Karolinska Institutet

How Reliable is Your Regression Model's Uncertainty Under Real-World Distribution Shifts?

Fredrik K. Gustafsson
Karolinska Institutet
www.fregu856.com

RISE Learning Machines Seminar
March 21, 2024

About Me

Since December 2023, I'm a postdoc in the group of Mattias Rantalainen at Karolinska Institutet in Stockholm (Department of Medical Epidemiology and Biostatistics), working on **machine learning** and **computer vision** for **computational pathology**.

My research focuses on how to build and evaluate **reliable machine learning** models, and has often included regression problems, uncertainty estimation methods or energy-based models.

Background:

- 2023: PhD in *Machine Learning*, Uppsala University.
  - Thesis: *Towards Accurate and Reliable Deep Regression Models*.
  - Supervisors: Thomas Schön & Martin Danelljan.
- 2018: MSc in *Electrical Engineering*, Linköping University.
- 2016: BSc in *Applied Physics and Electrical Engineering*, Linköping University.

About the Presentation

This presentation covers work done together with my supervisors while being a PhD student in Uppsala. The presentation is mainly based on our TMLR paper:

How Reliable is Your Regression Model's Uncertainty Under Real-World Distribution Shifts?
Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön
Transactions on Machine Learning Research (TMLR), 2023

Quite large parts of our previous CVPR Workshops paper will however also be covered, as this is relevant background material:

Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision
Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön
The Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), 2020

Summary of Contributions

How Reliable is Your Regression Model's Uncertainty Under Real-World Distribution Shifts?
Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön
Transactions on Machine Learning Research (TMLR), 2023

- We propose a benchmark for testing the reliability of regression uncertainty estimation methods under real-world distribution shifts.
- We then employ our benchmark to evaluate many of the most common uncertainty estimation methods, as well as two state-of-the-art uncertainty scores from out-of-distribution detection.
- We find that while all methods are well calibrated when there is no distribution shift, they become highly overconfident on many of the benchmark datasets – thus **uncovering important limitations** of current methods.
- This demonstrates that **more work is required** in order to develop truly reliable uncertainty estimation methods for regression.

Outline

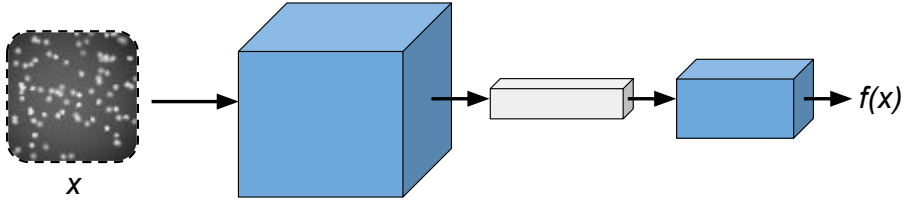
In a **supervised regression problem**, the task is to predict a *continuous* target value $y^\star \in \mathcal{Y} = \mathbb{R}^K$ for any given input $x^\star \in \mathcal{X}$. To solve this, we are also given a training set of i.i.d. input-target pairs, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, $(x_i, y_i) \sim p(x, y)$.

In this presentation, the focus will be on the 1D case, i.e. when $\mathcal{Y} = \mathbb{R}$.

The input space $\mathcal{X}$ will correspond to the space of images.

We view a **Deep Neural Network (DNN)** simply as a function $f_\theta : \mathcal{X} \to \mathcal{O}$, parameterized by $\theta \in \mathbb{R}^P$. This function maps inputs $x \in \mathcal{X}$ to outputs $f_\theta(x) \in \mathcal{O}$ in some output space $\mathcal{O}$.

We also divide the DNN f_θ into a *backbone feature extractor*, and one or more smaller *network heads*. The feature extractor takes x as input and outputs a feature vector $g(x)$, which is then fed into the network heads, producing the final output $f_\theta(x) \in \mathcal{O}$.



In a **supervised regression problem**, the task is to predict a *continuous* target value $y^\star \in \mathcal{Y} = \mathbb{R}^K$ for any given input $x^\star \in \mathcal{X}$. To solve this, we are also given a training set of i.i.d. input-target pairs, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, $(x_i, y_i) \sim p(x, y)$.

We view a **Deep Neural Network (DNN)** simply as a function $f_\theta : \mathcal{X} \to \mathcal{O}$, parameterized by $\theta \in \mathbb{R}^P$. This function maps inputs $x \in \mathcal{X}$ to outputs $f_\theta(x) \in \mathcal{O}$ in some output space $\mathcal{O}$.

The most common and straightforward deep regression approach is to let the DNN f_θ directly output predicted targets, $\hat{y}(x) = f_\theta(x)$, training the DNN by minimizing e.g. the L2 loss over the training data, $J(\theta) = \sum_{i=1}^N \big(y_i - f_\theta(x_i)\big)^2$.

Outline

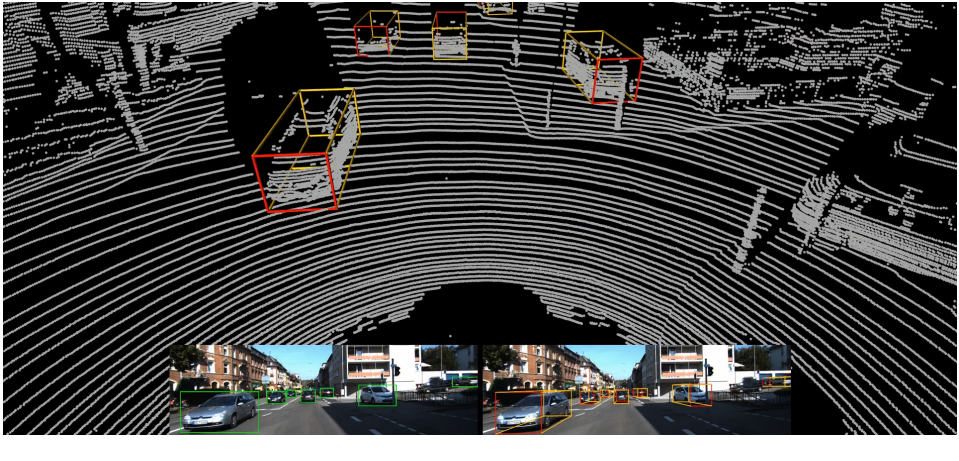
DNNs $f_\theta : \mathcal{X} \to \mathcal{O}$ have become the go-to approach within computer vision and many other domains due to their impressive predictive power. However, they generally fail to properly capture the uncertainty inherent in their predictions.

Bayesian deep learning is one approach that aims to address this issue in a principled manner. It deals with predictive uncertainty by decomposing it into the distinct types of *aleatoric* and *epistemic* uncertainty.

Aleatoric uncertainty captures *inherent* and *irreducible* ambiguity in the inputs x.

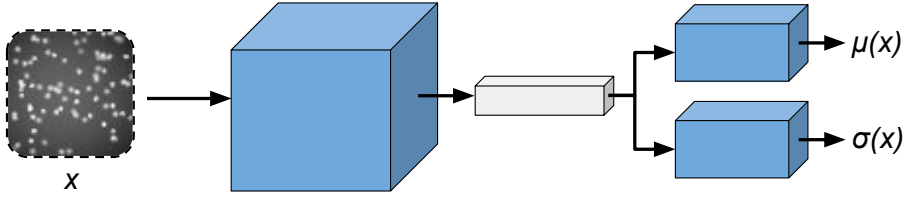
Epistemic uncertainty accounts for uncertainty in the DNN model parameters θ. More broadly, it is *reducible* uncertainty related to a lack of knowledge *("uncertainty due to things one could in principle know but does not in practice")*.

Input-dependent aleatoric uncertainty arises whenever the target y is expected to be inherently more uncertain for some inputs x than others. This is true e.g. in automotive 3D object detection, where it is inherently more difficult to estimate the 3D position and size of distant or partially occluded vehicles.



To estimate input-dependent aleatoric uncertainty, the DNN $f_\theta : \mathcal{X} \to \mathcal{O}$ can be used to specify a model $p(y|x;\theta)$ of the conditional target distribution.

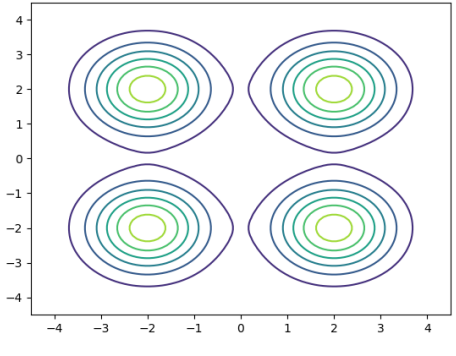
For example, a Gaussian model can be used, $p(y|x;\theta) = \mathcal{N}\big(y; \mu_\theta(x), \sigma^2_\theta(x)\big)$.



The mean can then be taken as a prediction, $\hat{y}(x) = \mu_\theta(x)$, whereas the variance $\sigma^2_\theta(x)$ naturally can be interpreted as a measure of aleatoric uncertainty for this prediction.

Using DNNs to specify models $p(y|x;\theta)$ of the conditional target distribution does however not capture **epistemic** uncertainty, as information about the uncertainty in the model parameters θ is disregarded.

Large epistemic uncertainty is present whenever a large set of model parameters explains the given training data (approximately) equally well.



This is often the case for DNNs, since the corresponding optimization landscapes are highly multi-modal.

Disregarding the epistemic model uncertainty can lead to highly confident yet incorrect predictions, especially for inputs x which are not well-represented by the training data.

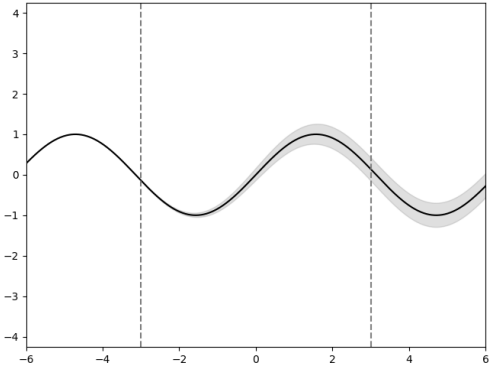
Epistemic uncertainty can be estimated in a principled manner by performing *approximate Bayesian inference*.

Instead of just finding a single point estimate $\hat{\theta}$ of the model parameters θ, by minimizing the negative log-likelihood $\mathcal{L}(\theta) = \sum_{i=1}^{N} -\log p(y_i|x_i;\theta)$ over the training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, Bayesian inference entails estimating the full *posterior distribution* $p(\theta|\mathcal{D})$.

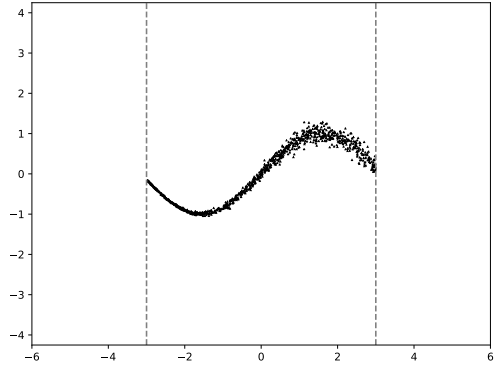
The posterior $p(\theta|\mathcal{D})$ is obtained from the data likelihood $\prod_{i=1}^{N} p(y_i|x_i;\theta)$ and a chosen prior $p(\theta)$ by applying Bayes' theorem, $p(\theta|\mathcal{D}) \propto \prod_{i=1}^{N} p(y_i|x_i;\theta)p(\theta)$.

Let us consider the following simple 1D regression problem:

$$p(y|x) = \mathcal{N}\big(y; \mu(x), \sigma^2(x)\big), \qquad \mu(x) = \sin(x), \qquad \sigma(x) = \frac{0.15}{1 + e^{-x}}.$$

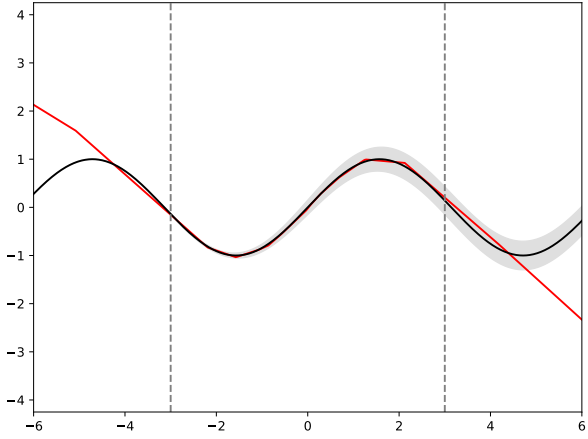


(a) True data generator $p(y|x)$.

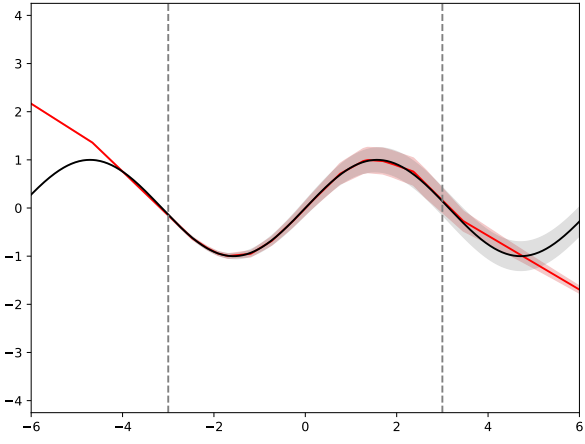


(b) Training dataset $\{(x_i, y_i)\}_{i=1}^{1000}$.

A DNN f_θ trained to directly output predicted targets, $\hat{y}(x) = f_\theta(x)$, is able to accurately regress the mean $\mu(x) = \sin(x)$ for $x \in [-3, 3]$. However, this model fails to capture any notion of uncertainty.

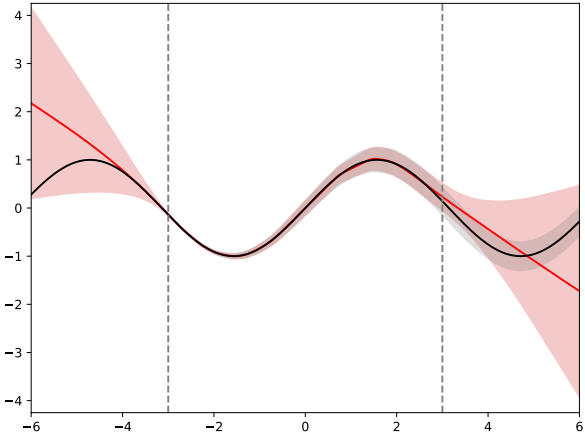


Instead, the DNN f_θ can be used to specify a Gaussian model $p(y|x;\theta) = \mathcal{N}(y; \mu_\theta(x), \sigma^2_\theta(x))$, trained by minimizing the NLL $\mathcal{L}(\theta)$. The model closely matches the true $p(y|x)$ for $x \in [-3, 3]$, accounting for *aleatoric* uncertainty.



For inputs $|x| > 3$ not seen during training, however, the estimated mean $\mu_\theta(x)$ deviates significantly from the true $\mu(x) = \sin(x)$, while the estimated uncertainty $\sigma^2_\theta(x)$ remains very small. That is, the model becomes **overconfident** for inputs $|x| > 3$.

The Gaussian DNN model $p(y|x;\theta) = \mathcal{N}\big(y; \mu_\theta(x), \sigma^2_\theta(x)\big)$ can instead be estimated via approximate Bayesian inference, in order to account for both *aleatoric* and *epistemic* uncertainty.



The model now predicts a more reasonable uncertainty $\sigma^2_\theta(x)$ in the region with no available training data.

While the estimated mean $\mu_\theta(x)$ still deviates from the true $\mu(x) = \sin(x)$ for $|x| > 3$, the uncertainty $\sigma^2_\theta(x)$ also increases accordingly – the model does *not* become overconfident.

Outline

Given a desired miscoverage rate $\alpha \in]0, 1[$, a **prediction interval** $C_\alpha(x^\star) = [L_\alpha(x^\star),\ U_\alpha(x^\star)] \subseteq \mathbb{R}$ is a function that maps the input $x^\star$ onto an interval that should cover the true regression target $y^\star$ with probability $1 - \alpha$.

For any set $\{(x_i^\star, y_i^\star)\}_{i=1}^{N^\star}$ of $N^\star$ examples, the empirical **interval coverage** is the proportion of inputs for which the prediction interval covers the target,

$$\text{Coverage}(C_\alpha) = \frac{1}{N^\star} \sum_{i=1}^{N^\star} \mathbb{I}\{y_i^\star \in C_\alpha(x_i^\star)\}. \tag{1}$$

If the coverage equals $1 - \alpha$, we say that the prediction intervals are **perfectly calibrated**. Unless stated otherwise, we here set $\alpha = 0.1$. The prediction intervals should thus obtain a coverage of 90%.

Prediction interval: $C_\alpha(x^\star) = [L_\alpha(x^\star),\ U_\alpha(x^\star)] \subseteq \mathbb{R}$.

Empirical interval coverage: $\mathrm{Coverage}(C_\alpha) = \frac{1}{N^\star} \sum_{i=1}^{N^\star} \mathbb{I}\{y_i^\star \in C_\alpha(x_i^\star)\}$.

With a trained **Gaussian DNN model** $p(y|x;\theta) = \mathcal{N}\big(y; \mu_\theta(x),\ \sigma_\theta^2(x)\big)$, a prediction interval $C_\alpha(x^\star)$ for a given input $x^\star$ can be constructed as,

$$C_\alpha(x^\star) = [\mu_\theta(x^\star) - \sigma_\theta(x^\star)\Phi^{-1}(1-\alpha/2),\ \mu_\theta(x^\star) + \sigma_\theta(x^\star)\Phi^{-1}(1-\alpha/2)],$$

where Φ is the CDF of the standard normal distribution.

With a trained Gaussian DNN model $p(y|x;\theta) = \mathcal{N}\big(y; \mu_\theta(x), \sigma^2_\theta(x)\big)$, a prediction interval $C_\alpha(x^\star)$ for a given input $x^\star$ can be constructed as,

$$C_\alpha(x^\star) = [\mu_\theta(x^\star) - \sigma_\theta(x^\star)\Phi^{-1}(1-\alpha/2),\ \mu_\theta(x^\star) + \sigma_\theta(x^\star)\Phi^{-1}(1-\alpha/2)], \quad (2)$$

where Φ is the CDF of the standard normal distribution.

With a trained **ensemble** $\{f_{\theta^{(m)}}\}_{m=1}^M$ of M such Gaussian DNN models, a single mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ can be computed as,

$$\hat{\mu}(x^\star) = \frac{1}{M}\sum_{m=1}^{M} \mu_{\theta^{(m)}}(x^\star), \qquad \hat{\sigma}^2(x^\star) = \frac{1}{M}\sum_{m=1}^{M}\Big(\big(\hat{\mu}(x^\star) - \mu_{\theta^{(m)}}(x^\star)\big)^2 + \sigma^2_{\theta^{(m)}}(x^\star)\Big),$$

and then plugged into (2) to construct a prediction interval $C_\alpha(x^\star)$ for the input $x^\star$.

Outline

The general idea of selective prediction is to give a model the option to abstain from outputting predictions for some inputs.

This is achieved by combining the prediction model f_θ with an uncertainty function $\kappa_f : \mathcal{X} \to \mathbb{R}$. Given an input $x^\star$, the prediction $f_\theta(x^\star)$ is output if the uncertainty $\kappa_f(x^\star) \leq \tau$, otherwise $x^\star$ is rejected and no prediction is made.

The **prediction rate** is the proportion of inputs for which a prediction is output,

$$\text{Predition Rate} = \frac{1}{N^\star} \sum_{i=1}^{N^\star} \mathbb{I}\{\kappa_f(x_i^\star) \leq \tau\}. \tag{3}$$

We combine selective prediction with standard regression methods. A prediction interval $C_\alpha(x^\star)$ and predicted target $\hat{y}(x^\star)$ are thus output if and only if $\kappa_f(x^\star) \leq \tau$. Our aim is for this to improve the calibration of the output prediction intervals.

For $\kappa_f(x)$, the variance $\hat{\sigma}^2(x)$ of a Gaussian ensemble could be used, for example.

One could also use some of the various uncertainty scores employed in the rich **out-of-distribution (OOD) detection** literature. In OOD detection, the task is to distinguish in-distribution inputs x, inputs which are similar to those of the training set $\{(x_i, y_i)\}_{i=1}^N$, from out-of-distribution inputs.

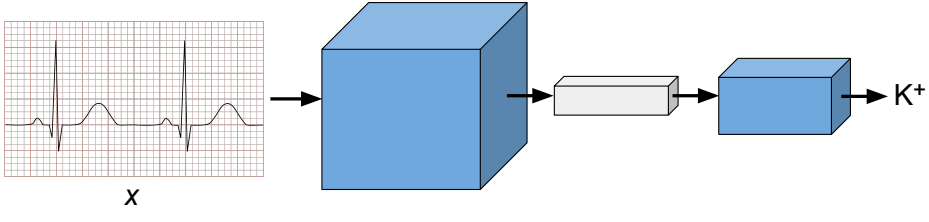
A principled approach to OOD detection would be to fit a model of $p(x)$ on the training set. Inputs x for which $p(x)$ is small are then deemed OOD. One can also fit a simple model to the feature vectors $g(x)$, modelling $p(x)$ indirectly.

Outline

How Reliable is Your Regression Model's Uncertainty Under Real-World Distribution Shifts?
Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön
Transactions on Machine Learning Research (TMLR), 2023

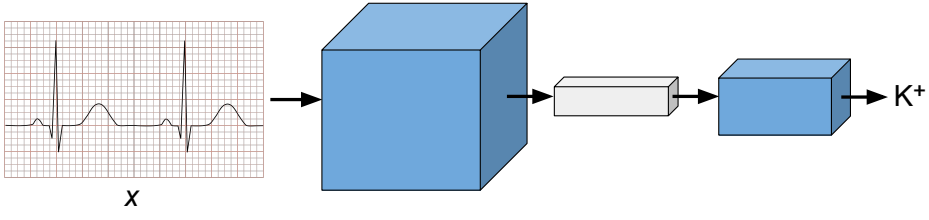
- We propose a benchmark for testing the reliability of regression uncertainty estimation methods under real-world distribution shifts.
- We then employ our benchmark to evaluate many of the most common uncertainty estimation methods, as well as two state-of-the-art uncertainty scores from out-of-distribution detection.
- We find that while all methods are well calibrated when there is no distribution shift, they become highly overconfident on many of the benchmark datasets – thus **uncovering important limitations** of current methods.
- This demonstrates that **more work is required** in order to develop truly reliable uncertainty estimation methods for regression.

Outline



Much of the work in the *How Reliable is Your Regression Model's Uncertainty Under Real-World Distribution Shifts?* paper was inspired and motivated by concurrent work on **ECG-based electrolyte prediction**.

Abnormal potassium (K^+) concentration levels in the human body can lead to serious heart conditions. If the concentration could be accurately monitored using an ECG-based regression model, potentially life-threatening conditions could be avoided.



We trained a DNN on this task and obtained reasonable regression accuracy *(to train the model, we utilized a large-scale dataset of over* 290 000 *ECGs from adult patients attending emergency departments at Swedish hospitals)*.

During this work, we started thinking more carefully about the question: **Would it be possible to actually deploy this model in clinical practice at the university hospital?** What requirements would such real-world deployment within a safety-critical domain put on this deep regression model?

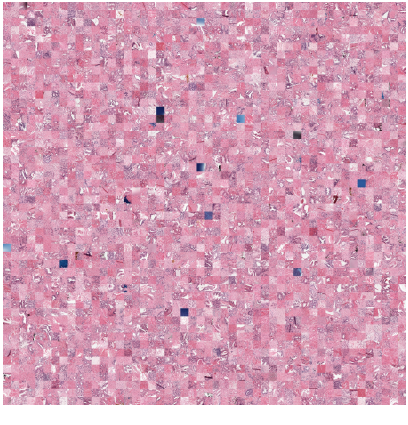
The model must at least be **well calibrated**. If it outputs a prediction and a 90% prediction interval for each input, 90% interval coverage should actually be achieved.

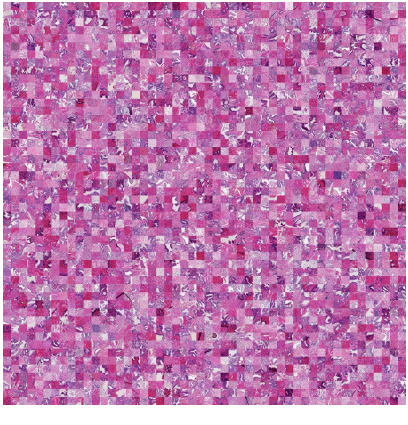
Otherwise, if the model becomes overconfident and outputs highly confident yet incorrect predictions, providing uncertainty estimates might just instill a false sense of security – arguably making the model even less suitable for safety-critical deployment.

Moreover, the model must remain well calibrated also under the wide variety of **distribution shifts** that might be encountered during practical deployment.

For example, a model trained on data collected solely at a large urban hospital in the year 2020, for instance, should output well-calibrated predictions also in 2024, for patients both from urban and rural areas.

One concrete example of a real-world distribution shift from the computational pathology domain – tissue image patches from a lab in the Netherlands vs from a lab in Sweden (publicly available data from *Artificial Intelligence for Diagnosis and Gleason Grading of Prostate Cancer: The PANDA Challenge* (Nature Medicine, 2022)):





Outline

7. Proposed Benchmark - Datasets

	Cells-Tails	ChairAngle-Gap	AssetWealth	Ventricular Volume	BrainTumour Pixels	SkinLesion Pixels	Histology NucleiPixels	AerialBuilding Pixels
Train								
	y = 100	y = 80.5	y = 1.594	y = 40.38	y = 252	y = 500	y = 1257	y = 1097
Test								
	y = 198	y = 43.8	y = 1.314	y = 85.93	y = 273	y = 516	y = 1156	y = 433

We collect 8 publicly available datasets for different image-based regression tasks, with various types of distribution shifts *(e.g., train on satellite images captured in densely populated American cities – test on images captured in a rural European area)*.

2 synthetic datasets, 6 real-world datasets. 6 592 - 20 614 training images.

We evaluate regression uncertainty estimation methods mainly in terms of **prediction interval coverage**, $\text{Coverage}(C_\alpha) = \frac{1}{N^\star} \sum_{i=1}^{N^\star} \mathbb{I}\{y_i^\star \in C_\alpha(x_i^\star)\}$.

If a method outputs a prediction $\hat{y}(x)$ and a 90% prediction interval $C_{0.1}(x)$ for each input x, does the method actually achieve 90% coverage on the *test* set? I.e., are the prediction intervals calibrated?

We also evaluate in terms of *average interval length* on the val set. This is a natural secondary metric, since a method that achieves a coverage close to $1 - \alpha$ but outputs extremely large intervals for all inputs x, would not be particularly useful in practice.

For methods based on selective prediction, the only difference is that predictions $\hat{y}(x)$ and prediction intervals $C_\alpha(x)$ are output only for some test inputs x (iff $\kappa_f(x) \leq \tau$). The prediction interval coverage is thus computed only on this subset of test.

For these methods, the proportion of inputs for which a prediction actually is output is another natural secondary metric. We thus also evaluate in terms of the *prediction rate* $\frac{1}{N^\star} \sum_{i=1}^{N^\star} \mathbb{I}\{\kappa_f(x_i^\star) \leq \tau\}$ on test.

If a coverage close to $1 - \alpha$ is achieved with a very low prediction rate, the method might still not be practically useful in certain applications.

Outline

We evaluate five common regression uncertainty estimation methods, which all output a 90% prediction interval $C_{0.1}(x)$ and a predicted target $\hat{y}(x) \in C_{0.1}(x)$ for each input.

Two of these methods we also combine with selective prediction, utilizing four different uncertainty functions $\kappa_f(x)$.

In total, we evaluate 10 different methods.

We calibrate the prediction intervals, for each of the 10 *methods, such that exactly* 90% *interval coverage is obtained on the val set. Ideally, the coverage should then not change from the val set to the test set.*

Conformal Prediction.

Ensemble.

Gaussian.

Gaussian Ensemble.

Quantile Regression.

Gaussian + Selective GMM.

- Combining a Gaussian model $p(y|x;\theta) = \mathcal{N}\big(y; \mu_\theta(x), \sigma^2_\theta(x)\big)$ with selective prediction. A GMM is fit to the feature vectors $\{g(x_i)\}_{i=1}^N$ of the training set. The GMM likelihood is then taken as the uncertainty score, $\kappa_f(x) = -\mathrm{GMM}\big(g(x)\big)$.

Gaussian + Selective kNN.

- The average distance from $g(x)$ to its k nearest neighbors among the train feature vectors $\{g(x_i)\}_{i=1}^N$ is taken as the uncertainty score, $\kappa_f(x) = \mathrm{kNN}\big(g(x)\big)$.

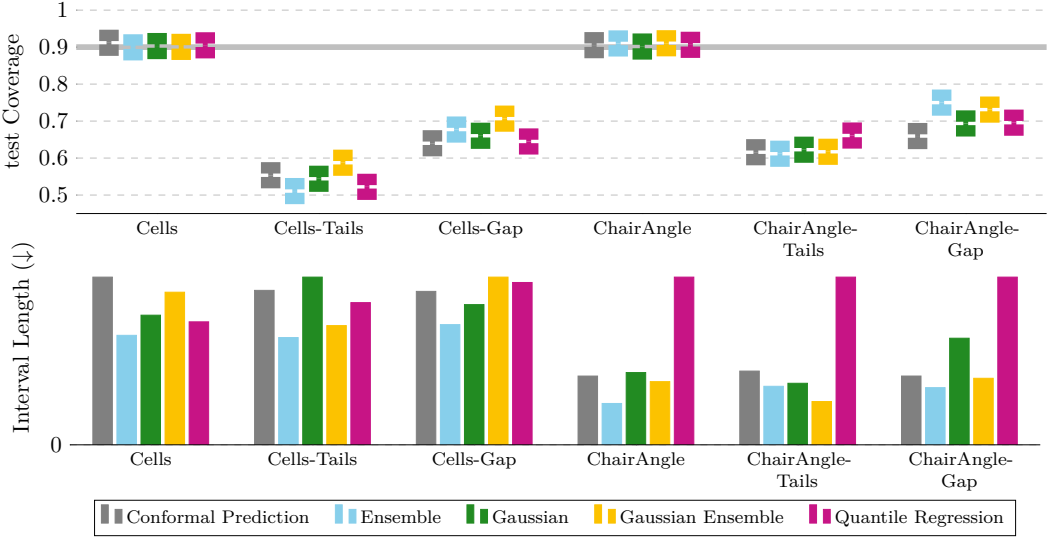
Gaussian + Selective Variance.

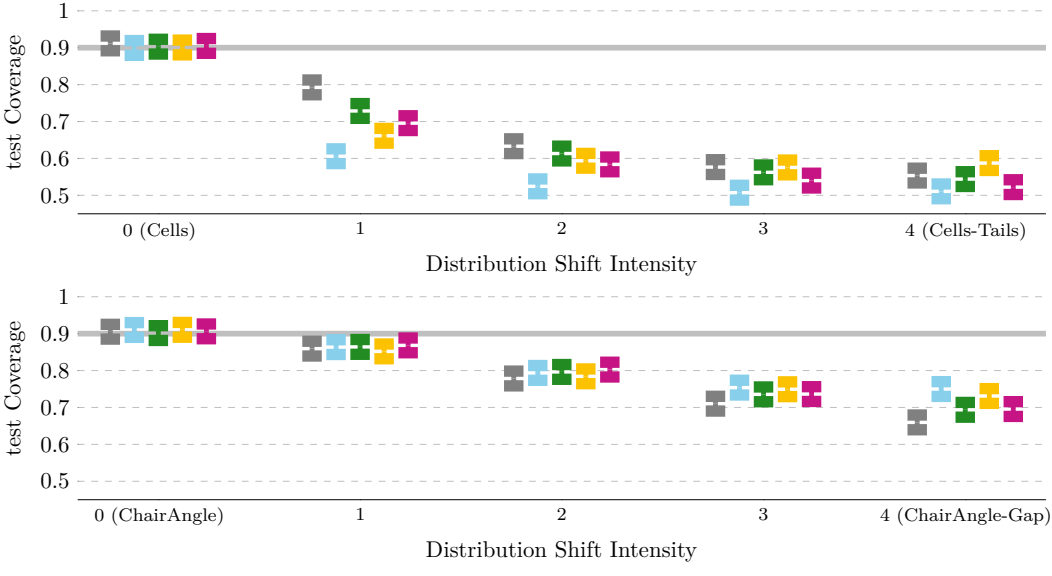
Gaussian Ensemble + Selective GMM.

Gaussian Ensemble + Selective Ensemble Variance

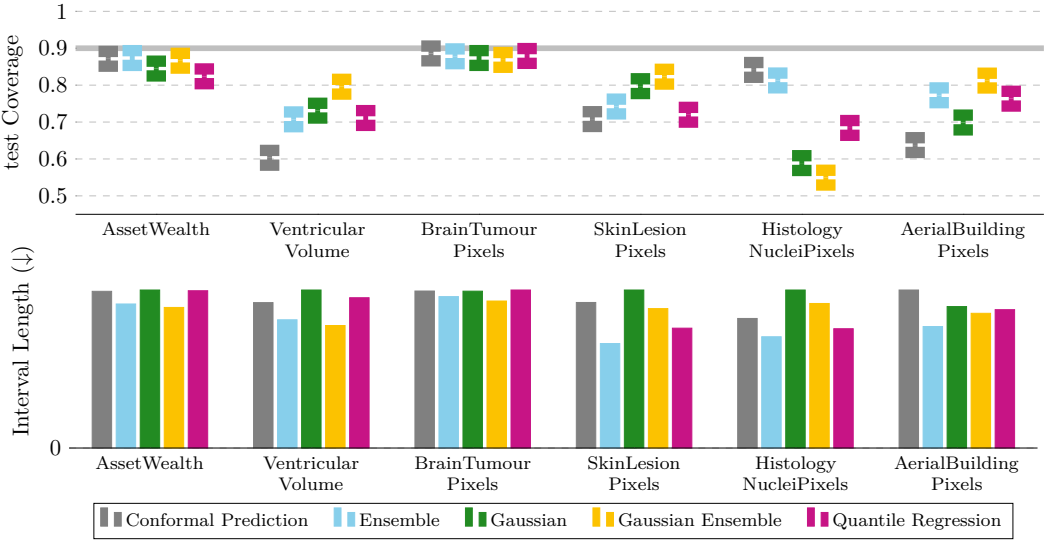
Outline

9. Results - Common Uncertainty Estimation Methods - Synthetic

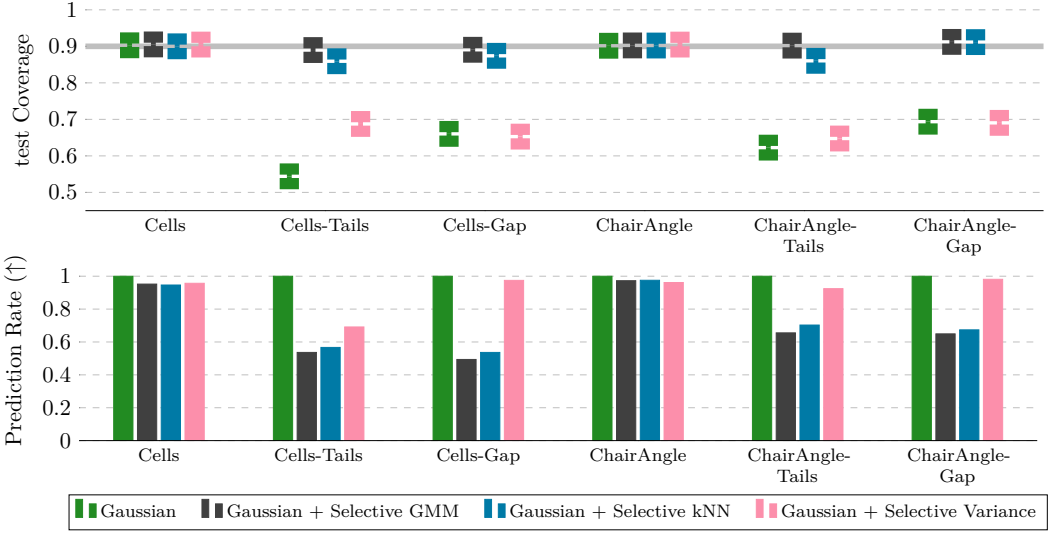



test Coverage
1
0.9
0.8
0.7
0.6
0.5
0 (Cells)
1
2
3
4 (Cells-Tails)
Distribution Shift Intensity
test Coverage
1
0.9
0.8
0.7
0.6
0.5
0 (ChairAngle)
1
2
3
4 (ChairAngle-Gap)
Distribution Shift Intensity

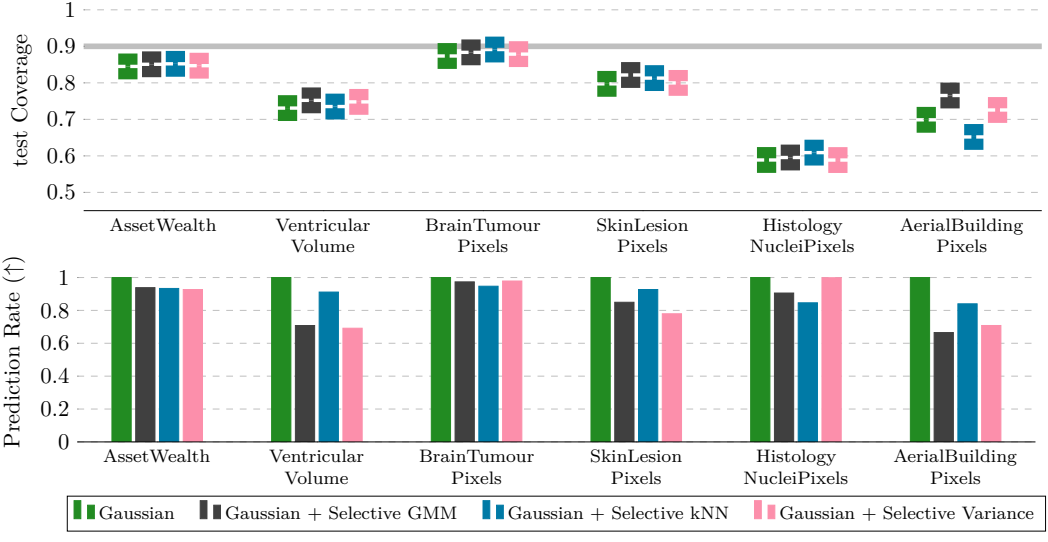
9. Results - Common Uncertainty Estimation Methods - Real-World



9. Results - Selective Prediction Methods - Synthetic Datasets



9. Results - Selective Prediction Methods - Real-World Datasets



Outline

(1/4) All methods are well calibrated on baseline datasets with no distribution shift, but become highly overconfident in many realistic scenarios. Uncertainty estimation methods must therefore be evaluated using sufficiently challenging benchmarks. Otherwise, one might be lead to believe that methods will be more reliable during real-world deployment than they actually are.

(2/4) Conformal prediction methods have commonly promoted theoretical coverage guarantees, but these depend on an assumption that is unlikely to hold in many practical applications. Consequently, also these methods can become highly overconfident in realistic scenarios. If the underlying assumptions are not examined critically by practitioners, such theoretical guarantees risk instilling a false sense of security – making these models even less suitable for safety-critical deployment.

(3/4) The clear performance difference between synthetic and real-world datasets observed for selective prediction methods based on feature-space density is an interesting direction for future work. If the reasons for this performance gap can be understood, an uncertainty estimation method that stays well calibrated across all datasets could potentially be developed.

(4/4) Selective prediction methods based on feature-space density perform well relative to other methods (as expected based on their state-of-the-art OOD detection performance), but are also overconfident in many cases. Only comparing the relative performance of different methods is therefore not sufficient. To track if actual progress is being made towards the ultimate goal of truly reliable uncertainty estimation methods, benchmarks must also evaluate method performance in an absolute sense.

Summary of Contributions (Repetition)

How Reliable is Your Regression Model's Uncertainty Under Real-World Distribution Shifts?
Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön
Transactions on Machine Learning Research (TMLR), 2023

- We propose a benchmark for testing the reliability of regression uncertainty estimation methods under real-world distribution shifts.
- We then employ our benchmark to evaluate many of the most common uncertainty estimation methods, as well as two state-of-the-art uncertainty scores from out-of-distribution detection.
- We find that while all methods are well calibrated when there is no distribution shift, they become highly overconfident on many of the benchmark datasets – thus **uncovering important limitations** of current methods.
- This demonstrates that **more work is required** in order to develop truly reliable uncertainty estimation methods for regression.

Contact & Feedback

Fredrik K. Gustafsson
fredrik.gustafsson@ki.se
www.fregu856.com

Please feel free to leave any type of **anonymous** feedback on this presentation:
www.fregu856.com/post/feedback