Università di Pisa

A Gentle Introduction to Conformal Prediction

Based on the paper by Angelopoulos & Bates

Daniele Borghesi
Nicolas Humberto Montes De Oca Ibañez

Department of Computer Science
Master's Degree in Data Science and Business Informatics
Statistics for Data Science (SDS)

July 16, 2025

Table of Contents

1. The Problem: Why We Need Uncertainty
2. The Core Recipe of Conformal Prediction
3. Experimental setup
4. Conformal procedures for classification
5. Classification results analysis
6. Conformal procedures for regression
7. Regression results analysis

From a Single Guess to a Set of Possibilities

- Modern Machine Learning models are powerful, but they usually give just a single 'best guess' prediction.
- But how much can we really trust that one answer? What if it's wrong? This is critical in high-risk fields like medical diagnostics.
- **The Idea**: Instead of one answer, what if the model provided a **set of plausible answers**?
- **The Goal**: We want this set to be statistically *valid*. This means it should be guaranteed to contain the true answer with a high probability we can choose, like 90%.
- **Here is how it works in practice**: Imagine a model classifying an image.

  Standard Model's Best Guess:

  → ‘"fox"‘

  Conformal Prediction's 90% Plausible Set:

  → ‘"fox", "coyote", "dog"‘

We trade a single, possibly wrong answer for a small, rigorous set of possibilities.

Score, Calibrate, Predict

1. **Define a Score**: First, define a score that measures how poorly the model performs. For classification, a common score is *1* minus the *softmax probability* assigned to the true class.

$$s_i = 1 - \hat{f}(X_i)_{Y_i}$$

2. **Calibrate**: Using a separate **calibration set**, we compute the scores for each data point. We then find a threshold $\hat{q}$ by calculating a specific quantile of these scores. This $\hat{q}$ represents our calibrated level of acceptable error.

$$\hat{q} = \text{The } \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \text{ quantile of scores } \{s_1, ..., s_n\}$$

3. **Form the Prediction Set**: For a new input X, we create a prediction set by including all possible answers y whose error score would be less than or equal to our threshold $\hat{q}$.

$$\mathcal{T}(X) = \{y : s(X, y) \leq \hat{q}\}$$

This set is guaranteed to contain the true answer with a probability of at least $1 - \alpha$.

Experimental Setup

Our setup evaluated conformal prediction for both classification and regression, primarily using the **IRIS dataset**. We consistently employed **Support Vector Machines (SVMs)** as base models, except for one quantile regression experiment.

Data splitting was: ≈**70% training**, ≈**10% calibration**, and ≈**20% test**.

Our comprehensive **evaluation and analysis** included:

- **Coverage Analysis**: Repeated data splitting to plot coverage distribution (e.g., histogram centered on 90% target).
- **Adaptiveness Checks**: Investigated validity across data "slices" (e.g., **FSC** and **SSC**).

Experiments followed two **execution modes**:

- **Coverage Analysis**: 100 repetitions, each with a unique random seed for data splitting.
- **Other Metrics (Set Size, FSC, SSC)**: Calculated using a single, fixed base seed for reproducibility.

Classification: a Smarter Way to Build Sets

- **The Problem**: The basic method can be rigid. We want sets that are small for easy inputs and larger for hard ones.
- **Intuition for a new Score**: We start adding the most likely answers from our model, one by one, until we've accumulated enough "confidence". The score for a calibration point (X, Y) is the total probability mass we need to accumulate to finally include the **true class** Y. A low score is good (the true class was one of the first).

$$s(X, Y) = \sum_{j=1}^{k} \hat{f}(X)_{\pi_j}, \quad \text{where } Y = \pi_k$$

- **The Final Prediction Set**: After calibrating to find our threshold $\hat{q}$, we apply the greedy logic to new data. We add the most likely classes one by one until their cumulative probability sum reaches our threshold $\hat{q}$.

$$\mathcal{T}(X) = \{\pi_1, ..., \pi_k\}, \quad \text{where } k = \inf\{k' : \sum_{j=1}^{k'} \hat{f}(X)_{\pi_j} \geq \hat{q}\}$$

Classification: Bayesian Uncertainty with Guarantees

- **The Idea**: We use a Bayesian model's rich uncertainty estimate (the posterior predictive density). By calibrating it with Conformal Prediction, we get rigorous guarantees. This method is Bayes-optimal, meaning it produces the smallest possible average set size.
- **The Score Function**: The score is defined as the **negative** of the posterior predictive density for the true class Y. A high score (less negative) means the model assigned a low density to the true class, indicating a poor prediction.

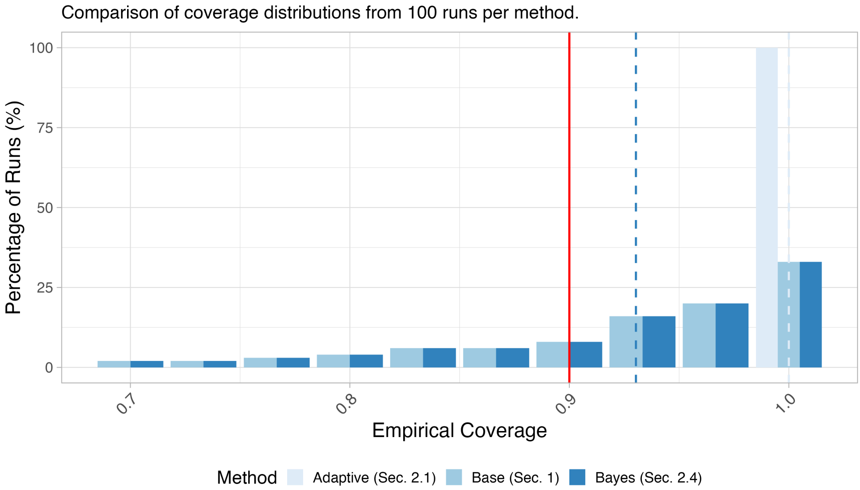
$$s(X, Y) = -\hat{f}(X)_Y$$

The negative sign ensures that a low density (a bad prediction) results in a high score, as required by the Conformal Prediction.

- **The Prediction Set**: After finding the calibrated threshold $\hat{q}$ (which will be a negative number), we form the prediction set. It includes all classes y whose posterior predictive density is greater than our new threshold $-\hat{q}$.

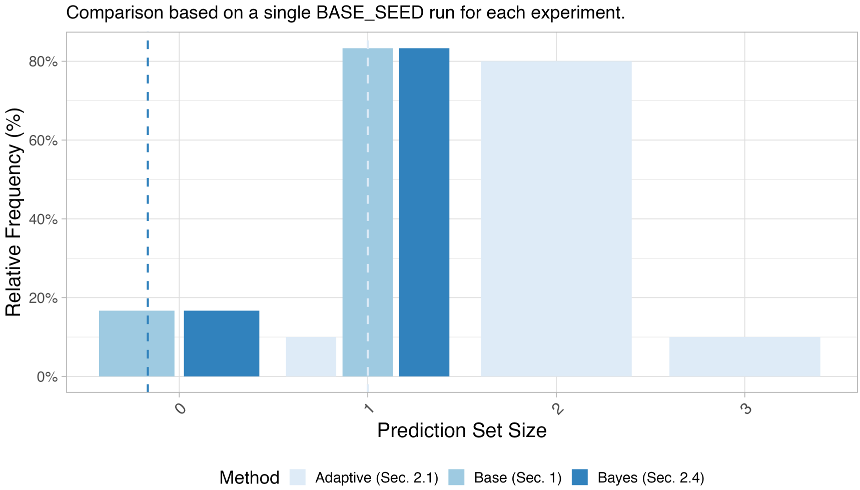
$$\mathcal{T}(X) = \{y : \hat{f}(X)_y > -\hat{q}\}$$

Classification: distribution of Coverage over 100 runs



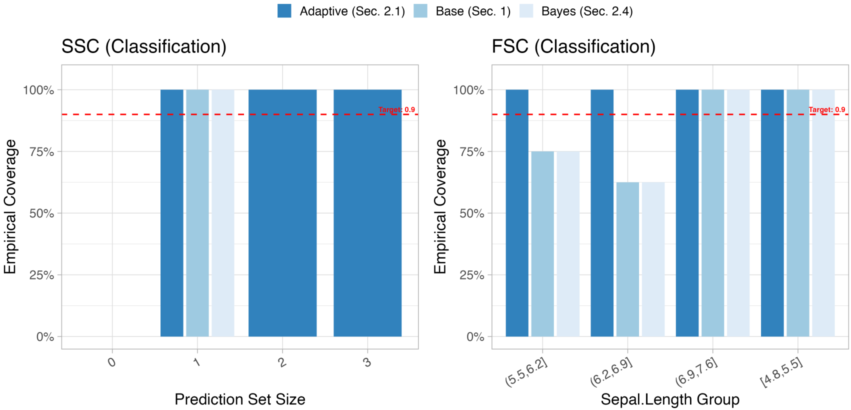
The Adaptive method consistently achieves higher and more stable coverage, while Basic and Bayes methods vary significantly.

Classification: Distribution of Prediction Set Sizes



Basic and Bayes methods produce smaller sets but can be empty; the Adaptive method avoids empty sets at the cost of larger sizes.

Classification: Coverage Stratified (SSC and FSC)



The Adaptive method maintains the target coverage across all feature groups, unlike Basic and Bayes which fail for some. Adaptive guarantees coverage for all set sizes, while Basic and Bayes fail for empty sets (size 0).

Classification: Key Results and Discussion

Metric	Basic	Adaptive	Bayes
Avg. Marginal Coverage	91.5%	**100%**	91.5%
Avg. Set Size	**0.83**	2.0	**0.83**
FSC (min. coverage)	62.5%	100%	62.5%
SSC (min. coverage)	0%	100%	0%

- **Performance of Basic vs. Bayes**: The Basic (Sec. 1) and Conformalizing Bayes (Sec. 2.4) methods show nearly identical performance.
  - They meet the average marginal coverage goal but are **unstable**.
  - Their conditional coverage fails, with FSC dropping as low as **62.5%**
- **Superiority of the Adaptive Method**: The Adaptive approach (Sec. 2.1) proves to be far more **robust**.
  - It achieves **100% coverage** in all tested scenarios
  - This reliability comes at the cost of a larger average prediction set size (2.0)
- **Conclusion**: For applications requiring **uncertainty guarantees**, the Adaptive method is the clear choice, despite its lower efficiency (larger sets).

Regression: Reliable Prediction Intervals

- **The Idea**: We shift from classification to regression. We start with a **Quantile Regression** model that predicts an initial interval (e.g., from the 5th to the 95th percentile). We then use Conformal Prediction to calibrate this interval and guarantee its coverage.
- **The Score Function**: The score measures how far the true value Y falls outside the predicted interval. If Y is inside the interval, the score is ≤ 0; otherwise, it's the distance from the nearest interval boundary.

$$s(X, Y) = \max\{\hat{t}_{\alpha/2}(X) - Y, \quad Y - \hat{t}_{1-\alpha/2}(X)\}$$

- **The Final Prediction Interval**: After finding the calibration term $\hat{q}$, we correct the initial interval by expanding it on both sides. This new interval is guaranteed to contain the true value with $1 - \alpha$ probability.

$$\mathcal{T}(X) = [\hat{t}_{\alpha/2}(X) - \hat{q}, \quad \hat{t}_{1-\alpha/2}(X) + \hat{q}]$$

Regression: a Simple Alternative for Intervals

- **The Idea**: Instead of predicting an interval directly, the model outputs a single point prediction ($\hat{f}(X)$) and a single value for its uncertainty ($u(X)$), such as the standard deviation. We then calibrate this uncertainty value.
- **Score Function**: The score is the absolute error normalized by the model's own uncertainty estimate. It measures the "relative error" and penalizes the model for being confidently wrong.

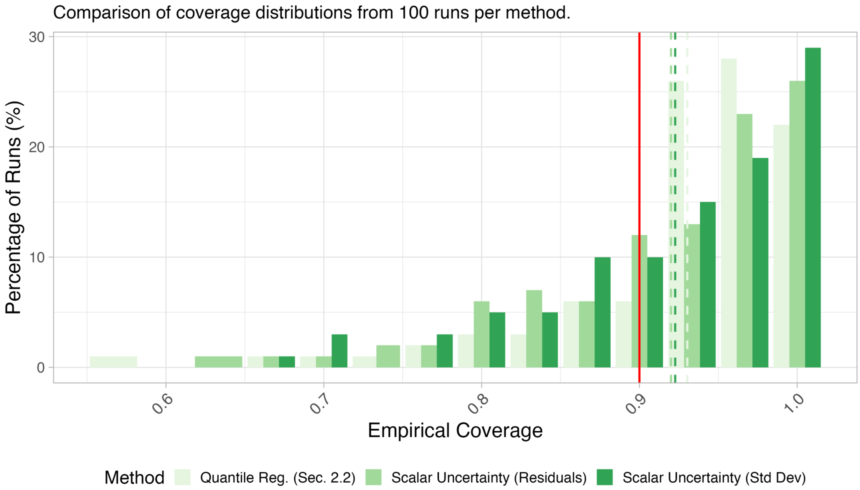
$$s(X, Y) = \frac{|Y - \hat{f}(X)|}{u(X)}$$

- **Final Prediction Interval**: After finding the calibration factor $\hat{q}$, we build a symmetric interval around our point prediction. The interval's radius is the model's uncertainty scaled by our calibrated factor $\hat{q}$.

$$\mathcal{T}(X) = [\hat{f}(X) - u(X)\hat{q},\ \hat{f}(X) + u(X)\hat{q}]$$

- **Generalization**: $u(X)$ is not only standard deviation. It can be any scalar uncertainty!

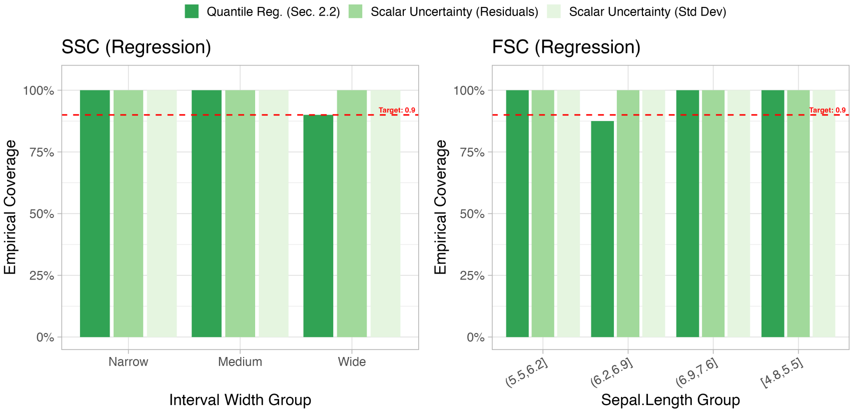
Regression: Coverage Distribution



Scalar and **Stddev** achieve a similar coverage (around 92%).

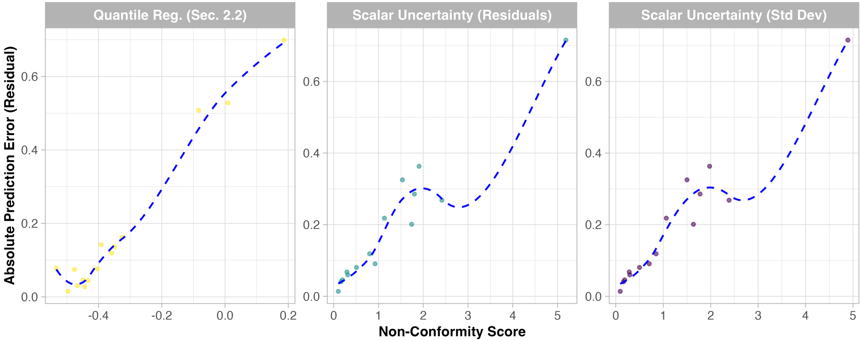
Quantile Reg. method is slightly better, with an average coverage of 93%.

Regression: Comparative FSC and SSC analysis



Quantile Reg. method falls below target in the group $(6.2, 6.9]$ and in the *Large* interval group. **Residuals** and **Stddev** methods maintain perfect 100% coverage across all groups.

Additivity Correlation (Non-conformity vs Error)



A positive correlation indicates good adaptiveness: larger errors correspond to higher scores.

- All methods show strong positive correlation between non-conformity scores and absolute prediction errors.
- For Quantile Regression, many non-conformity scores are zero (or very low), consistent with its definition for points inside the predicted interval.

Regression: Key Results and Discussion

Metric	**Quantile Reg.**	**Scalar Uncert.**	**Stddev Uncert.**
Avg. Marginal Coverage	93.0%	**92.0%**	**92.3%**
Avg. Interval Width	**1.40**	1.55	$1.12x10^{\mathbf{6}}$
FSC (min. Feature coverage)	96.67%	**100%**	**100%**
SSC (min. Size coverage)	96.67%	**100%**	**100%**

Comparative Discussion:

- All the methodologies achieve an average coverage over 90%. **Quantile Regression**, however, provides the best one average, with an average of 93%.
- In the single run, the minimum subgroup coverage (FSC, SSC) reveals potential weaknesses in the **Quantile Regression** method, suggesting lower reliability for some feature or size groups.
- The mean width of the intervals generated by one of the estimator-based methodologies (**standard deviation**) is anomalous. This highlights how the estimation of uncertainty through predictive models can vary greatly, even leading to totally unsuitable values.

Questions?

Do you have any questions?

- Montes de Oca Ibanez Humberto Nicolas (n.montesdeocaiban@studenti.unipi.it)
- Borghesi Daniele (d.borghesi@studenti.unipi.it)

THANKS!