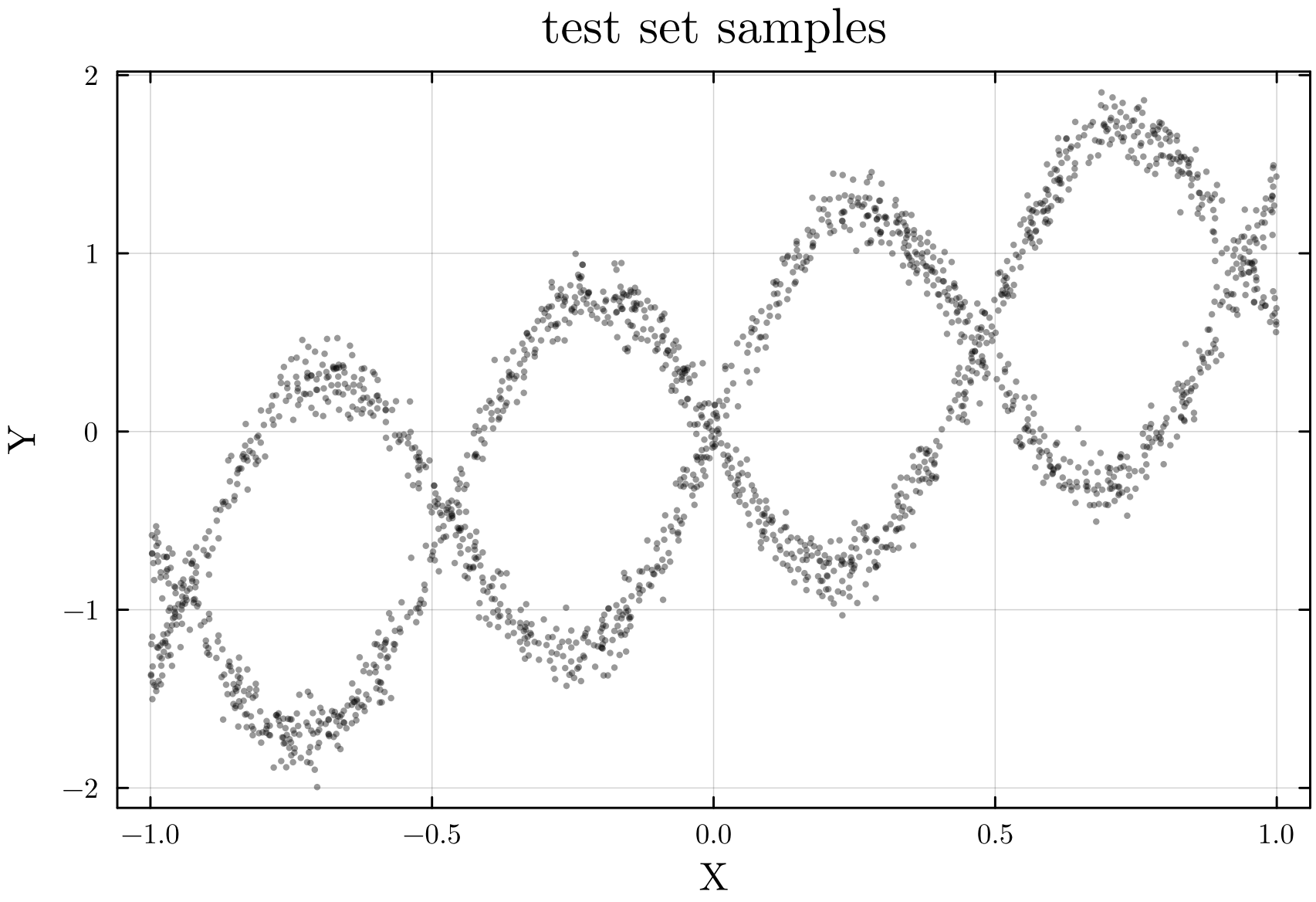
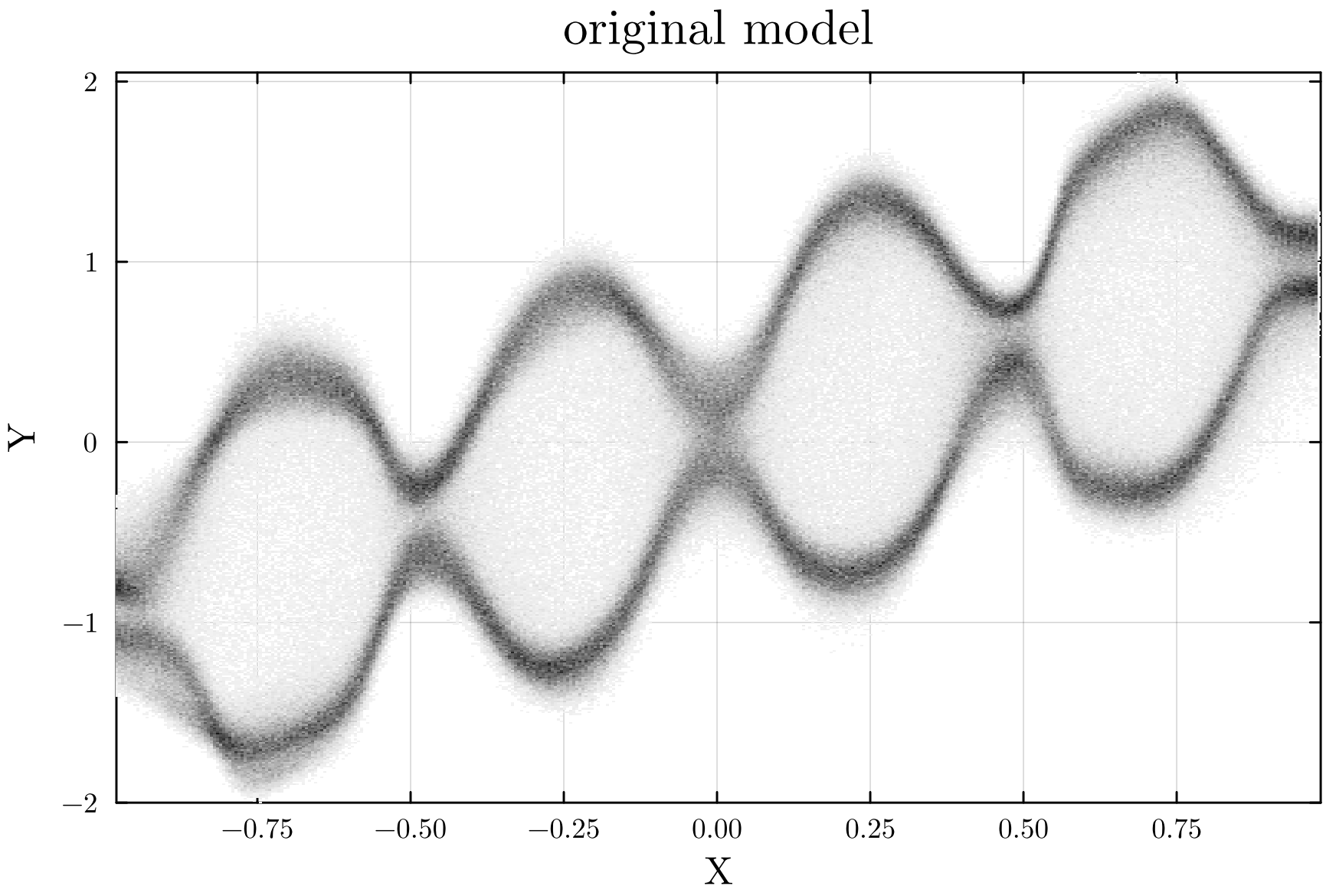
Calibration

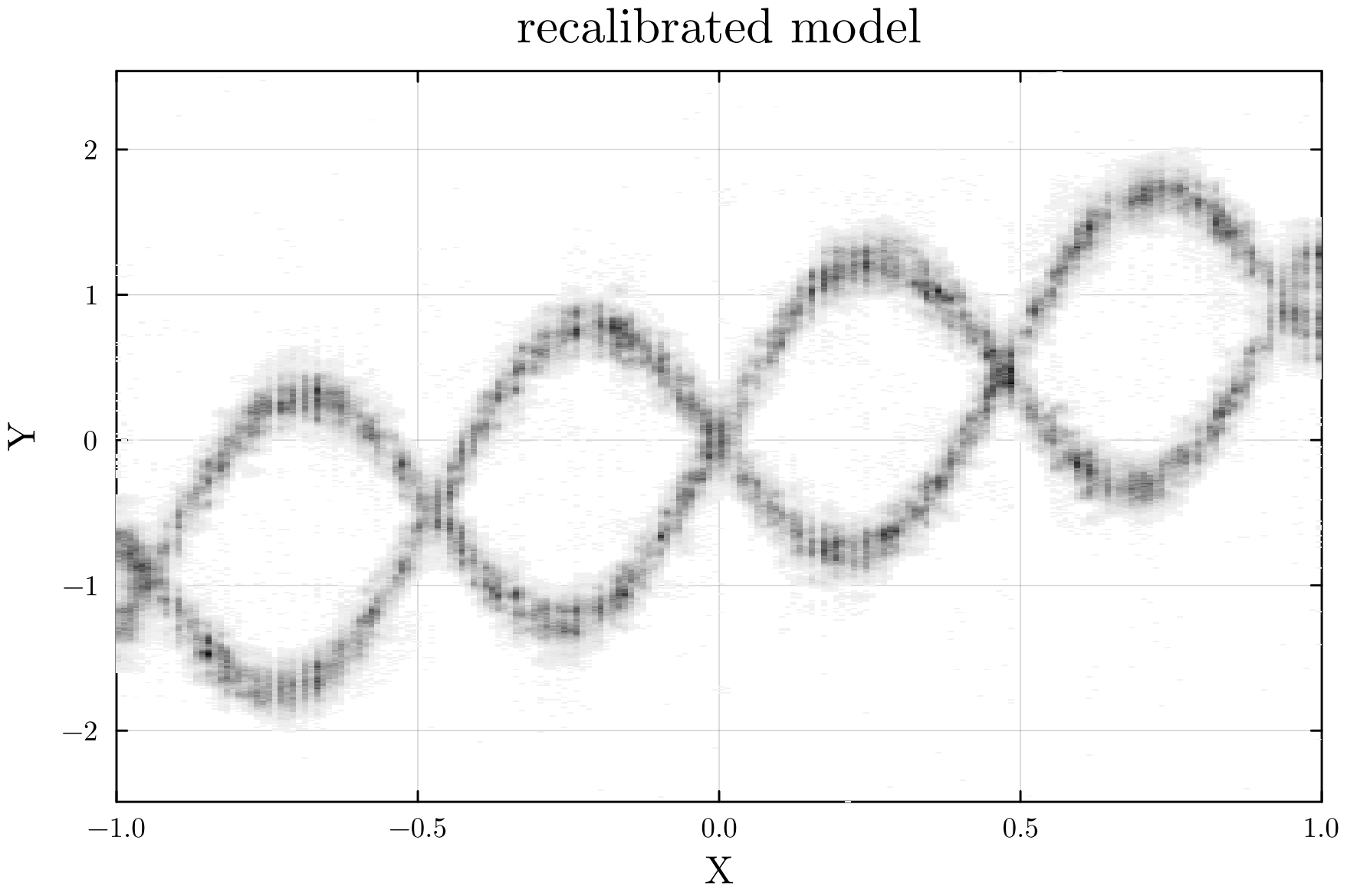
A Nonparametric Distribution Regression Re-calibration Method

Ádám Jung
Domokos M. Kelen
András A. Benczúr

2025 Spring -


test set samples
2
1
0
−1
−2
Y
−1.0
−0.5
0.0
0.5
1.0
X


original model
2
1
0
−1
−2
Y
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
X


recalibrated model
2
1
0
−1
−2
Y
−1.0
−0.5
0.0
0.5
1.0
X

Intuition of calibration

Two ways to improve the expected score (e.g. NLL) of a model :

i) Make the predictions **sharper**, i.e. better discriminate the events for which the conditional distribution of the target is different $\implies$ lower residual entropy.

ii) Make better **calibrated** predictions, i.e. there is less divergence between the predictions and observations.

i) Make the predictions **sharper**, i.e. better discriminate the events for which the conditional distribution of the target is different $\implies$ lower residual entropy.

$\Updownarrow$

Examples of models with lack of sharpness :

- A model that uses only a subset of the features.
- A model with too strong regularization.

ii) Make better **calibrated** predictions, i.e. there is less divergence between the predictions and observations.

$$\Updownarrow$$

- If only the predictions and observations are visible
- original features hidden,
- impossible to distinguish real observations from generated ones

Definitions

Let the features X, response Y and predictions Q follow the joint distribution

$$(X, Y, Q) \sim \mathbb{P}_{X,Y,Q} \ .$$

- Note that the prediction Q is treated as a random variable.
- A realization of Q is a distribution of Y.

With this notation our ultimate goal is to have

$$Q|X = \mathbb{P}_{Y|X} \ . \tag{1}$$

Calibration can be defined via

$$Q = \mathbb{P}_{Y|Q} \ , \tag{2}$$

and quantified with

$$\mathbb{E}\left[d\left(\mathbb{P}_{Y|Q} \middle\| Q\right)\right] \ . \tag{3}$$

An example

- Assume Q is a random element of a μ, σ^2 parameter family.
- Then the parameter is a random vector.
- Eq. (2) $\iff$ *When the predicted mean and variance is* μ, σ^2 *the target should have* $\mathbb{E}[Y] = \mu, \mathrm{Var}(Y) = \sigma^2$.

Lack of sharpness can be quantified as

$$\mathbb{E}\left[H(\mathbb{P}_{Y|Q}) - H(\mathbb{P}_{Y|X})\right] \ , \tag{4}$$

which is actually the conditional mutual information

$$I(Y; X|Q) \geq 0 \ , \tag{5}$$

0 iff Y and X are independent, given Q.

Two ways to improve the expected score (e.g. NLL) of a model :

i) Make the predictions **sharper**, i.e. better discriminate the events for which the conditional distribution of the target is different $\implies$ lower residual entropy.

ii) Make better **calibrated** predictions, i.e. there is less divergence between the predictions and observations.

Expected Score decomposition (based on [3, Lemma 4.1.])

$$\mathbb{E}\left[S(Q,Y)\right] = \underbrace{\mathbb{E}\left[d\left(\mathbb{P}_{Y|Q} \,\middle\|\, Q\right)\right]}_{\text{calibration error}} + \underbrace{I(Y;X|Q)}_{\text{lack of sharpness}} + \underbrace{\mathbb{E}\left[H(\mathbb{P}_{Y|X})\right]}_{\text{aleatoric uncertainty}} .$$

$$\text{calibration error} + \text{lack of sharpness} = \mathbb{E}\left[d\left(\mathbb{P}_{Y|X} \,\middle\|\, Q\right)\right]$$

Goal

Maintain the given sharpness, and eliminate the calibration error.

Solution

Distribution regression, where the features are predictions :

1) Split data indices to $\mathcal{D}_{train} \cup \mathcal{D}_{calibration} \cup \mathcal{D}_{test}$.

2) Estimate $\mathbb{P}_{Y|X}$ using $\{(x_i, y_i) \mid i \in \mathcal{D}_{train}\}$.

3) Make predictions $q_i \sim Q|\{X = x_i\}$ for the calibration and test set.

4) On the dataset $\{(q_i, y_i) \mid i \in \mathcal{D}_{calibration}\}$ train a regression model $\tilde{Q}$ that estimates $\mathbb{P}_{Y|Q}$.

5) Make predictions $\tilde{q}_i \sim \tilde{Q}|\{Q = q_i\}$ on the test set :

$$X = x_i \quad \rightsquigarrow \quad q_i \sim Q|\{X = x_i\} \quad \rightsquigarrow \quad \tilde{q}_i \sim \tilde{Q}|\{Q = q_i\}$$

Implementation details

4) On the dataset $\{(q_i, y_i) \mid i \in \mathcal{D}_{calibration}\}$ train a regression model $\tilde{Q}$ that estimates $\mathbb{P}_{Y|Q}$.

We estimate $\mathbb{P}_{Y|Q}$ via CKMEs. For that we should have a kernel on the new *feature* Q.

Nonparametric kernel on distributions

$$k(q_1, q_2) = \exp\left\{-\frac{1}{\sigma_k}\left(\mathbb{E}|M_1 - M_2| - \frac{\mathbb{E}|M_1 - M_1'| + \mathbb{E}|M_2 - M_2'|}{2}\right)\right\}$$

where $M_1, M_1' \sim q_1$, $M_2, M_2' \sim q_2$ all independent and $\sigma_k > 0$.

Validation

Validate via the hypothesis test [4] and expected score [1].

Use of Large Language Models

Models used : ChatGPT and Gemini, to

- find relevant published results,
- provide sentence-level writing assistance.

References

[1] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477) :359–378, 2007.

[2] Tilmann Gneiting and Johannes Resin. Regression diagnostics meets forecast evaluation : conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17(2), January 2023.

[3] Sebastian G. Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond, 2024.

[4] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests beyond classification, 2022.