Beyond (?) the pinball loss: Quantile Methods for Calibrated Uncertainty Quantification

Ignacio Hounie
ihounie@seas.upenn.edu

March 1, 2022

Outline

Definitions

Methods
- A fancy loss
- A Proper Scoring Rule
- KDE

Experiments

Results

Concluding Thoughts

References

Setting

- Supervised learning

  $\Rightarrow$ Let $\mathbf{X}, \mathbf{Y} \sim \mathbb{F}_{X,Y}$ denote random variables over $\mathcal{X} \times \mathcal{Y}$, $\mathcal{Y}$ an interval in $\mathbb{R}$ (i.e regression setting)
- We assume there exists a true conditional distribution $\mathbb{F}_{\mathbf{Y}|x}$ over $\mathcal{Y}$
- $Q_p(x)$ denotes the true p-th conditional quantile of this distribution i.e. $\mathbb{F}_{\mathbf{Y}|x}\left(Q_p(x)\right) = p$
- Conditional quantile estimator $\hat{Q}_p : \mathcal{X} \times (0, 1) \rightarrow \mathcal{Y}$

Pinball Loss.

- The p-th quantile minimizes the *pinball loss* $\ell_p : \mathcal{Y} \times \mathcal{Y} \to R$. Given a target $Q_p(x) = y$ and a prediction $\hat{Q}_p(x) = \hat{y}$:

Pinball Loss.

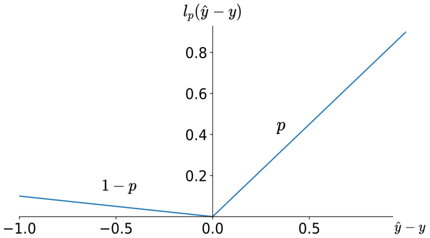
- The p-th quantile minimizes the *pinball loss* $\ell_p : \mathcal{Y} \times \mathcal{Y} \to R$.
  Given a target $Q_p(x) = y$ and a prediction $\hat{Q}_p(x) = \hat{y}$:

$$\ell_p(y, \hat{y}) = (\hat{y} - y)(\mathbb{I}\{y \leq \hat{y}\} - p) = \begin{cases} (1-p)(\hat{y} - y) & y < \hat{y} \\ -p(\hat{y} - y) & y \geq \hat{y} \end{cases}$$

Pinball Loss.

- The p-th quantile minimizes the *pinball loss* $\ell_p : \mathcal{Y} \times \mathcal{Y} \to R$.
  Given a target $Q_p(x) = y$ and a prediction $\hat{Q}_p(x) = \hat{y}$:

$$\ell_p(y, \hat{y}) = (\hat{y} - y)(\mathbb{I}\{y \leq \hat{y}\} - p) = \begin{cases} (1-p)(\hat{y} - y) & y < \hat{y} \\ -p(\hat{y} - y) & y \geq \hat{y} \end{cases}$$



Pinball Loss.

- The p-th quantile minimizes the *pinball loss* $\ell_p : \mathcal{Y} \times \mathcal{Y} \to R$.
  Given a target $Q_p(x) = y$ and a prediction $\hat{Q}_p(x) = \hat{y}$:
  $$\ell_p(y, \hat{y}) = (\hat{y} - y)(\mathbb{I}\{y \leq \hat{y}\} - p) = \begin{cases} (1-p)(\hat{y} - y) & y < \hat{y} \\ -p(\hat{y} - y) & y \geq \hat{y} \end{cases}$$
- Let $R(\hat{y}) = \mathbb{E}_{y \sim \mathbb{F}_{Y|x}} \ell_p(y, \hat{y})$ denote the statistical risk
- Assuming $R(\hat{y})$ is differentiable
  $$\frac{\partial R(\hat{y})}{\partial \hat{y}} = (1-p)\mathbb{F}_{Y|x}(\hat{y}) - p\left(1 - \mathbb{F}_{Y|x}(\hat{y})\right) = \mathbb{F}_{Y|x}(\hat{y}) - p$$
  $$\implies \left.\frac{\partial R(\hat{y})}{\partial \hat{y}}\right|_{\hat{y}=y} = 0$$

Pinball Loss.

- Stronger Results
  - The pinball loss is a Proper Scoring Rule for estimating quantile p.

Pinball Loss.

- Stronger Results
  - The pinball loss is a Proper Scoring Rule for estimating quantile p.
- A quantile prediction loss $g_p(\hat{y}, y)$ is a (strictly) proper scoring rule for quantile p iff the true quantile y (uniquely) minimizes g_p

Pinball Loss.

- Stronger Results
  - The pinball loss is a Proper Scoring Rule for estimating quantile p.
- A quantile prediction loss $g_p(\hat{y}, y)$ is a (strictly) proper scoring rule for quantile p iff the true quantile y (uniquely) minimizes g_p
- Proper scoring rules have tons of nice properties:
  - Form a non-negative convex cone
  - Admit an integral representation
  - Can define information measures and Bregman divergences under some conditions.

$\Rightarrow$ See [Buja et al., 2005] for a wonderful characterization of proper scoring rules for binary classification.

Proper Scoring rules and the Pinball Loss.

- Even Stronger Results:
- [Schervish et al., 2018] (Theorem 1): Any real valued quantile prediction loss g_p is a (strictly) proper scoring rule for quantile p iff there exists a (strictly) increasing function s such that:

$$g_p(y, \hat{y}) - g_p(y, y) = \begin{cases} p[s(y) - s(\hat{y})] & \text{if } y > \hat{y} \\ (1-p)[s(\hat{y}) - s(y)] & \text{if } \hat{y} < y \end{cases}$$

Proper Scoring rules and the Pinball Loss.

- Even Stronger Results:
- [Schervish et al., 2018] (Theorem 1): Any real valued quantile prediction loss g_p is a (strictly) proper scoring rule for quantile p iff there exists a (strictly) increasing function s such that:

$$g_p(y, \hat{y}) - g_p(y, y) = \begin{cases} p[s(y) - s(\hat{y})] & \text{if } y > \hat{y} \\ (1-p)[s(\hat{y}) - s(y)] & \text{if } \hat{y} < y \end{cases}$$

- Which can be re-written as:

$$g_p(y, \hat{y}) = g_p(y, y) + \ell_p(s(y), s(\hat{y}))$$

Proper Scoring rules and the Pinball Loss.

- Even Stronger Results:
- [Schervish et al., 2018] (Theorem 1): Any real valued quantile prediction loss g_p is a (strictly) proper scoring rule for quantile p iff there exists a (strictly) increasing function s such that:

$$g_p(y,\hat{y}) - g_p(y,y) = \begin{cases} p[s(y) - s(\hat{y})] & \text{if } y > \hat{y} \\ (1-p)[s(\hat{y}) - s(y)] & \text{if } \hat{y} < y \end{cases}$$

- Which can be re-written as:

$$g_p(y,\hat{y}) = g_p(y,y) + \ell_p(s(y), s(\hat{y}))$$

$\Rightarrow$ since $g_p(y,y)$ does not depend on $\hat{y}$ we end up minimising the composition of the pinball loss with an increasing function.

Proper Scoring rules and the Pinball Loss.

- Even Stronger Results:
- [Schervish et al., 2018] (Theorem 1): Any real valued quantile prediction loss g_p is a (strictly) proper scoring rule for quantile p iff there exists a (strictly) increasing function s such that:
- Which can be re-written as:

$$g_p(y, \hat{y}) = g_p(y, y) + \ell_p(s(y), s(\hat{y}))$$

$\Rightarrow$ since $g_p(y, y)$ does not depend on $\hat{y}$ we end up minimising the composition of the pinball loss with an increasing function.

$\Rightarrow$ In fact all scoring rules of this form are also (strictly) proper scoring rules for probability prediction of binary variables [Buja et al., 2005].

Is minimising the pinball loss enough?

- All proper scoring rules minimise both Calibration and Sharpness because, by definition they are minimised under the true distribution.

  $\Rightarrow$ However, this balance/trade-off (between penalising Calibration and Sharpness) is fixed, which according to [Chung et al., 2020] becomes relevant when minimising it empirically.

- All proper scoring rules minimise both Calibration and Sharpness because, by definition they are minimised under the true distribution.

  $\Rightarrow$ However, this balance/trade-off (between penalising Calibration and Sharpness) is fixed, which according to [Chung et al., 2020] becomes relevant when minimising it empirically.

- We can re-write the pinball loss as:

$$\ell_p(y, \hat{y}) = p\hat{y} + (y - \hat{y})\mathbb{I}_{y \leq \hat{y}} - py$$

- $p\hat{y}$ penalizes larger quantile predictions, i.e. sharpness
- $(y - \hat{y})\mathbb{I}_{y \leq \hat{y}}$ penalizes calibration
- py does not depend on predictions.

[Chung et al., 2020] propose a “tunable” loss function.

Probability Forecaster.

- Same Notions of calibration as in probability forecasts.
- $\mathcal{F}(\mathcal{Y})$ that maps an input $x \in \mathcal{X}$ to a continuous CDF $h(y)$ over $\mathcal{Y}$.
- *Perfect Probability Forecast* outputs the true conditional CDF $h^*(x) = \mathbb{F}_{Y|x}$
- Conditional Quantile regression "at all quantiles" (for all p) is equivalent to Inverse CDF estimation

$\Rightarrow$ We will refer to the family of quantile estimates as

$$\hat{Q} : \mathcal{X} \times (0,1) \to \mathcal{Y} = \{\hat{Q}_p(x),\ p \in\ [0,1]\}$$

$\Rightarrow$ This is what all the experiments in [Chung et al., 2020] actually estimate/model

(Individual) Calibration

- What we defined in class:
  - Classifier $\hat{y} : \mathcal{X} \to \mathcal{Y}$
  - Confidence predictor $\hat{p} : \mathcal{X} \to [0, 1]$

$$P(\hat{y}(x) = y \mid \hat{p}(x) = c) = c$$

- Quantile Regression

$$\begin{aligned} \text{for all } p \in \ (0,1),\ x \in \mathcal{X} \\ \hat{Q}_p(x) = Q_p(x) \\ \Leftrightarrow \mathbb{F}_{\mathbf{Y}|x}\left(\hat{Q}_p(x)\right) = p \end{aligned}$$

  - This is equivalent to marginal coverage.

$\Rightarrow$ For $\hat{Q}$ to be calibrated, this has to hold for all p

$\Rightarrow$ "Individual Fairness" [Kearns et al., 2019]

Individual Calibration (at all quantiles) is impossible

- Impossibility results cited
  - [Vovk, 2012]: (we saw it in class) Conditional Conformal Inference: at almost all nonatomic points of x, the prediction interval has infinite expected length.
  - [Vovk et al., 2005]: Probabilistic prediction (without assumptions on distributions) is impossible under Finite $\mathcal{X}$ and $\mathcal{Y}$, and Finite training set without repetition of x.
  - [Zhao et al., 2020] Individual calibration for probability forecasters is impossible (and unverifiable) for finite datasets (without assumptions on distributions).

Relaxing Calibration

- We define the observed probability of $\hat{Q}_p$ as:

$$p_{\mathrm{obs}}(p) := \mathbb{F}_{\mathbf{Y}|x}\left(\hat{Q}_p(x)\right)$$

Relaxing Calibration

- We define the observed probability of $\hat{Q}_p$ as:

$$p_{\mathrm{obs}}(p) := \mathbb{F}_{\mathbf{Y}|x}\left(\hat{Q}_p(x)\right)$$

- Mean Calibration: i.e. averaging over all x

$$\mathbb{E}_{x \sim \mathbb{F}_X}[p_{obs}(x)] = p$$

$\Rightarrow$ This is equivalent to marginal coverage.

Relaxing Calibration

- We define the observed probability of $\hat{Q}_p$ as:

$$p_{\mathrm{obs}}(p) := \mathbb{F}_{\mathbf{Y}|x}\left(\hat{Q}_p(x)\right)$$

- Mean Calibration: i.e. averaging over all x

$$\mathbb{E}_{x\sim\mathbb{F}_X}[p_{obs}(x)] = p$$

$\Rightarrow$ This is equivalent to marginal coverage.

- Group Calibration i.e. averaging over groups
  - Consider groups, i.e. measurable subsets $\mathcal{S}_i \subset \mathcal{X}$, $i = 0, \ldots, k$.
    We want mean calibration to hold when conditioning on each group:

$$\mathbb{E}_{x\sim\mathbb{F}_{X|X\in S_i}}[p_{obs}(x)] = p \quad i = 0, \ldots, k$$

Relaxing Calibration

- We define the observed probability of $\hat{Q}_p$ as:

$$p_{\mathrm{obs}}(p) := \mathbb{F}_{\mathbf{Y}|x}\left(\hat{Q}_p(x)\right)$$

- Mean Calibration: i.e. averaging over all x

$$\mathbb{E}_{x\sim\mathbb{F}_X}[p_{obs}(x)] = p$$

$\Rightarrow$ This is equivalent to marginal coverage.

- Group Calibration i.e. averaging over groups
  - Consider groups, i.e. measurable subsets $\mathcal{S}_i \subset \mathcal{X}$, $i = 0, \ldots, k$.
    We want mean calibration to hold when conditioning on each group:

$$\mathbb{E}_{x\sim\mathbb{F}_{X|X\in S_i}}[p_{obs}(x)] = p \quad i = 0, \ldots, k$$

- Adversarial Group Calibration "Average calibration within all subsets of the dataset with sufficiently many points" as a proxy for

$$\text{Group Calibration for all } \mathcal{S} \subset \mathcal{X} \text{ s.t. } P_{x\sim\mathbb{F}_X}(x \in \mathcal{S}) > 0$$

Relaxing Calibration

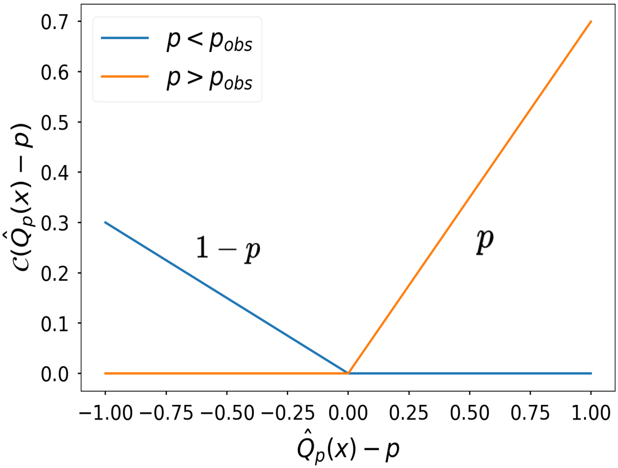
- From mean to adversarial Groups: Increasingly stringent (in x) definitions of calibration, may be suitable for different applications.
- There are also PAC notions of calibration in QR [Zhao et al., 2020].
- Other definitions in other contexts. E.g.: [Kearns et al., 2019] average over different tasks.

A Tunable Loss function

- Calibration objective:

$$\mathcal{C}\left(\hat{Q}_p\right) = \mathbb{I}\left\{p_{obs} < p\right\} * \mathbb{E}\left[Y - \hat{Q}_p \mid Y > \hat{Q}_p\right] * (1 - p_{obs}) \\ + \mathbb{I}\left\{p_{obs} > p\right\} * \mathbb{E}\left[\hat{Q}_p - Y \mid \hat{Q}_p > Y\right] * p_{obs}$$

A Tunable Loss function



A Tunable Loss function

- Calibration objective:

- It is minimised by the true quantiles.
- It is not decomposable in individual samples.
- It is not a proper scoring function

A Tunable Loss function

- Sharpness objective
- Predict quantiles at p and $1 - p$

$$\mathcal{P}\left(\hat{Q}_p, p\right) = \mathbb{E}\left[\left|\hat{Q}_p - \hat{Q}_{1-p}\right|\right]$$

A Tunable Loss function

- Sharpness objective
- Predict quantiles at p and $1 - p$

$$\mathcal{P}\left(\hat{Q}_p, p\right) = \mathbb{E}\left[\left|\hat{Q}_p - \hat{Q}_{1-p}\right|\right]$$

- You should only penalise

$$\mathbb{E}\left[\left|\hat{Q}_p - \hat{Q}_{1-p}\right|\right] \neq \mathbb{E}\left[\left|Q_p - Q_{1-p}\right|\right] = 1 - 2p$$

A Tunable Loss function

- Tunable loss:

$$\mathcal{L}\left(\hat{Q}_p, \hat{Q}_{1-p}\right) = (1-\lambda)\left[\mathcal{C}\left(\hat{Q}_p\right) + \mathcal{C}\left(\hat{Q}_{1-p}\right)\right] + \lambda\mathcal{P}\left(\hat{Q}_p, \hat{Q}_{1-p}\right)$$

A Tunable Loss function

- Tunable loss:

$$\mathcal{L}\left(\hat{Q}_p, \hat{Q}_{1-p}\right) = (1-\lambda)\left[\mathcal{C}\left(\hat{Q}_p\right) + \mathcal{C}\left(\hat{Q}_{1-p}\right)\right] + \lambda\mathcal{P}\left(\hat{Q}_p, \hat{Q}_{1-p}\right)$$

- λ is set by doing cross validation

- [Chung et al., 2020] train a model that ouputs all quantiles by optimizing

$$\mathbb{E}_{p\sim\mathrm{Unif}(0,1)}\mathcal{L}\left(\hat{Q}_p, \hat{Q}_{1-p}\right)$$

A Tunable Loss function

- This loss not a proper scoring rule
- It is not decomposable in individual samples
- It is not minimised under the true quantiles.

A proper scoring rule

- The interval (Winkler) Score [Winkler, 1972]:

$$S_\alpha(\hat{l}, \hat{u}; y) = (\hat{u} - \hat{l}) + \frac{2}{\alpha}\left[(\hat{l} - y)\mathbb{I}\{y < \hat{l}\} + (y - \hat{u})\mathbb{I}\{y > \hat{u}\}\right]$$

A proper scoring rule

- The interval (Winkler) Score [Winkler, 1972]:

$$S_\alpha(\hat{l}, \hat{u}; y) = (\hat{u} - \hat{l}) + \frac{2}{\alpha}\left[(\hat{l} - y)\mathbb{I}\{y < \hat{l}\} + (y - \hat{u})\mathbb{I}\{y > \hat{u}\}\right]$$

$$= \frac{2}{\alpha}\left[(\tilde{u} - y)\left(\mathbb{I}\{y \leq \hat{u}\} - \left(1 - \frac{\alpha}{2}\right)\right) + (\tilde{l} - y)\left(\mathbb{I}\{y \leq \hat{y}\} - \frac{\alpha}{2}\right)\right]$$

$\Rightarrow$ It is the scaled sum of pinball losses:

$$S_\alpha(\hat{l}, \hat{u}; y) = \frac{2}{\alpha}\left[\ell_{1-\alpha}(y, \tilde{u}) + \ell_\alpha(y, \tilde{l})\right]$$

A proper scoring rule

- The interval (Winkler) Score [Winkler, 1972]:

$$S_\alpha(\hat{l}, \hat{u}; y) = (\hat{u} - \hat{l}) + \frac{2}{\alpha}\left[(\hat{l} - y)\mathbb{I}\{y < \hat{l}\} + (y - \hat{u})\mathbb{I}\{y > \hat{u}\}\right]$$

$$= \frac{2}{\alpha}\left[(\tilde{u} - y)\left(\mathbb{I}\{y \leq \hat{u}\} - \left(1 - \frac{\alpha}{2}\right)\right) + (\tilde{l} - y)\left(\mathbb{I}\{y \leq \hat{y}\} - \frac{\alpha}{2}\right)\right]$$

$\Rightarrow$ It is the scaled sum of pinball losses:

$$S_\alpha(\hat{l}, \hat{u}; y) = \frac{2}{\alpha}\left[\ell_{1-\alpha}(y, \tilde{u}) + \ell_\alpha(y, \tilde{l})\right]$$

$\Rightarrow$ "bring to light a proper scoring rule that has largely been neglected for the purpose of learning quantiles. While some previous works utilize the interval score to evaluate interval predictions [some citations], to the best of our knowledge, no previous work has focused on simultaneously optimizing it and shown a thorough experimental evaluation as we provide"

A proper scoring rule

- The interval (Winkler) Score [Winkler, 1972]:

$$S_\alpha(\hat{l}, \hat{u}; y) = (\hat{u} - \hat{l}) + \frac{2}{\alpha}\left[(\hat{l} - y)\mathbb{I}\{y < \hat{l}\} + (y - \hat{u})\mathbb{I}\{y > \hat{u}\}\right]$$

$$= \frac{2}{\alpha}\left[(\tilde{u} - y)\left(\mathbb{I}\{y \leq \hat{u}\} - \left(1 - \frac{\alpha}{2}\right)\right) + (\tilde{l} - y)\left(\mathbb{I}\{y \leq \hat{y}\} - \frac{\alpha}{2}\right)\right]$$

$\Rightarrow$ It is the scaled sum of pinball losses:

$$S_\alpha(\hat{l}, \hat{u}; y) = \frac{2}{\alpha}\left[\ell_{1-\alpha}(y, \tilde{u}) + \ell_\alpha(y, \tilde{l})\right]$$

$\Rightarrow$ If training over all quantiles the α-th pinball loss gets scaled by:

$$\frac{2}{\alpha} + \frac{2}{1 - \alpha}$$

Model Agnostic QR: KDE + Regression model

- Conditional Density estimation
  - Assumes smoothness i.e. $x_j \approx x_k$ then $\mathbb{F}_{\mathbf{Y}|x_j} \approx \mathbb{F}_{\mathbf{Y}|x_k}$
- Under assumptions on the bandwidth the kernel density estimation Converges Uniformly to CDF [Stute, 1986]
- MAQR:
  - Utilize conditional density estimators to collect a dataset of quantile estimates
  - Fit a regressor on those quantiles (to get their inverses)

- Expected Calibration Error:
  - Quantile predictor $\hat{Q}_p$
  - For N samples the empirical observed probability is:

$$\hat{p}_{obs}(p) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left\{ y_i \leq \hat{Q}_p\left(x_i\right) \right\}$$

$$ECE(\hat{Q}_p) = \left| p_{obs} - p \right|$$

  - Family of Quantile predictors $\hat{Q} = \hat{Q}_p, \; p \in [0, 1]$
  - [Chung et al., 2020] Average over m quantiles

$$\mathrm{ECE}(Q) = \frac{1}{m} \sum_{j=1}^{m} \left| \hat{p}_{obs}\left(p_j\right) - p_j \right|, \text{ where } p_j \sim \mathrm{Unif}(0, 1)$$

Setup

- Models trained for all quantiles sampling p uniformly
  - Cali: Using their loss
  - SQR [Tagasovska and Lopez-Paz, 2019]: Pinball Loss
  - Interval Score
  - MAQR: KDE + Regressor
  - mPAIC [Zhao et al., 2020]: Randomised Quantile predictors trained on NLL+ECE

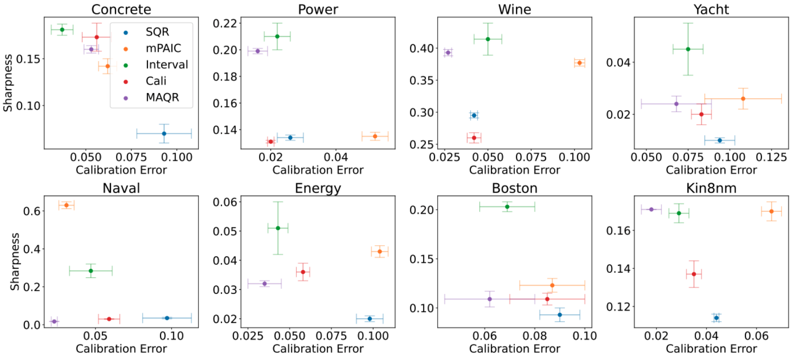
(Toy) Datasets

- UCI:
  - 8 regression datasets
  - Tiny and Low dimensional: dimension ≤ 20
- Nuclear fission: windowed time series. (iid assumption does not hold)
  - 16 regression tasks/outputs
  - Better but still low input dimension: 468

(Toy) Models

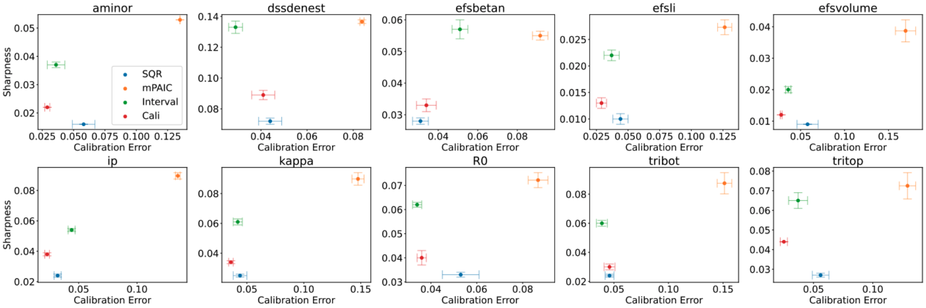
- UCI:
  - 2 layer Neural Network with 64 hidden neurons
- Nuclear fission
  - Deep learning: 3 hidden layers w/100 hidden neurons each

UCI


Concrete
Power
Wine
Yacht
Naval
Energy
Boston
Kin8nm
SQR
mPAIC
Interval
Cali
MAQR
Sharpness
Calibration Error

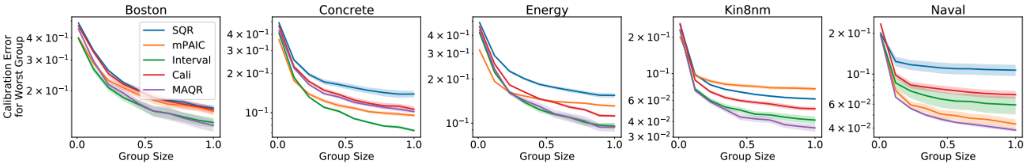
	SQR	*mPAIC*	***Interval***	***Cali***	***MAQR***
concrete	2.038 ± 0.225	1.157 ± 0.069	0.943 ± 0.053	1.465 ± 0.086	$\mathbf{0.672 \pm 0.118}$
power	0.834 ± 0.022	0.917 ± 0.021	0.620 ± 0.010	0.699 ± 0.019	$\mathbf{0.592 \pm 0.009}$
wine	3.242 ± 0.166	3.168 ± 0.019	2.197 ± 0.045	2.498 ± 0.135	$\mathbf{2.052 \pm 0.052}$
yacht	0.314 ± 0.061	0.197 ± 0.036	0.190 ± 0.021	0.298 ± 0.063	$\mathbf{0.086 \pm 0.016}$
naval	0.097 ± 0.011	3.112 ± 0.053	0.620 ± 0.114	1.560 ± 0.268	$\mathbf{0.044 \pm 0.001}$
energy	0.290 ± 0.016	0.223 ± 0.017	0.182 ± 0.026	0.204 ± 0.018	$\mathbf{0.101 \pm 0.006}$
boston	1.833 ± 0.299	1.395 ± 0.176	1.010 ± 0.118	1.449 ± 0.259	$\mathbf{0.864 \pm 0.287}$
kin8nm	1.241 ± 0.041	1.347 ± 0.031	0.776 ± 0.017	1.121 ± 0.072	$\mathbf{0.691 \pm 0.015}$

Figure 10: **UCI Interval Score** Full interval score results of UCI experiments from Section 4.1. Mean score across 5 trials is given, along with ± 1 standard error. The best mean has been bolded. *MAQR* tends to achieve the best interval score, which is surprising given that *Interval* utilizes the same model class to optimize the interval score directly.



aminor
dssdenest
efsbetan
efsli
efsvolume
ip
kappa
R0
tribot
tritop
Sharpness
Calibration Error
SQR
mPAIC
Interval
Cali

Adversarial Calibration



Concluding Thoughts

- QR is a widely used Uncertainty Quantification method
- Can be used to obtain prediction regions.
- But unlike the conformal setting iid-ness is assumed.
- Estimating at all quantiles is just estimating the inverse conditional CDF.
- How does this relate to Takeuchi's optimal regions in the conformal setting?

Concluding Thoughts

- Not so happy about
  - Toy datasets and models are used for benchmarking

    $\Rightarrow$ Although the usual computer vision benchmarks are used in other papers. And some fairness and causal inference related datasets.
  - KDE for those problems seems to work better but won't work in high dimensional settings.
  - fitting all quantiles may not be traditional QR
  - Designing penalised losses that may work in practice but without much connection to theory

References I

Buja, A., Stuetzle, W., and Shen, Y. (2005).

Loss functions for binary class probability estimation and classification: Structure and applications.

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. (2020).

Beyond pinball loss: Quantile methods for calibrated uncertainty quantification.

CoRR, abs/2011.09588.

Kearns, M., Roth, A., and Sharifi-Malvajerdi, S. (2019).

Average individual fairness: Algorithms, generalization and experiments.

Schervish, M. J., Kadane, J. B., and Seidenfeld, T. (2018).

Characterization of proper and strictly proper scoring rules for quantiles.

References II

Stute, W. (1986).
On Almost Sure Convergence of Conditional Empirical Distribution Functions.
The Annals of Probability, 14(3):891 – 901.

Tagasovska, N. and Lopez-Paz, D. (2019).
Single-model uncertainties for deep learning.

Vovk, V. (2012).
Conditional validity of inductive conformal predictors.
In *Asian conference on machine learning*, pages 475–490. PMLR.

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

References III

Winkler, R. L. (1972).
A decision-theoretic approach to interval estimation.
Journal of the American Statistical Association, 67(337):187–191.

Zhao, S., Ma, T., and Ermon, S. (2020).
Individual calibration with randomized forecasting.