2022

[1] Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique
[2] TotalEnergies OneTech.

Context

Uncertainty Quantification plays an essential role in risk assessment and decision-making.

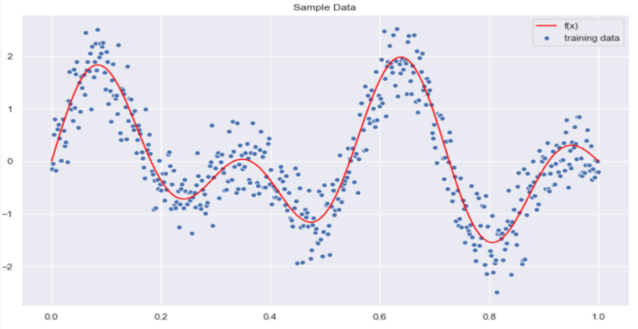
Example: While estimating natural gas reserves, companies should comply with Securities and Exchange Commission rules.

1P	**2P**	**3P**
90% wells produce more than 1P predictions (**proven**)	50% wells produce more than 2P predictions (**probable**)	10% wells produce more more than 3P predictions (**possible**)

However, many approaches and ML models do not fit or may require huge amount of data to predict uncertainty (e.g. jackknife, bootstrap).

Introduction

Consider a standard problem of Kriging with Gaussian Processes [Rasmussen and Williams, 2005]: d-dimensional input dataset $\mathbf{X} = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}\right)$ containing n observations with an output vector $\boldsymbol{y} = \left(y^{(1)}, \dots, y^{(n)}\right)$.



Modelling with Gaussian Processes

Ingredients: A training set $(\mathbf{X}, \boldsymbol{y}) = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$, a family of covariance function $\{\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}\}$ with $\dim(\boldsymbol{\theta}) = d$ and a new point to predict $\boldsymbol{x}_{\text{new}}$.

Modelling with Gaussian Processes

Ingredients: A training set $(\mathbf{X}, \boldsymbol{y}) = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$, a family of covariance function $\{\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}\}$ with $\dim(\boldsymbol{\theta}) = d$ and a new point to predict $\boldsymbol{x}_{\mathrm{new}}$.

Assumption: Assumption of the Gaussian Processes prior.

$$\left(Y(\boldsymbol{x}^{(i)})\right)_{i=1}^{n} | \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}(\mathbf{F}\beta, \mathbf{K}),$$

Modelling with Gaussian Processes

Ingredients: A training set $(\mathbf{X}, \boldsymbol{y}) = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$, a family of covariance function $\{\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}\}$ with $\dim(\boldsymbol{\theta}) = d$ and a new point to predict $\boldsymbol{x}_{\mathrm{new}}$.

Assumption: Assumption of the Gaussian Processes prior.

$$\left(Y(\boldsymbol{x}^{(i)})\right)_{i=1}^{n} | \beta, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}(\mathbf{F}\beta, \mathbf{K}),$$

where $\boldsymbol{m} = \mathbf{F}\beta$ is the trend and $\mathbf{K} = \left(\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)\right)_{i,j=1}^{n} + \sigma_\epsilon^2 \mathbf{I}_n$ is the covariance matrix.

Modelling with Gaussian Processes

Ingredients: A training set $(\mathbf{X}, \boldsymbol{y}) = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$, a family of covariance function $\{\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}\}$ with $\dim(\boldsymbol{\theta}) = d$ and a new point to predict $\boldsymbol{x}_{\mathrm{new}}$.

Assumption: Assumption of the Gaussian Processes prior.

$$\left(Y(\boldsymbol{x}^{(i)})\right)_{i=1}^{n} |\beta, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}(\mathbf{F}\beta, \mathbf{K}),$$

where $\boldsymbol{m} = \mathbf{F}\beta$ is the trend and $\mathbf{K} = \left(\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)\right)_{i,j=1}^{n} + \sigma_\epsilon^2 \mathbf{I}_n$ is the covariance matrix.

Theorem: The *posterior* predictive distribution is Gaussian

$$Y(\boldsymbol{x}_{\mathrm{new}})|\mathbf{X}, \boldsymbol{y}, \beta, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}\left(\tilde{y}(\boldsymbol{x}_{\mathrm{new}}), \tilde{\sigma}^2(\boldsymbol{x}_{\mathrm{new}})\right)$$

Modelling with Gaussian Processes

Ingredients: A training set $(\mathbf{X}, \boldsymbol{y}) = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$, a family of covariance function $\{\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}\}$ with $\dim(\boldsymbol{\theta}) = d$ and a new point to predict $\boldsymbol{x}_{\mathrm{new}}$.

Assumption: Assumption of the Gaussian Processes prior.

$$\left(Y(\boldsymbol{x}^{(i)})\right)_{i=1}^{n} |\beta, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}(\mathbf{F}\beta, \mathbf{K}),$$

where $\boldsymbol{m} = \mathbf{F}\beta$ is the trend and $\mathbf{K} = \left(\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)\right)_{i,j=1}^{n} + \sigma_\epsilon^2 \mathbf{I}_n$ is the covariance matrix.

Theorem: The *posterior* predictive distribution is Gaussian

$$Y(\boldsymbol{x}_{\mathrm{new}})|\mathbf{X}, \boldsymbol{y}, \beta, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}\left(\tilde{y}(\boldsymbol{x}_{\mathrm{new}}), \tilde{\sigma}^2(\boldsymbol{x}_{\mathrm{new}})\right)$$

Result: Uncertainties are fully characterized

$$\mathcal{PI}_{1-\alpha}(\boldsymbol{x}_{\mathrm{new}}) = \left[\tilde{y}(\boldsymbol{x}_{\mathrm{new}}) + q_{\alpha/2}\, \tilde{\sigma}(\boldsymbol{x}_{\mathrm{new}});\ \tilde{y}(\boldsymbol{x}_{\mathrm{new}}) + q_{1-\alpha/2}\, \tilde{\sigma}(\boldsymbol{x}_{\mathrm{new}})\right]$$

Learning Gaussian Processes model i

The nugget effect $\widehat{\sigma}_{\epsilon}^{2}$ can be estimated by a sequential approach [Iooss and Marrel, 2019].

The hyperparameters $(\sigma^2, \boldsymbol{\theta})$ can be estimated either by:

Learning Gaussian Processes model i

The nugget effect $\widehat{\sigma}_{\epsilon}^{2}$ can be estimated by a sequential approach [Iooss and Marrel, 2019].

The hyperparameters $(\sigma^2, \boldsymbol{\theta})$ can be estimated either by:

- The **Maximum Likelihood (ML)** that maximizes the likelihood so that the optimized model produces observed data with the highest probability.

$$(\widehat{\sigma}_{ML}^{2}, \widehat{\boldsymbol{\theta}}_{ML}) \in \operatorname{argmin}_{\sigma^2, \boldsymbol{\theta}} \ell(\sigma^2, \boldsymbol{\theta} \mid \boldsymbol{y}) = \boldsymbol{y}^{\top} \overline{\mathbf{K}} \boldsymbol{y} + \log\left(\det \mathbf{K}\right).$$

where $\overline{\mathbf{K}} = \mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{F}\left(\mathbf{F}^{\top}\mathbf{K}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^{\top}\mathbf{K}^{-1}$.

Learning Gaussian Processes model ii

- **The Mean Squared Error Cross-Validation (MSE-CV)** [Bachoc, 2013] that minimizes the MSE when predicting $y^{(i)}$ using all other points $(\mathbf{X}_{-i}, \boldsymbol{y}_{-i})$ (Leave-One-Out)

$$(\hat{\sigma}^2_{MSE}, \hat{\boldsymbol{\theta}}_{MSE}) \in \operatorname{argmin}_{\sigma^2, \boldsymbol{\theta}} \; \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - \tilde{y}_i \right)^2 = \boldsymbol{y}^\top \overline{\mathbf{K}} \, \mathrm{Diag}\left(\overline{\mathbf{K}}\right)^{-2} \overline{\mathbf{K}} \; \boldsymbol{y}.$$

Learning Gaussian Processes model ii

- **The Mean Squared Error Cross-Validation (MSE-CV)** [Bachoc, 2013] that minimizes the MSE when predicting $y^{(i)}$ using all other points $(\mathbf{X}_{-i}, \boldsymbol{y}_{-i})$ (Leave-One-Out)

$$(\hat{\sigma}^2_{MSE}, \hat{\boldsymbol{\theta}}_{MSE}) \in \text{argmin}_{\sigma^2,\boldsymbol{\theta}} \ \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \tilde{y}_i\right)^2 = \boldsymbol{y}^\top \overline{\mathbf{K}}\,\text{Diag}\left(\overline{\mathbf{K}}\right)^{-2}\overline{\mathbf{K}}\ \boldsymbol{y}.$$

The regression coefficients $\widehat{\beta}$ are estimated by **Generalized Least Squares** method

$$\widehat{\beta} = \left(\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{K}^{-1}\boldsymbol{y}$$

once the hyperparameters $(\widehat{\sigma}^2_{ML}, \widehat{\boldsymbol{\theta}}_{ML})$ or $(\widehat{\sigma}^2_{MSE}, \widehat{\boldsymbol{\theta}}_{MSE})$ are obtained.

Well-specified vs misspecified models

Formal definition of a well-specified model: The model is said to be *well-specified* if there exists a couple of hyperparameters $(\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0)$ such that $\boldsymbol{y}$ is considered as a realization of a GP model with covariance function $\boldsymbol{k}_{\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0}$.

Well-specified vs misspecified models

Formal definition of a well-specified model: The model is said to be *well-specified* if there exists a couple of hyperparameters $(\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0)$ such that $\boldsymbol{y}$ is considered as a realization of a GP model with covariance function $\boldsymbol{k}_{\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0}$.

Informal definition of a well-specified model: The model is said to be *well-specified* if $\boldsymbol{y}$ satisfies the normality given the obtained hyperparameters by MLE method $(\widehat{\sigma}_{ML}^2, \widehat{\boldsymbol{\theta}}_{ML})$.
This result can be verified only empirically (e.g. graphically or using Shapiro test on the predictive distribution).

Well-specified vs misspecified models

Formal definition of a well-specified model: The model is said to be *well-specified* if there exists a couple of hyperparameters $(\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0)$ such that $\boldsymbol{y}$ is considered as a realization of a GP model with covariance function $\boldsymbol{k}_{\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0}$.

Informal definition of a well-specified model: The model is said to be *well-specified* if $\boldsymbol{y}$ satisfies the normality given the obtained hyperparameters by MLE method $(\widehat{\sigma}_{ML}^2, \widehat{\boldsymbol{\theta}}_{ML})$.
This result can be verified only empirically (e.g. graphically or using Shapiro test on the predictive distribution).

Misspecified model: when the model is not well-specified.

Challenges

In some cases, the model or the data may not honour all assumptions, hence no guarantees that Prediction intervals are well estimated by the **Maximum Likelihood** method [Bachoc, 2013].

Challenges

In some cases, the model or the data may not honour all assumptions, hence no guarantees that Prediction intervals are well estimated by the **Maximum Likelihood** method [Bachoc, 2013].

The **MSE-CV** method is more efficient when the model is misspecified and adapted for point-wise prediction [Bachoc, 2013] but the predictive variance may not be well estimated.

The main ideas of the paper

- Consider the estimation of covariance hyperparameters in a misspecified model setting.
- Propose a CV-based approach for a robust estimation of the model hyperparameters.
- Improve the quality of the estimated Predictive Intervals to achieve a nominal confidence level.

Estimating Prediction Intervals bounds by Cross-Validation i

Quick Reminder of the Leave-One-Out method:

Estimating Prediction Intervals bounds by Cross-Validation i

Quick Reminder of the Leave-One-Out method:

Assume $\widehat{\sigma}_\epsilon^2$ is known or has been estimated.

Estimating Prediction Intervals bounds by Cross-Validation i

Quick Reminder of the Leave-One-Out method:

Assume $\widehat{\sigma}_\epsilon^2$ is known or has been estimated.

For given hyperparameters $(\sigma^2, \boldsymbol{\theta})$, "build" n GP models where each model has been trained on $(\mathbf{X}_{-i}, \boldsymbol{y}_{-i})$ to predict, at each point $\boldsymbol{x}^{(i)}$,

$$\tilde{y}_i = \tilde{y}(\boldsymbol{x}^{(i)}) \text{ the predictive mean,}$$
$$\tilde{\sigma}_i^2 = \tilde{\sigma}^2(\boldsymbol{x}^{(i)}) \text{ the predictive variance,}$$
$$\widetilde{y}_i^a = \tilde{y}_i + q_a \times \tilde{\sigma}_i \text{ the PI bound for a given rate } a.$$

Estimating Prediction Intervals bounds by Cross-Validation i

Quick Reminder of the Leave-One-Out method:

Assume $\widehat{\sigma}_{\epsilon}^2$ is known or has been estimated.

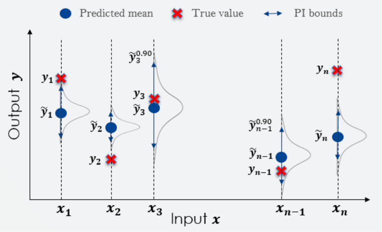
For given hyperparameters $(\sigma^2, \boldsymbol{\theta})$, "build" n GP models where each model has been trained on $(\mathbf{X}_{-i}, \boldsymbol{y}_{-i})$ to predict, at each point $\boldsymbol{x}^{(i)}$,

$$\tilde{y}_i = \tilde{y}(\boldsymbol{x}^{(i)}) \text{ the predictive mean,}$$
$$\tilde{\sigma}_i^2 = \tilde{\sigma}^2(\boldsymbol{x}^{(i)}) \text{ the predictive variance,}$$
$$\widetilde{y}_i^a = \tilde{y}_i + q_a \times \tilde{\sigma}_i \text{ the PI bound for a given rate } a.$$

In reality, we do not "build" n models, we have direct formulas [Dubrule, 1983] to estimate $\tilde{y}_i$ and $\tilde{\sigma}_i^2$.

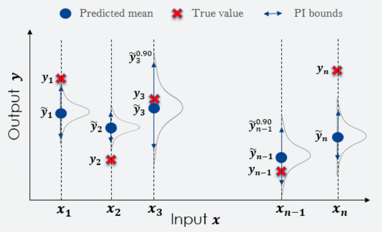
Estimating Prediction Intervals bounds by Cross-Validation ii

How to insure that the PI bound $\widetilde{\boldsymbol{y}}^{a} = (\widetilde{y}_i^{a})_{i=1}^{n}$ covers exactly $a \times 100\%$ (e.g. 95%) of true values ?



Estimating Prediction Intervals bounds by Cross-Validation ii

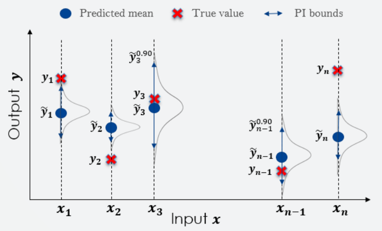
How to insure that the PI bound $\widetilde{\boldsymbol{y}}^a = (\widetilde{y}_i^a)_{i=1}^n$ covers exactly $a \times 100\%$ (e.g. 95%) of true values ?



By modifying the predictive distribution at each point.

Estimating Prediction Intervals bounds by Cross-Validation ii

How to insure that the PI bound $\widetilde{\boldsymbol{y}}^a = (\widetilde{y}_i^a)_{i=1}^n$ covers exactly $a \times 100\%$ (e.g. 95%) of true values ?

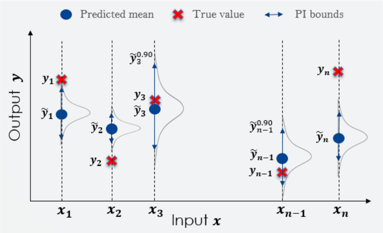


By modifying the predictive distribution at each point.

i.e. By calibrating the hyperparameters of the model.

Estimating Prediction Intervals bounds by Cross-Validation ii

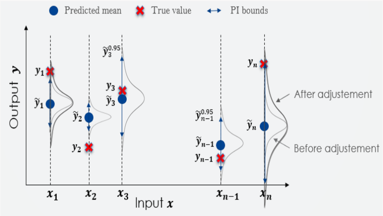
How to insure that the PI bound $\widetilde{\boldsymbol{y}}^a = (\widetilde{y}_i^a)_{i=1}^n$ covers exactly $a \times 100\%$ (e.g. 95%) of true values ?



By modifying the predictive distribution at each point.

i.e. By calibrating the hyperparameters of the model.

i.e. By optimizing model's hyperparameters with respect to a special metric ψ_a.



Robust Prediction Intervals Estimation i

Consider

$$\psi_a\left(\sigma^2, \boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\frac{y^{(i)} - \tilde{y}_i}{\tilde{\sigma}_i} \leq q_a\right\}$$ (i.e. the number of predictions falling below q_a).

Robust Prediction Intervals Estimation i

Consider

$$\psi_a\left(\sigma^2, \boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\frac{y^{(i)} - \tilde{y}_i}{\tilde{\sigma}_i} \leq q_a\right\} \quad \text{(i.e. the number of predictions falling below } q_a\text{).}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\frac{\left(\overline{\mathbf{K}}\boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{K}}\right)_{i,i}}} \leq q_a\right\} \quad \text{(Virtual Cross-Validation formulas of Dubrule [1983])}$$

Robust Prediction Intervals Estimation i

Consider

$$\begin{aligned}\psi_a\left(\sigma^2, \boldsymbol{\theta}\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\frac{y^{(i)} - \tilde{y}_i}{\tilde{\sigma}_i} \leq q_a\right\} \quad \text{(i.e. the number of predictions falling below } q_a\text{).}\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\frac{\left(\overline{\mathbf{K}}\boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{K}}\right)_{i,i}}} \leq q_a\right\} \quad \text{(Virtual Cross-Validation formulas of Dubrule [1983])}\\ &\simeq \psi_a^{(\delta)}\left(\sigma^2, \boldsymbol{\theta}\right) \quad \text{(for a continuous function } \psi_a^{(\delta)} \text{ converging point-wise to } \psi_a\text{).}\end{aligned}$$

Robust Prediction Intervals Estimation i

Consider

$$\begin{aligned}\psi_a\left(\sigma^2,\boldsymbol{\theta}\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\frac{y^{(i)}-\tilde{y}_i}{\tilde{\sigma}_i}\leq q_a\right\} \quad \text{(i.e. the number of predictions falling below } q_a\text{).}\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\frac{\left(\overline{\mathbf{K}}\boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{K}}\right)_{i,i}}}\leq q_a\right\} \quad \text{(Virtual Cross-Validation formulas of Dubrule [1983])}\\ &\simeq \psi_a^{(\delta)}\left(\sigma^2,\boldsymbol{\theta}\right) \quad \text{(for a continuous function } \psi_a^{(\delta)} \text{ converging point-wise to } \psi_a\text{).}\end{aligned}$$

If $\psi_a\left(\sigma^2,\boldsymbol{\theta}\right)=a$ for some hyperparameters $\left(\sigma^2,\boldsymbol{\theta}\right)$ **then** your model has learned to estimate the PI bound $\widetilde{\boldsymbol{y}}^a$ such that $a\times 100\%$ of true values $\boldsymbol{y}$ are below $\widetilde{\boldsymbol{y}}^a$.

Robust Prediction Intervals Estimation ii

Challenge 1: $\psi_a^{(\delta)}$ is piece-wise constant, there would be an infinite number of solutions in the search space $\mathcal{H}$.

Robust Prediction Intervals Estimation ii

Challenge 1: $\psi_a^{(\delta)}$ is piece-wise constant, there would be an infinite number of solutions in the search space $\mathcal{H}$.

Solution: Reformulate the problem i.e. choose the closest solution $(\sigma^2, \boldsymbol{\theta})$ to $(\sigma_0^2, \boldsymbol{\theta}_0)$ (ML or MSE-CV solution) using a similarity measure d (Wasserstein distance [Masarotto et al., 2019])

$$\underset{(\sigma^2, \boldsymbol{\theta})}{\arg\min}\ d^2\left((\sigma^2, \boldsymbol{\theta}), (\sigma_0^2, \boldsymbol{\theta}_0)\right).$$

Robust Prediction Intervals Estimation ii

Challenge 1: $\psi_a^{(\delta)}$ is piece-wise constant, there would be an infinite number of solutions in the search space $\mathcal{H}$.

Solution: Reformulate the problem i.e. choose the closest solution $(\sigma^2, \boldsymbol{\theta})$ to $(\sigma_0^2, \boldsymbol{\theta}_0)$ (ML or MSE-CV solution) using a similarity measure d (Wasserstein distance [Masarotto et al., 2019])

$$\underset{(\sigma^2, \boldsymbol{\theta})}{\arg\min} \; d^2\left((\sigma^2, \boldsymbol{\theta}), (\sigma_0^2, \boldsymbol{\theta}_0)\right).$$

Challenge 2: The resolution of this problem may be too costly when the dimension d is high.

Robust Prediction Intervals Estimation ii

Challenge 1: $\psi_a^{(\delta)}$ is piece-wise constant, there would be an infinite number of solutions in the search space $\mathcal{H}$.

Solution: Reformulate the problem i.e. choose the closest solution $(\sigma^2, \boldsymbol{\theta})$ to $(\sigma_0^2, \boldsymbol{\theta}_0)$ (ML or MSE-CV solution) using a similarity measure d (Wasserstein distance [Masarotto et al., 2019])

$$\underset{(\sigma^2, \boldsymbol{\theta})}{\arg\min} \; d^2\left((\sigma^2, \boldsymbol{\theta}), (\sigma_0^2, \boldsymbol{\theta}_0)\right).$$

Challenge 2: The resolution of this problem may be too costly when the dimension d is high.

Solution: Reduce the dimension of the search space $\mathcal{H}$ by applying *the relaxation* method i.e. fix $\boldsymbol{\theta}_0 \in (\widehat{\boldsymbol{\theta}}_{ML}, \widehat{\boldsymbol{\theta}}_{MSE})$, $\lambda \in (0, +\infty)$ and solve for σ^2

$$\psi_a^{(\delta)}(\sigma^2, \lambda\boldsymbol{\theta}_0) = a$$

Robust Prediction Intervals Estimation iii

- Take the smallest variance $\sigma^2_{\mathrm{opt}}(\lambda)$ that satisfies $\psi_a^{(\delta)}(\sigma^2, \lambda\boldsymbol{\theta}_0) = a$.
  (In the kriging framework, σ^2 should be as small as possible).

Robust Prediction Intervals Estimation iii

- Take the smallest variance $\sigma^2_{\mathrm{opt}}(\lambda)$ that satisfies $\psi_a^{(\delta)}(\sigma^2, \lambda\boldsymbol{\theta}_0) = a$.
  (In the kriging framework, σ^2 should be as small as possible).

- Plug this solution in the following minimization problem

$$\operatorname*{arg\,min}_{\lambda\in(0,+\infty)} d^2\left((\sigma^2_{\mathrm{opt}}(\lambda), \lambda\boldsymbol{\theta}_0), (\sigma^2_0, \boldsymbol{\theta}_0)\right)$$

Robust Prediction Intervals Estimation iii

- Take the smallest variance $\sigma^2_{\mathrm{opt}}(\lambda)$ that satisfies $\psi_a^{(\delta)}(\sigma^2, \lambda\boldsymbol{\theta}_0) = a$.
  (In the kriging framework, σ^2 should be as small as possible).

- Plug this solution in the following minimization problem

$$\underset{\lambda\in(0,+\infty)}{\arg\min}\ d^2\big((\sigma^2_{\mathrm{opt}}(\lambda), \lambda\boldsymbol{\theta}_0), (\sigma_0^2, \boldsymbol{\theta}_0)\big)$$

Proposition: Under appropriate hypotheses, this problem admits at least a solution λ^*.

Robust Prediction Intervals Estimation iii

- Take the smallest variance $\sigma^2_{\mathrm{opt}}(\lambda)$ that satisfies $\psi_a^{(\delta)}(\sigma^2, \lambda\boldsymbol{\theta}_0) = a$.
  (In the kriging framework, σ^2 should be as small as possible).

- Plug this solution in the following minimization problem

$$\underset{\lambda\in(0,+\infty)}{\arg\min}\ d^2\left((\sigma^2_{\mathrm{opt}}(\lambda), \lambda\boldsymbol{\theta}_0), (\sigma_0^2, \boldsymbol{\theta}_0)\right)$$

Proposition: Under appropriate hypotheses, this problem admits at least a solution λ^*.

- Update the model by considering the hyperparameters $(\sigma^2_{\mathrm{opt}}(\lambda), \lambda^*\boldsymbol{\theta}_0)$ and $\widehat{\boldsymbol{\beta}}^*_{\mathrm{opt}}$ (with GLS formulas).

General comments on the method

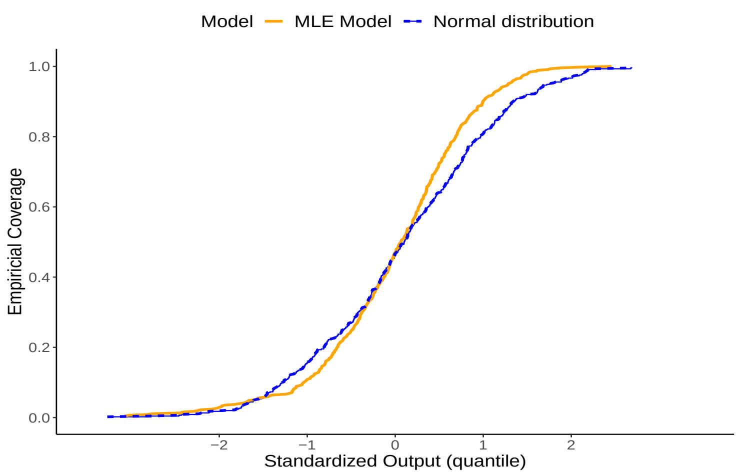
Comment 1: When there is no nugget effect, the proposition does not hold for Matérn kernels with $\nu > 2$.
We can justify it by the fact that Matérn kernels with $\nu > 2$ are less robust for uncertainty quantification.

Comment 2: The function $\mathcal{L}(\lambda) = d^2\big((\sigma^2_{\mathrm{opt}}(\lambda), \lambda\boldsymbol{\theta}_0), (\sigma^2_0, \boldsymbol{\theta}_0)\big)$ is continuous and coercive, thus, a minimize λ^* exists.

Comment 3: The problem can be solved numerically using the golden-section search method.

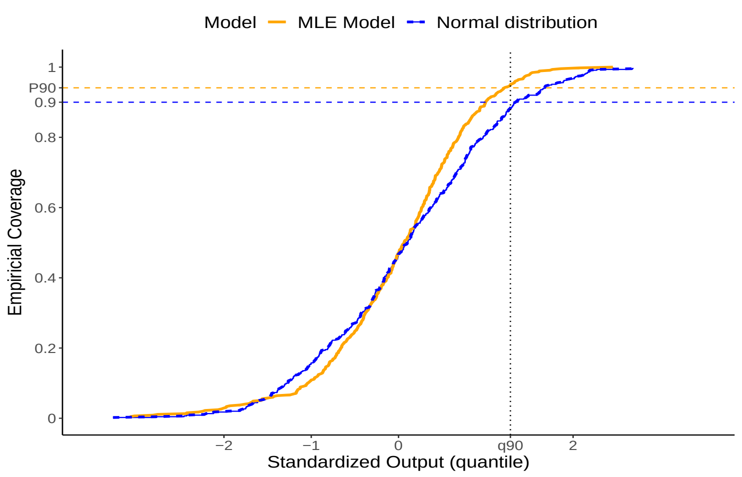
What happens exactly to the predictive distribution? i

Consider the LOO standardized predictive distribution of $\tilde{\boldsymbol{y}} = \left((y^{(i)} - \tilde{y}_i)/\tilde{\sigma}_i\right)_{i=1}^{n}$.



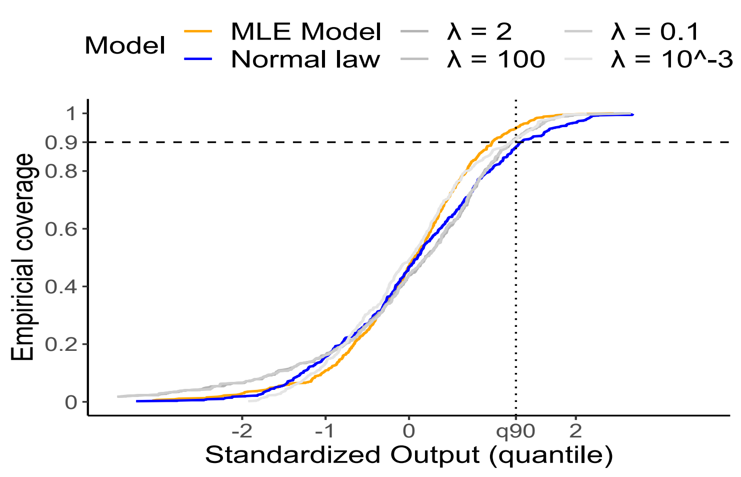
What happens exactly to the predictive distribution? ii

The ML model overestimates the PI bound of level 90% (here the empirical coverage $P90 = 94\%$).



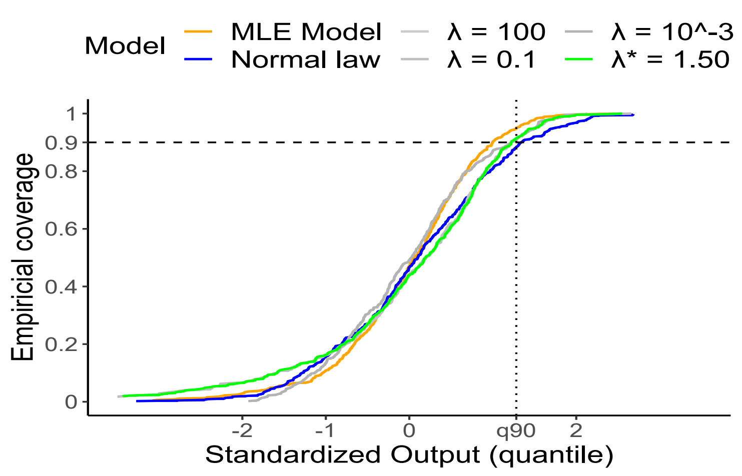
What happens exactly to the predictive distribution? iii

Target the *true* PI bound of level 90%. Infinite distributions that coincide with the standard normal distribution on point $(q_a, a) = (1.28, 0.90)$ are possible.



What happens exactly to the predictive distribution? iv

Pick the optimal distribution (obtained from λ^*) that is close to MLE wrt Wasserstein distance .



Robust Prediction Intervals Estimation iv

The main interest of the **Robust Prediction Interval Estimation (RPIE)** method:

- A GP model $\mathbf{GP}_{\alpha/2}$ able to predict the bound $\tilde{Y}_{\alpha/2}$ such that $\alpha/2 \times 100\%$ of true values are below $\tilde{Y}_{\alpha/2}$.

- A GP model $\mathbf{GP}_{1-\alpha/2}$ able to predict the bound $\tilde{Y}_{1-\alpha/2}$ such that $(1-\alpha/2) \times 100\%$ of true values are below $\tilde{Y}_{1-\alpha/2}$.

Result: Prediction Intervals respecting as best as possible the optimal coverage rate $1-\alpha$.

Benchmark: the Full-Bayesian GP

The Full-Bayesian GP assumes a *prior* on the hyperparameters $(\sigma^2, \boldsymbol{\theta})$ and

Benchmark: the Full-Bayesian GP

The Full-Bayesian GP assumes a *prior* on the hyperparameters $(\sigma^2, \boldsymbol{\theta})$ and

$$p(y_{\mathrm{new}} \mid \boldsymbol{y}) = \iint p(y_{\mathrm{new}} \mid \boldsymbol{y}, \sigma^2, \boldsymbol{\theta}) p(\sigma^2, \boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\sigma^2 \mathrm{d}\boldsymbol{\theta},$$

where $p\big(y_{\mathrm{new}} \mid \sigma^2, \boldsymbol{\theta}\big)$ is given by the posterior predictive distribution of the GP model.

Benchmark: the Full-Bayesian GP

The Full-Bayesian GP assumes a *prior* on the hyperparameters $(\sigma^2, \boldsymbol{\theta})$ and

$$p(y_{\mathrm{new}} \mid \boldsymbol{y}) = \iint p(y_{\mathrm{new}} \mid \boldsymbol{y}, \sigma^2, \boldsymbol{\theta}) p(\sigma^2, \boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\sigma^2 \mathrm{d}\boldsymbol{\theta},$$

where $p\big(y_{\mathrm{new}} \mid \sigma^2, \boldsymbol{\theta}\big)$ is given by the posterior predictive distribution of the GP model.

The predictive distribution is estimated as

$$p(y_{\mathrm{new}} \mid \boldsymbol{y}) \simeq \frac{1}{N} \sum_{i=1}^{N} p(y_{\mathrm{new}} \mid \boldsymbol{y}, \sigma_i^2, \boldsymbol{\theta}_i),$$

where $(\sigma_i^2, \boldsymbol{\theta}_i)$ is the *i*-th sample drawn from the posterior distribution $p(\sigma^2, \boldsymbol{\theta} \mid \boldsymbol{y})$ by MCMC.

$\mathcal{PI}_{1-\alpha}$ are obtained from the empirical $\alpha/2$- and $1 - \alpha/2$-quantiles of the sample $\big(Y_i(\boldsymbol{x}_{\mathrm{new}})\big)_{i=1}^{N}$.

Numerical results: evaluation metrics

The Leave-One-Out Coverage Probability $\tilde{\mathbb{P}}_{1-\alpha}$ on training set and CP on testing set :

$$\tilde{\mathbb{P}}_{1-\alpha} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y^{(i)} \in \mathcal{PI}_{1-\alpha}(\boldsymbol{x}^{(i)})\},$$

$$\mathrm{CP}_{1-\alpha} = \frac{1}{n_{test}}\sum_{i=1}^{n_{test}} \mathbf{1}\{y_{test}^{(i)} \in \mathcal{PI}_{1-\alpha}\left(\boldsymbol{x}_{test}^{(i)}\right)\}$$

The mean (MPIW) and standard-deviation (SdPIW) of Prediction Intervals widths:

$$\mathrm{MPIW}_{1-\alpha} = \frac{1}{n_{test}}\sum_{i=1}^{n_{test}} \left|\mathcal{PI}_{1-\alpha}(\boldsymbol{x}_{test}^{(i)})\right|$$

$$\mathrm{SdPIW}_{1-\alpha} = \left(\frac{1}{n_{test}}\sum_{i=1}^{n_{test}} \left[\left|\mathcal{PI}_{1-\alpha}(\boldsymbol{x}_{test}^{(i)})\right| - \mathrm{MPIW}_{1-\alpha}\right]^2\right)^{1/2}.$$

Toy example: the Morokoff & Caflisch function i

We consider the Morokoff and Caflisch [1995] function defined on $[0,1]^d$ by

$$f(\boldsymbol{x}) = \frac{1}{2}\left(1+\frac{1}{d}\right)^d \prod_{i=1}^{d}(x_i)^{1/d}.$$

$\mathbf{X}$ has $n = 600$ observations and $d = 10$ variables, with a train-test split rate of 75-25%.

$\boldsymbol{y}$ is generated as $y^{(i)} = f(\boldsymbol{x}^{(i)}) + \epsilon^{(i)}$ where $\epsilon^{(i)}$ are sampled i.i.d. over $\mathcal{N}(0, \sigma_\epsilon^2 = 10^{-4})$.

Covariance model family $\boldsymbol{k}$: Matérn 5/2 anisotropic geometric correlation model.

Targeted confidence level: $1 - \alpha = 90\%$.

Toy example: the Morokoff & Caflisch function ii

	Before RPIE		After RPIE		Full-Bayesian
	MLE	MSE-CV	MLE	MSE-CV	-
$\tilde{\mathbb{P}}_{1-\alpha}$	93.6	98.3	90.0	90.0	93.8
CP$_{1-\alpha}$	94.0	98.0	92.6	87.3	93.3
MPIW$_{1-\alpha}$	$1.68\ 10^{-1}$	$1.81\ 10^{-1}$	$5.51\ 10^{-2}$	$5.78\ 10^{-2}$	$1.66\ 10^{-1}$
SdPIW$_{1-\alpha}$	$9.61\ 10^{-3}$	$4.16\ 10^{-2}$	$1.29\ 10^{-2}$	$1.41\ 10^{-2}$	$9.27\ 10^{-3}$
Ct	1min 16s	24min 18s	3min 55s	27min 43s	4h 43min 38s

$\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in % on the training set; CP$_{1-\alpha}$: CP in % on the testing set; MPIW: Mean of Prediction Interval widths; SdPIW: standard deviation of Prediction Interval widths and Ct: computational time.

Industrial case: Predicting uncertainty for gas production i

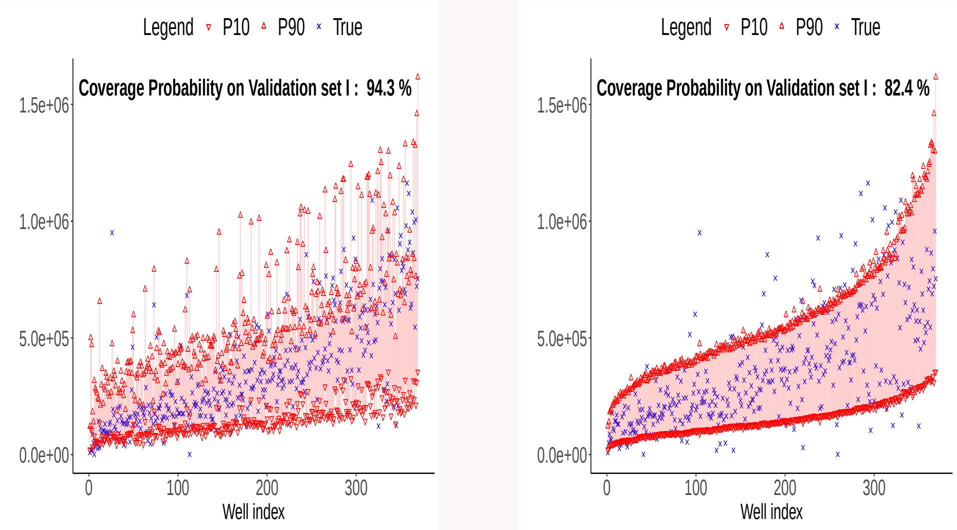
A field dataset containing $n = 1850$ wells with $d = 11$ dimensional inputs $\mathbf{X}$, the output $\boldsymbol{y}$ is the Cumulative Production of natural gas over 12 months.

Targeted confidence level $1 - \alpha = 80\%$

	MLE before RPIE	MLE after RPIE
$\tilde{\mathbb{P}}_{1-\alpha}$	91.1	79.9
$\mathrm{CP}_{1-\alpha}$	94.3	83.2
$\mathrm{MPIW}_{1-\alpha}$	1.53	1.40
$\mathrm{SdPIW}_{1-\alpha}$	$2.20\ 10^{-1}$	$1.40\ 10^{-2}$
Ct	17min 47s	53min 21s

$\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in % on the training set; $\mathrm{CP}_{1-\alpha}$: CP in % on Validation set; $\mathrm{MPIW}_{1-\alpha}$: Mean of Prediction Interval widths; $\mathrm{SdPIW}_{1-\alpha}$: standard deviation of Prediction Interval widths and Ct: computational time.

Industrial case: Predicting uncertainty for gas production ii



(a) Before the RPIE on log output

(b) After the RPIE on log output

Conclusion

- Consider the estimation of covariance hyperparameters in a misspecified model setting to improve Prediction Intervals.

- The RPIE method gives better Prediction Intervals estimation if compared to Maximum Likelihood or Full-Bayesian approaches.

- Categorial inputs should be considered in a future work with group kernels [Roustant et al., 2020]

References

F. Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66: 55–69, 2013.

O. Dubrule. Cross validation of kriging in a unique neighborhood. *Journal of the International Association for Mathematical Geology*, 15(6):687–699, 12 1983.

B. Iooss and A. Marrel. Advanced methodology for uncertainty propagation in computer experiments with large number of inputs. *Nuclear Technology*, 205(12):1588–1606, 2019. doi: 10.1080/00295450.2019.1573617.

V. Masarotto, V. M. Panaretos, and Y. Zemel. Procrustes metrics on covariance operators and optimal transportation of gaussian processes. *Sankhya A*, 81(1):172–213, Feb 2019. ISSN 0976-8378. doi: 10.1007/s13171-018-0130-1.

W. J. Morokoff and R. E. Caflisch. Quasi-monte carlo integration. *Journal of computational physics*, 122:218–230, 1995.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

O. Roustant, E. Padonou, Y. Deville, A. Clément, G. Perrin, J. Giorla, and H. Wynn. Group kernels for Gaussian process metamodels with categorical inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):775–806, 2020. doi: 10.1137/18M1209386.

Securities and Exchange Commission. Modernization of oil and gas reporting, revisions and additions to the definition section in rule 4-10 of regulation s-x, Jan. 2010. `https://www.sec.gov/rules/final/2008/33-8995.pdf`.