Uncertainty Quantification Metrics for Deep Regression

Ziliang Xiong, Simon Kristofferson Lind, Per-Erik Forssén, Volker Krüger

LINKÖPING UNIVERSITY

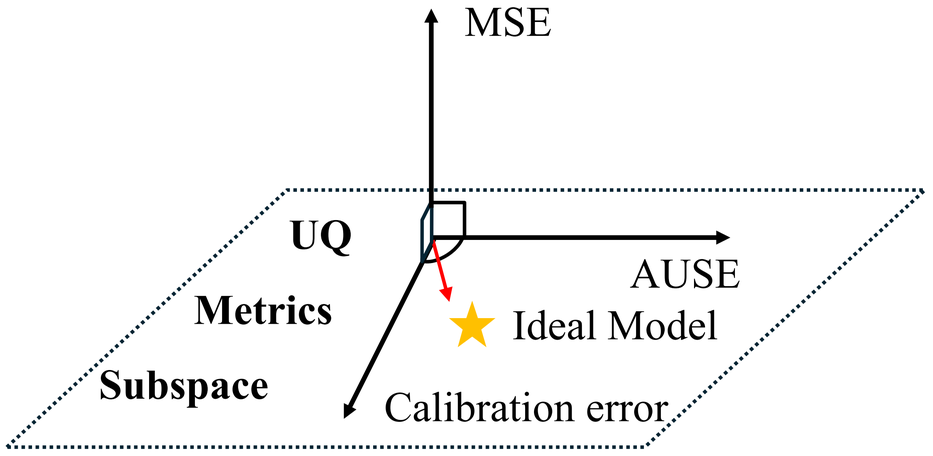
LUND UNIVERSITY

Research Goals

Here we address metrics for uncertainty quantification with focus on **regression tasks.** We try to answer these questions:

- *What do different evaluation metrics measure?*
- *Are they stable and robust with limited data?*
- *What are their strengths and weaknesses?*

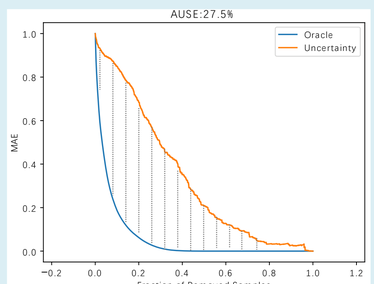
Our results indicate that **Calibration Error is the most stable and interpretable metric, but AUSE and NLL also have their respective use cases**. We discourage the usage of Spearman's Rank Correlation and recommend replacing it with AUSE.



Uncertainty Evaluation Metrics

Area Under Sparsification Error Curve

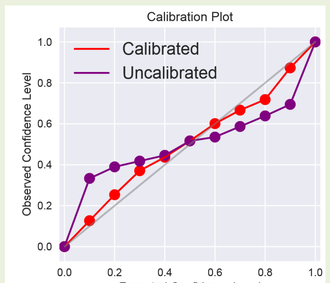
It assesses how well the predicted uncertainty coincides with the prediction error on a test set.



$$\mathrm{AUSE}(S) = \int_0^1 \frac{\mathrm{MAE}(S_{\vee}^{U}(\alpha))}{\mathrm{MAE}(S)} - \frac{\mathrm{MAE}(S_{\vee}^{\mathrm{AE}}(\alpha))}{\mathrm{MAE}(S)} d\alpha$$

Calibration Error

The difference between cumulative density function and empirical frequency.



$$\mathrm{cal}(\hat{p}_1, \cdots, \hat{p}_N) = \sum_{j=1}^{M} w_j (p_j - \hat{p}_j)^2$$

Empirical frequency

$$\hat{p}_j = \frac{|\{y_i \mid P(y < y_i|\theta) \leq p_j, (x_i, y_i) \in S\}|}{N}$$

Rank based | Density based

Spearman's Rank Correlation

Correlation between rank of AE and uncertainty scalar.

$$\rho_{R(L_1),R(L_2)} = \frac{\mathrm{cov}(R(L_1), R(L_2))}{\sigma_{R(L_1)} \sigma_{R(L_2)}}$$

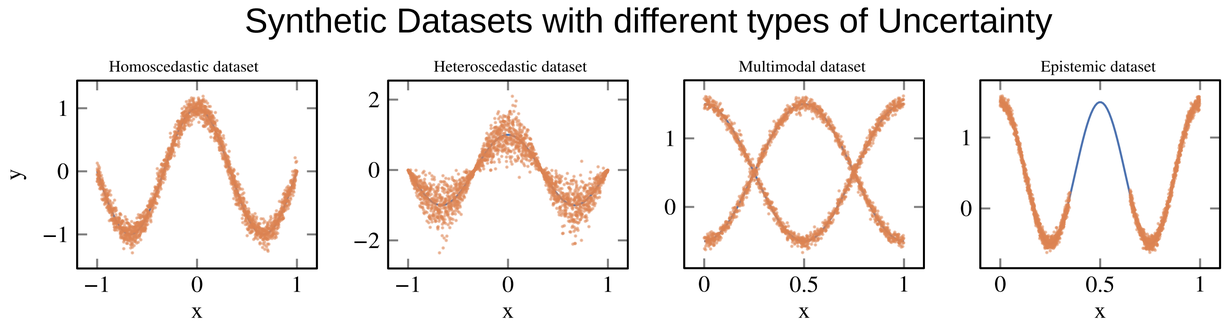
Negative Log Likelihood

NLL is minimized if and only if prediction is equal to the true underlying data distribution.

$$\mathrm{NLL}(S) = -\sum_{i=1}^{N} \log p(y_i|x_i;\theta)$$

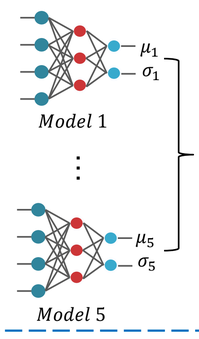
Methods

Synthetic Datasets with different types of Uncertainty



Two Typical Uncertainty-aware Regression Models

Deep Ensemble



Gaussian mixture approximated by one Gaussian $\hat{\mu}, \hat{\sigma}$

Energy based Regression

$$p(y|x;\theta) = \frac{e^{f(x,y;\theta)}}{\int e^{f(x,\tilde{y};\theta)} d\tilde{y}}$$

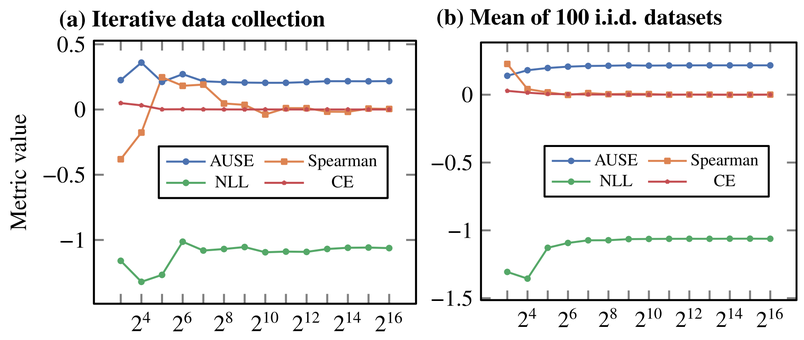
$$\mathcal{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \log \int e^{f(x,\tilde{y};\theta)} d\tilde{y} - f(x_i, y_i; \theta)$$

Selected Results

Stability on varying test set sizes

Experiment 1: How quickly does each metric converge to its expected value?

Experiment 2: Are the estimates unbiased for small test set sizes?



Conclusions

1. Stability: CE > AUSE > NLL > Spearman
2. Spearman converges to zero, should not be used with dense samples
3. All metrics converge beyond a dataset size of 1024
4. No metric exhibits any meaningful bias beyond a dataset size of 2^6

Metrics under different types of uncertainty

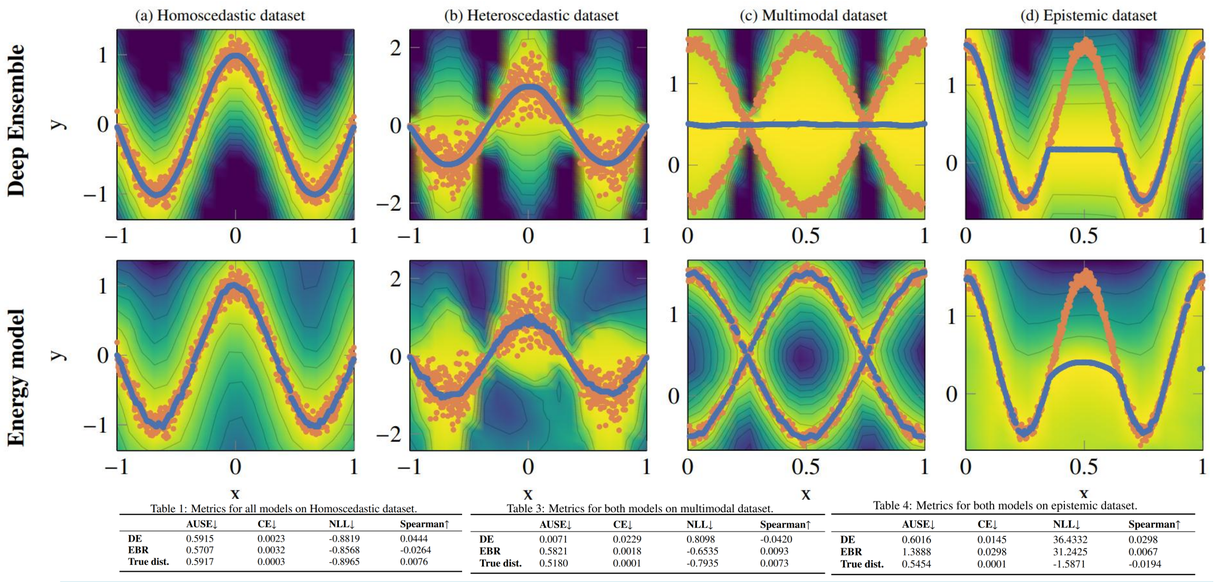


Table 1: Metrics for all models on Homoscedastic dataset.

	AUSE↓	CE↓	NLL↓	Spearman↑
DE	0.5915	0.0023	-0.8819	0.0444
EBR	0.5707	0.0032	-0.8568	-0.0264
True dist.	0.5917	0.0003	-0.8965	0.0076

Table 3: Metrics for both models on multimodal dataset.

	AUSE↓	CE↓	NLL↓	Spearman↑
DE	0.0071	0.0229	0.8098	-0.0420
EBR	0.5821	0.0018	-0.6535	0.0093
True dist.	0.5180	0.0001	-0.7935	0.0073

Table 4: Metrics for both models on epistemic dataset.

	AUSE↓	CE↓	NLL↓	Spearman↑
DE	0.6016	0.0145	36.4332	0.0298
EBR	1.3888	0.0298	31.2425	0.0067
True dist.	0.5454	0.0001	-1.5871	-0.0194

Conclusions

1. Homo and Hetero: Perfect Predicted distribution returns good CE, NLL but not good ranking.
2. Multimodal: AUSE for DE model is good, but calibration and NLL suggest that the model fails to capture generating distribution.

⇒ Perfect Calibration is not equaivalent to perfect Ranking.

Interpretability

- **CE:** lower bound of 0, and is highly interpretable; It requires the least amount of samples to be stable
- **AUSE:**unbounded and thus it lacks interpretability; robust ranking metrics, suitable for practical tasks
- **NLL:**less stable with small test set
- **Spearman**: Unsuitable for dense samples

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM