New Advances in Conformal Prediction

Axel Benyamine

May 8, 2025

Introduction

In machine learning and statistical inference, the ability to quantify uncertainty is paramount, particularly in high-stakes applications such as medical diagnostics, financial risk assessment, and autonomous systems. Traditional methods for uncertainty quantification often rely on parametric assumptions or asymptotic approximations, which may fail in nonparametric settings or finite-sample regimes. Unlike most methods that rely on data distribution assumptions and algorithm characteristics, Conformal Prediction (CP) offers a powerful framework for constructing distribution-free and algorithm-free prediction intervals with finite-sample validity, requiring only data exchangeability [Lei et al., 2018]. By leveraging arbitrary predictive models to generate "conformity scores," CP provides calibrated uncertainty estimates without assumptions on the underlying data distribution.

A critical challenge arises when the test data deviates from the training distribution, a phenomenon known as distribution shift. While CP guarantees marginal coverage under exchangeability, its validity may degrade under covariate shift, where the input distribution changes but the conditional distribution of outputs remains consistent. Recent work has extended CP to handle such shifts via weighted conformal inference, which adjusts the calibration process using likelihood ratios between training and test distributions [Tibshirani et al., 2019]. However, accurately estimating these likelihood ratios remains challenging, and the impact of estimation errors on coverage guarantees has not been thoroughly quantified.

Similarly, outlier detection, a task inherently linked to identifying distributional anomalies, typically relies on hypothesis testing to flag observations that deviate significantly from a reference distribution. Conformal inference naturally lends itself to this task by furnishing p-values that measure the extremity of test points relative to calibration data [Bates et al., 2021]. Yet, existing methods do not address how these p-values should be constructed when the test distribution differs from the training distribution.

This work aims at bridging these gaps by developing new theoretical guarantees and practical methodologies for conformal prediction under distribution shift with applications to outlier detection. Our contributions are threefold:

- Coverage Error Bound for Likelihood Ratio Estimation: We derive a coverage lower bound showing precisely how likelihood ratio estimation errors impact weighted conformal prediction coverage.
- Weighted Conformal p-Values for Outlier Detection: We introduce new conformal p-values that maintain validity for outlier detection even when test and training distributions differ, with theoretical guarantees for both marginal and calibration-conditional approaches.
- Real-World Validation: We experiment the integrated framework on a financial dataset with temporal distribution shifts. By combining weighted conformal inference with adaptive p-value thresholds, we show improved outlier detection type I error, while also exposing the practical limitations of estimating the likelihood ratio.

Our results build on the interplay between conformal prediction and distribution shift, offering practitioners a flexible tool for reliable uncertainty quantification in nonstationary environments. The theoretical foundations provide new insights into how distribution shift adaptation enhances outlier detection, while our empirical findings highlight critical trade-offs between type I and type II errors in practical applications.

Contents

1 Foundations of Conformal Prediction

Given some data $(Z_i)_{1\leq i\leq n} = (X_i, Y_i)_{1\leq i\leq n}$, the goal of CP is to provide a function C that maps the new covariate X_{n+1} into an interval with coverage at least $1-\alpha$. In other words, we want

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1-\alpha$$

This objective is called finite-sample coverage as it provides non-asymptotic guarantees.

1.1 Data Exchangeability and Conformity Scores

At the core of conformal prediction lies the principle of **exchangeability** of the data: a condition that ensures the joint distribution of data remains invariant under permutations. Exchangeability underpins the validity of conformal prediction, enabling finite-sample coverage guarantees without assumptions on the underlying data distribution.

Definition 1. ***Exchangeability***
A sequence of random variables $Z_1, \ldots, Z_n$ is exchangeable if their joint distribution P satisfies

$$P(Z_1, \ldots, Z_n) = P(Z_{\sigma(1)}, \ldots, Z_{\sigma(n)})$$

for any permutation σ.

Exchangeability serves as a weaker generalization of the traditional independent identically distributed (i.i.d.) assumption. Indeed, while i.i.d. random variables are obviously exchangeable, the exchangeability allows for more freedom while maintaining the fact that no specific point has more "importance" than others. Exchangeability explicitly allows for dependencies within sequences, making it broadly applicable to real-world data where strict independence rarely holds.

When measuring how well new observations "conform" to historical data through a score, exchangeability ensures that each score is equally likely to occupy any rank within the set of scores. We call such a score a **conformity score** (conformity scores can also referred as "nonconformity scores"). These scores aim at quantifying the "atypicality" of observations relative to a reference dataset, where higher values indicate greater deviation from expected patterns.

The most direct way to create a conformity score is to consider the residuals of a trained model $\hat{\mu}$. For example, given $\hat{\mu}_n$ an estimator of $\mathbb{E}(Y|X)$ trained on the dataset $\mathcal{D}_{train} = (Z_i)_{i=1,\ldots,n}$, we can define a score function for each data point $Z = (X, Y)$ by $S : X, \mathcal{D}_{train} \to |Y - \hat{\mu}(X)|$, measuring how well (X, Y) could follow the same distribution than $\mathcal{D}_{train}$.

The exchangeability of the data then implies the exchangeability of the scores which will be used to provide quantiles for the new point score $S(Z_{n+1}, \mathcal{D}_{train})$.

Remark 1.1. Another type of exchangeability, now regarding the algorithm $\mathcal{A}$ that maps data points to the estimator $\hat{\mu}$, is also required in CP. In fact, the algorithm has to treat the data symmetrically to ensure the observation of the fitted estimator(s) benefit from the exchangeability of the training data. For simplicity, and because we aim to focus on the relaxation of the data exchangeability assumption, we will restrict our study to symmetric algorithms. However, algorithms may benefit from non-symmetry, for example by giving higher weights to the recent data in time-dependent settings.

1.2 Full Conformal Prediction

Full conformal prediction represents the original and most rigorous formulation of conformal prediction, providing a prediction interval for Y_{n+1} given the training data $Z_{1:n}$ and the new covariate X_{n+1}. Its core idea is to consider all possible values y for Y_{n+1} and test which ones would make the augmented sample $(Z_1, ..., Z_n, (X_{n+1}, y))$ appear exchangeable.

For a test input X_{n+1}, the method computes scores $V_{y,i} = S(Z_i, Z_{1:n} \cup (X_{n+1}, y))$ for all $i = 1, \ldots, n$, $V_{y,n+1} = S((X_{n+1}, y), Z_{1:n} \cup (X_{n+1}, y))$ and includes $y \in \mathbb{R}$ in the prediction interval if:

$$V_{y,n+1} \leq \text{Quantile}\left(1 - \alpha; \{V_{y,1}, \ldots, V_{y,n}\} \cup \{\infty\}\right),$$

where the α-level quantile is adjusted to account for finite-sample calibration.

This condition comes from the fact that, if $y = Y_{n+1}$, all the data $Z_{1:n+1}$ is exchangeable and the position of $V_{y,n+1}$ among all score values $(V_{y,i})_{i=1}^{n+1}$ will be uniformly distributed in $[n+1]$ the set of integers between 1 and $n+1$. It would then be in position inferior than $\lceil (1-\alpha)(n+1) \rceil$ with probability at least $1 - \alpha$.

With the finite-sample adjustment, the condition is equivalent to

$$V_{y,n+1} \leq \text{Quantile}\left(\frac{\lceil (1-\alpha)(n+1) \rceil}{n+1}; \{V_1, \ldots, V_n\} \cup \{\infty\}\right),$$

which leads to the following algorithm formulation.

Algorithm 1 Full Conformal Prediction (from [Lei et al., 2018])

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$, new point X_{n+1} at which to construct the prediction interval, and values $\mathcal{Y}_{\text{trial}} = \{y_1, y_2, \ldots\}$ to act as trial values
Output: Prediction interval $C_{\text{conf}}(X_{n+1})$
for $y \in \mathcal{Y}_{\text{trial}}$ **do**
  $V_{y,i} = S(Z_i; Z_{1:n} \cup (X_{n+1}, z))$, $i = 1, \ldots, n$, and $V_{y,n+1} = S((X_{n+1}, y); Z_{1:n} \cup (X_{n+1}, y))$ $\quad (*)$
  $\pi(y) = (1 + \sum_{i=1}^{n} \mathbb{1}\{V_{y,i} \leq V_{y,n+1})\}/(n+1)$
end for
Return $C_{\text{conf}}(X_{n+1}) = \{y \in \mathcal{Y}_{\text{trial}} : (n+1)\pi(y) \leq \lceil (1-\alpha)(n+1) \rceil\}$

Remark 1.2. Using the traditional residual score in **Algorithm 1**, line $(*)$ becomes

$$\begin{cases} \hat{\mu}_y = \mathcal{A}\big(\{(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)\}\big) \\ V_{y,i} = |Y_i - \hat{\mu}_y(X_i)|, i = 1, \ldots, n, \text{ and } V_{y,n+1} = |y - \hat{\mu}_y(X_{n+1})| \end{cases}$$

which emphasizes more explicitly that $|\mathcal{Y}_{\text{trial}}|$ different models have to be fitted, one for each y value we are considering.

1.3 Split Conformal Prediction

To mitigate computational costs, split conformal prediction partitions the data into a training set $\mathcal{D}_{train}$ and a calibration set $\mathcal{D}_{cal}$ (usually fixing $\mathcal{D}_{train} = \mathcal{D}_{cal} = \frac{n}{2}$). A model $\hat{\mu}$ is fitted on $\mathcal{D}_{train}$, and conformity scores $V_i = S(Z, \mathcal{D}_{train})$ are computed for each $Z \in \mathcal{D}_{cal}$. For a new input X_{n+1}, the prediction interval becomes:

$$C(X_{n+1}) = [\hat{\mu}(X_{n+1}) - \hat{q}_{1-\alpha}, \hat{\mu}(X_{n+1}) + \hat{q}_{1-\alpha}],$$

where $\hat{q}_{1-\alpha}$ is the $(1-\alpha)(1+1/n_{cal})$-th quantile of the calibration scores. This approach reduces computation from $O(n)$ trainings (required in "full" conformal methods) to a single model fit, enabling scalability to large datasets. However, it introduces a trade-off: while efficient, the split may reduce statistical power compared to full conformal approaches that utilize all data for both training and calibration.

Algorithm 2 Split Conformal Prediction (from [Lei et al., 2018])

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$
Output: Prediction band, over $x \in \mathbb{R}^d$
Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{I}_1$, $\mathcal{I}_2$
$V_i = S(Z_i; (Z_j)_{j \in \mathcal{I}_1})$, $i \in \mathcal{I}_2$
$d =$ the kth smallest value in $\{V_i : i \in \mathcal{I}_2\}$, where $k = \lceil (n/2+1)(1-\alpha) \rceil$
Return $C_{\text{split}}(x) = [\hat{\mu}(x) - d, \hat{\mu}(x) + d]$, for all $x \in \mathbb{R}^d$

We can see here that Split CP is just a special case of Full CP where the regression algorithm $\mathcal{A}$ is constant and returns a model $\hat{\mu}_n$ trained on $Z_{1:n}$. Therefore, the coverage proof of Split CP will immediately follow from the Full CP proof applied for $\frac{n}{2}$.

1.4 Coverage Guarantees

The hallmark of conformal prediction is its finite-sample marginal coverage guarantee:

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha,$$

which holds without distributional assumptions beyond exchangeability.

Theorem 1. ***Full CP provides finite-sample marginal coverage*** *(from [Lei et al., 2018])*

If $Z_1, \ldots, Z_n, Z_{n+1}$ are exchangeable, then the conformal interval $C(X_{n+1})$ computed in ***Algorithm 1*** *satisfies*

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$$

In addition, if there is almost surely no ties between the score values $V_{Y_{n+1},1}, \ldots, V_{Y_{n+1},n+1}$, then $C(X_{n+1})$ also satisfies

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \leq 1 - \alpha + \frac{1}{n+1}$$

The proof hinges on the rank uniformity of the test score $V_{Y_{n+1},n+1}$ among the augmented set $\{V_{Y_{n+1},1}, \ldots, V_{Y_{n+1},n+1}\}$. By construction, $V_{Y_{n+1},n+1}$ has equal probability to occupy any position in the sorted scores, ensuring the α-threshold excludes the worst α-proportion of cases.

Remark 1.3. For continuous distributions, the condition for the upper-bound is always valid, but discrete settings may require randomized tie-breaking to achieve tight bounds. Both guarantees remain valid even when the score function S is poorly calibrated, though interval width depends critically on the choice of S.

Remark 1.4. The marginal coverage $\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1-\alpha$ should not be confused with the conditional coverage $\mathbb{P}\left(Y_{n+1} \in C(x) | X_{n+1} = x\right) \geq 1-\alpha$ for all x. In fact, while the latter implies the former, the marginal coverage is only valid on average with respect to X_{n+1}'s distribution and not for every value X_{n+1} can take.

Proof. of Theorem 1

By definition of **Algorithm 1**, we have the event equality

$$\{Y_{n+1} \in C(X_{n+1})\} = \{V_{Y_{n+1},n+1} \leq \text{Quantile}\left(1-\alpha; V_{Y_{n+1},1:n} \cup \{\infty\}\right)\}$$

We now use $\forall v \in \mathbb{R}, E \subset \mathbb{R}, \beta \leq 1$, the following equivalence:

$$v \leq q := \text{Quantile}\left(\beta; E \cup \{\infty\}\right) \iff v \leq q' := \text{Quantile}\left(\beta; E \cup v\right)$$

While the indirect implication is obvious, the direct implications follows from the fact that replacing $+\infty$ by v might shift q towards the left but we would still have $q' \subset [v, q]$ as $v \leq q$ implies that $\frac{1}{|E|}\sum_{e \in E} \mathbb{1}(e < v) < \beta$.

Having now, $\{Y_{n+1} \in C(X_{n+1})\} = \{V_{Y_{n+1},n+1} \leq \text{Quantile}\left(1-\alpha; V_{Y_{n+1},1:n+1}\right)\}$, Y_{n+1} belongs to the interval if the rank of $V_{Y_{n+1},n+1}$ among all $V_{Y_{n+1},1:n+1}$ is lower than $\lceil (1-\alpha)(n+1) \rceil$ which, by exchangeability, happens with probability at least $\frac{\lceil (1-\alpha)(n+1) \rceil\}}{n+1} \geq 1-\alpha$.

If there is alsmost surely no ties, the probability is exactly equal to $\frac{\lceil (1-\alpha)(n+1) \rceil\}}{n+1} \leq 1-\alpha+\frac{1}{n+1}$. □

Remark 1.5. [Tibshirani et al., 2019] present an alternative proof of **Theorem** 1 by considering the event E_v that $\{V_{Y_{n+1},1}, \ldots, V_{Y_{n+1},n+1}\} = \{v_1, v_{n+1}\}$. Assuming there are almost surely no ties, it can be shown quite easily that, under E_v, V_{n+1} follows a uniform distribution on the set $\{v_1, v_{n+1}\}$, which can be rewritten as

$$V_{n+1} | E_v \sim \frac{1}{n+1}\sum_{i=1}^{n+1} \delta_{v_i}$$

Having, under E_v, $\{V_1, \ldots, V_{n+1}\} = \{v_1, v_{n+1}\}$, we then have

$$V_{n+1} | E_v \sim \frac{1}{n+1}\sum_{i=1}^{n+1} \delta_{V_i}$$

Using this, and the definition of quantiles, they show the coverage when conditioning on E_v and by marginalizing, obtain the general result.

We notice here that $\frac{1}{n+1}$ weights are "applied" to each score V_i. All these weights are equal because we are here under the exchangeability assumption but we will see in the following section that, in non-exchangeable settings, choosing specific weights for every score can bring back the coverage guarantees of the exchangeability settings.

Remark 1.6. In the rest of this essay, we will no longer use **Full CP** due to its higher computational price. Using **Split CP**, the scores empirical distribution becomes $\frac{1}{n+1}\sum_{i=1}^{n} \delta_{V_i} + \frac{1}{n+1}\delta_{\infty}$ where δ_{∞} accounts for the, non computed, test score. As shown in the *proof of Theorem 1*, this empirical distribution yields to the same prediction intervals, even in the context of **Full CP**.

1.5 Conformity Scores in Practice

Even thought the CP marginal coverage holds for any score function (that relies on a regression model trained symmetrically with respect to the training points), nothing guarantees us that the intervals' lengths will be short enough for the results to be exploitable.

Consequently, the choice of conformity score critically influences, through the length of the intervals, the adaptivity and efficiency of conformal prediction intervals. Below, we detail three widely used scores -**absolute residuals**, **rescaled residuals** and **conformalized quantile regression (CQR)**- explaining their computation, theoretical properties, and practical trade-offs.

1.5.1 Absolute Residuals

The residual score is the simplest conformity measure, defined as the absolute difference between observed and predicted values:

$$S((X,Y),\mathcal{D}_{train}) = |Y - \hat{\mu}(X)|,$$

where $\hat{\mu}$ is a point predictor (e.g., linear regression, neural network) trained on $\mathcal{D}_{\text{train}}$. This score evaluates raw prediction error, producing intervals of constant width: $\hat{\mu}(X_{n+1}) \pm \hat{q}_{1-\alpha}$. For implementation, $\hat{\mu}$ is fit once on $\mathcal{D}_{\text{train}}$, and residuals are computed on $\mathcal{D}_{\text{cal}}$. While model-agnostic and computationally efficient, constant-width intervals perform poorly under heteroskedasticity, as they over-cover in low-variance regions and under-cover in high-variance regions. However, residual scores excel in homoskedastic settings (e.g., simulated data with additive Gaussian noise), usually having very competitive interval lengths while being computationally efficient.

1.5.2 Rescaled Residuals

The rescaled residuals score, designed by [Lei et al., 2018], addresses heteroskedasticity by normalizing residuals with an estimate of their conditional mean absolute deviation (MAD). The score is defined as:

$$S((X,Y),\mathcal{D}_{train}) = \frac{|Y - \hat{\mu}(X)|}{\hat{\sigma}(X)},$$

where $\hat{\sigma}(X)$ aims at estimating the conditional MAD $\mathbb{E}[|Y - \hat{\mu}(X)| \; |X]$. In order to compute both $\hat{\mu}$ and $\hat{\sigma}$ models, one can first train $\hat{\mu}$ on $\mathcal{D}_{\text{train}}$ to estimate $\mathbb{E}[Y|X]$ and then train $\hat{\sigma}$ on $\mathcal{D}_{\text{train}}$ (or a separate split) to predict $|Y - \hat{\mu}(X)|$, making this method slightly more computationally expensive than the absolute residuals.

The main interest of the rescaled residuals score is that the prediction interval length adapts to local uncertainty: $C(X_{n+1}) = \hat{\mu}(X_{n+1}) \pm \hat{q}_{1-\alpha} \cdot \hat{\sigma}(X_{n+1})$. While rescaled residuals perform well in heteroskedastic regimes, they require an accurate scale estimation $\hat{\sigma}(X)$ to avoid biases interval widths and are very sensitive to $\hat{\sigma}$'s miscalibration.

1.5.3 Conformalized Quantile Regression (CQR)

CQR integrates quantile regression with conformal prediction to construct adaptive intervals without explicit variance modeling. The score is:

$$S((X,Y),\mathcal{D}_{train}) = \max\left\{\hat{q}_{\alpha/2}(X) - Y, Y - \hat{q}_{1-\alpha/2}(X)\right\},$$

where $\hat{q}_\tau(X)$ estimates the τ-th conditional quantile of $Y|X$.

The CQR score, designed by [Romano et al., 2019], quantifies how far Y deviates from the central prediction interval $[\hat{q}_{\alpha/2}(X), \hat{q}_{1-\alpha/2}(X)]$, where the bulk of the conditional distribution $Y|X$ is expected

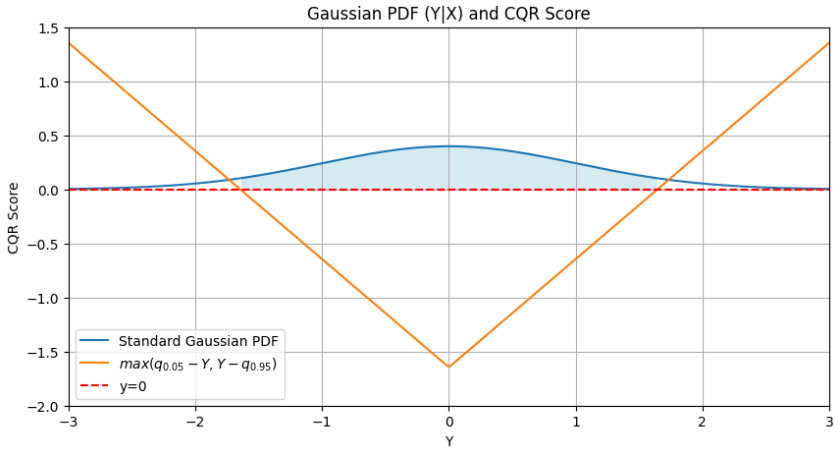


Figure 1: An example of CQR score when $Y|X \sim \mathcal{N}(0,1)$

to lie. The score is negative if Y falls inside this interval, with magnitude proportional to its distance from the nearest quantile boundary, and negative otherwise.

In order to compute the two quantile models (or a single model with two outputs), [Koenker & Bassett, 1978] showed that we can use the pinball loss:

$$l_\beta(\theta, y) = \begin{cases} \beta(y-\theta) & \text{if } y \geq \theta, \\ (1-\beta)(\theta - y) & \text{if } y < \theta. \end{cases}$$

which satisfies the property $\text{Quantile}(\beta, y_{1:n}) = \text{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} l_\beta(\theta, y_i)$.

Having $\hat{q}_{1-\alpha} = \text{Quantile}(1-\alpha, S((X_i, Y_i), \mathcal{D}_{train})_{1:n_{cal}})$, the CQR score outputs $C(X_{n+1}) = [\hat{q}_{\alpha/2}(X_{n+1}) - \hat{q}_{1-\alpha}, \hat{q}_{1-\alpha/2}(X_{n+1}) + \hat{q}_{1-\alpha}]$ eventually having $C(X_{n+1})$ narrower than $[\hat{q}_{\alpha/2}(X_{n+1}), \hat{q}_{1-\alpha/2}(X_{n+1})]$ if most scores are negatives.

By adapting to local uncertainty through the presence of $\hat{q}_{\alpha/2}(X_{n+1})$ and $\hat{q}_{1-\alpha/2}(X_{n+1})$ in the prediction interval, CQR performs very well in heteroskedastic settings (e.g., medical data with outcome variance depending on patient age) but demands accurate quantile estimation. However, the computational cost is usually higher than both residual-based methods due to the training of the quantiles models. While CQR scores are, as rescaled residuals were with the MAD, sensitive to the quantiles estimation, it is more robust than rescaled residuals and still manages to achieve the coverage guarantees with insufficient training data.

1.5.4 Comparison of Scores under Heteroskedastic and Homoskedastic designs

We generate datasets with $n = 1000$ and $n = 100$ data points (with $|\mathcal{D}_{train}| = |\mathcal{D}_{cal}| = \frac{n}{2}$) in one homoskedastic setting (having $Y \sim \sin(2\pi X) + \mathcal{N}(0, 0.5)$) and one heteroskedastic setting (having $Y \sim \sin(2\pi X) + \mathcal{N}(0, 0.05 + 0.45X^2)$). All the setups are tested on 200 points.

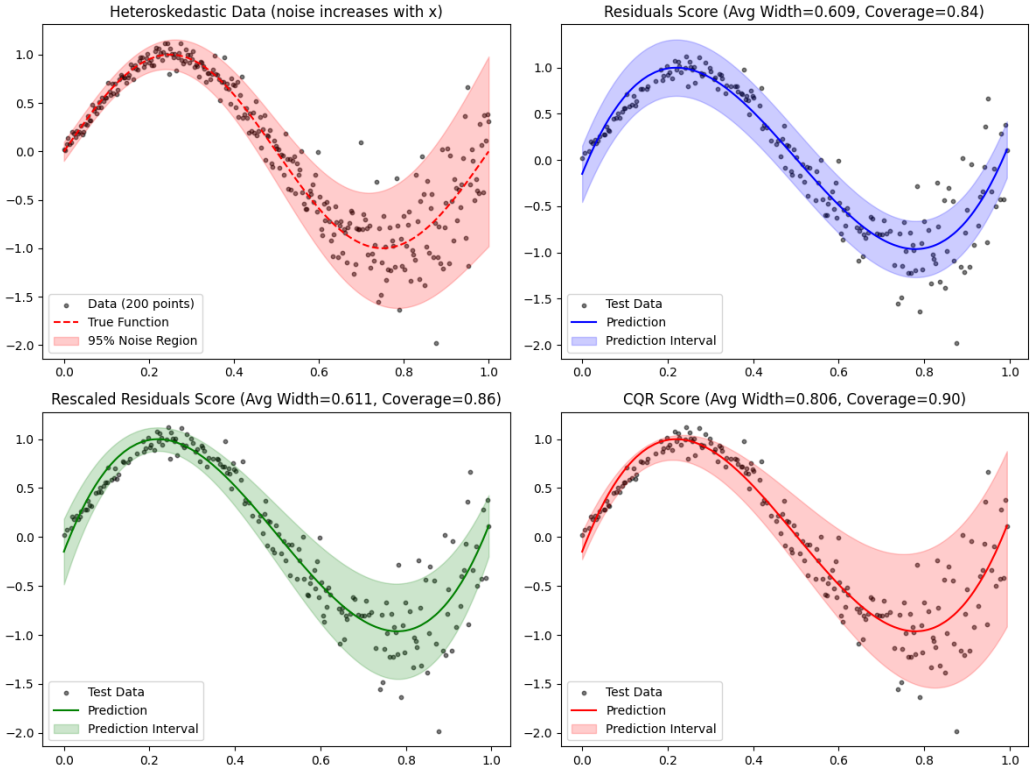
Homoskedastic Design

As shown in **figure 2a**, the absolute residuals seems to be the best method when having a limited number of training data, achieving the desired coverage with the shorter average width. This must be the consequence of lacking accuracy on the scale model (for rescaled residuals) and the quantiles models (for CQR). However, when augmenting the number of data points **figure 2b** both absolute residuals scores and CQR methods are very promising, with CQR having a slightly narrower interval.

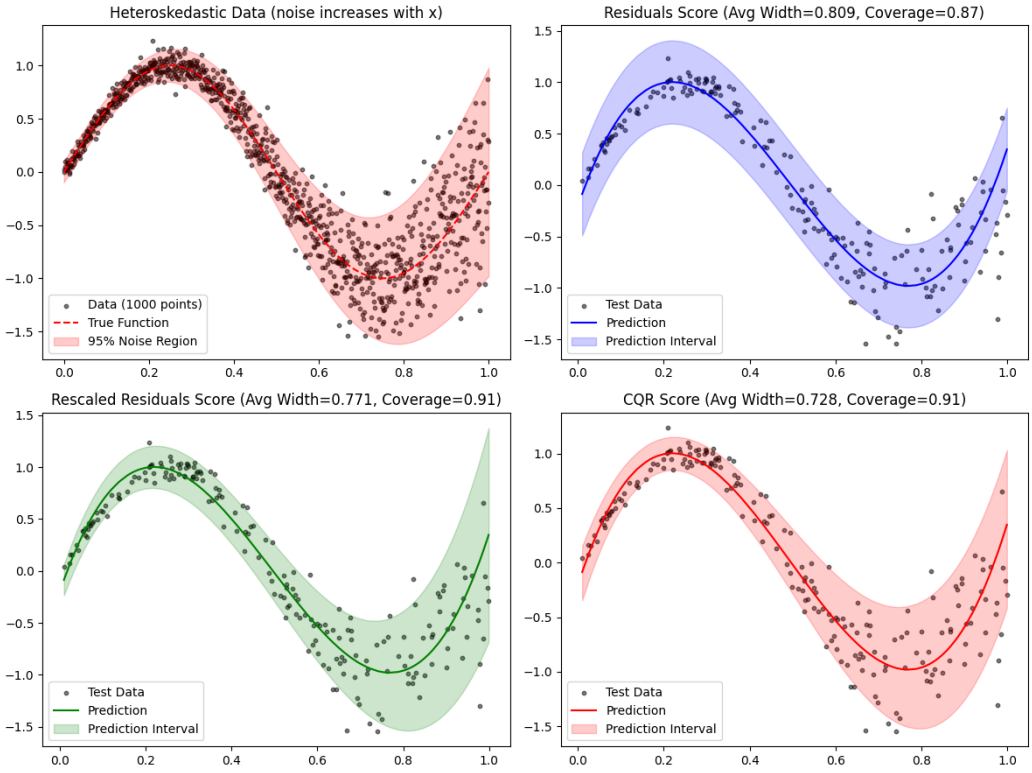
Heteroskedastic Design

In the context of heteroskedasticity, **figure 3** clearly demonstrate the importance of having a locally adapted interval length, thus not making intervals artificially wide in areas with lower noise. While both rescaled residuals and CQR methods provide good average widths, CQR performs better in both low data volume and high data volume setups: achieving, unlike rescaled residuals, 90% coverage with low data volume and providing narrower intervals in the high data volume setup.

Overall, these results highlight CQR's robust performance across data conditions, with rescaled residuals showing particular weakness in homoskedastic settings. The absolute residuals method, while simple, proves competitive in homoskedastic scenarios with sufficient data and outperforms CQR when having insufficient data.

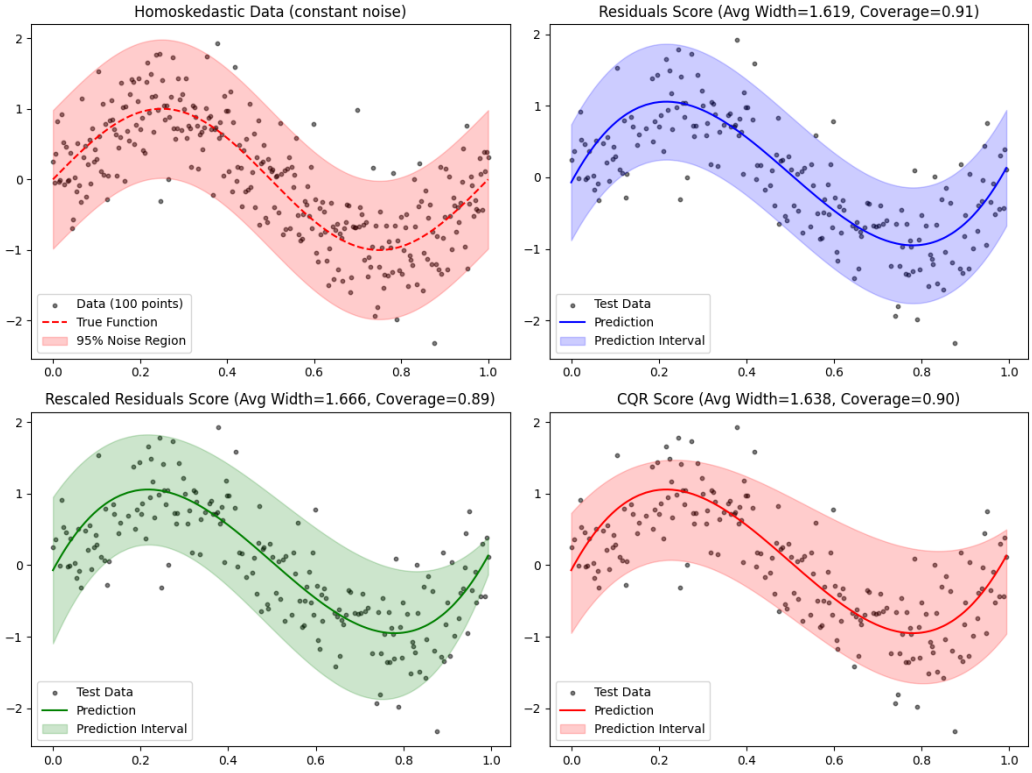


(a) With 100 training points

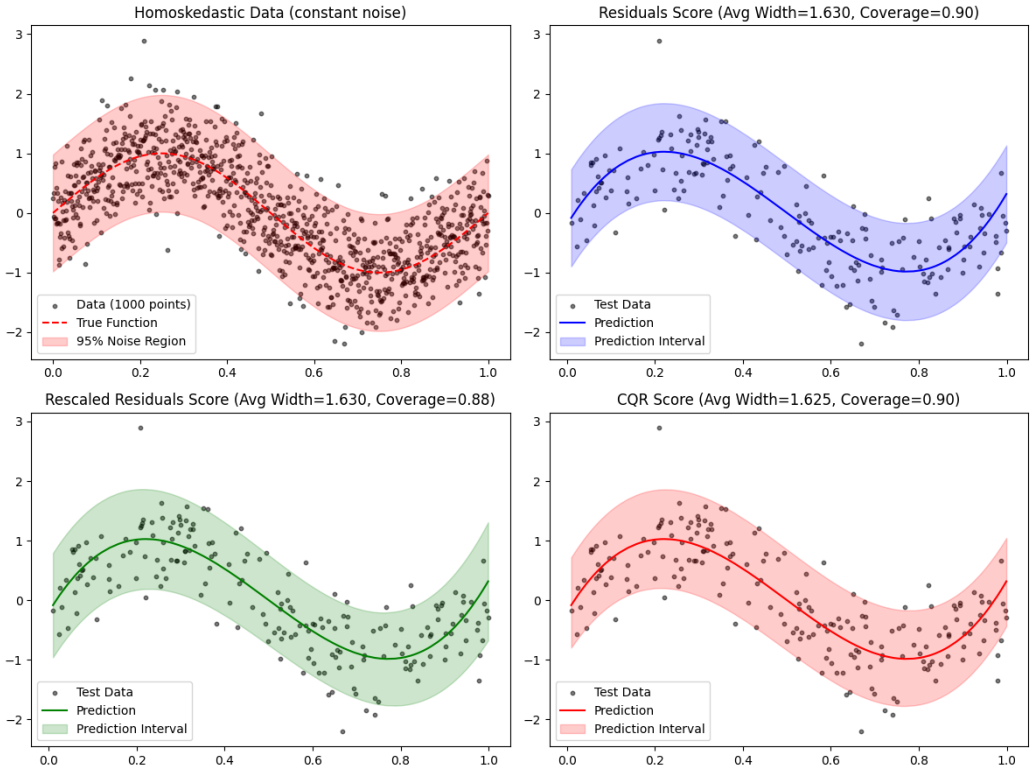


(b) With 1000 training points

Figure 2: Comparison of Score Functions in a Heteroskedastic design $Y \sim \sin(2\pi X) + \mathcal{N}(0, 0.05 + 0.45X^2)$



(a) With 100 training points



(b) With 1000 training points

Figure 3: Comparison of Score Functions in a Homoskedastic design $Y \sim \sin(2\pi X) + \mathcal{N}(0, 0.5)$

2 Conformal Prediction Under Distribution Shift

2.1 Covariate Shift and Exchangeability Breakdown

Conformal prediction hinges on the exchangeability of training and test data, a condition violated under distribution shift. In many real-world applications, the test covariate distribution P_X^{test} differs from the training distribution P_X^{train} while the conditional distribution $P_{Y|X}$ remains unchanged. This phenomenon, termed **covariate shift**, arises in settings such as medical diagnostics (where patient demographics evolve) or financial time series (subject to temporal drift). Standard conformal prediction intervals, which assume exchangeability, lose validity under such shifts, as their coverage guarantees degrade when test points are drawn from a distribution misaligned with the calibration data.

Formally, we have

$$\{(X_i, Y_i)\}_{i=1}^n \sim P^{\text{train}} = P_X^{\text{train}} \cdot P_{Y|X}$$

and

$$(X_{n+1}, Y_{n+1}) \sim P^{\text{test}} = P_X^{\text{test}} \cdot P_{Y|X}$$

with independence between all the points.

The goal remains the construction a prediction band $C(X_{n+1})$ satisfying:

$$P\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha,$$

despite $P_X^{\text{test}} \neq P_X^{\text{train}}$.

2.2 Weighted Conformal Prediction

To address covariate shift, [Tibshirani et al., 2019] introduced likelihood ratio weighting to recalibrate the calibration scores. Weighted CP addresses the non-uniform rank distribution of the test score V_{n+1} under distribution shift by reweighting it with the likelihood ratio between both distributions (as $P_{Y|X}$ remains unchanged, we only need to mimic the distribution $P_X^{\text{train}}(X_{n+1})$).

Let $w(x) = \frac{dP_X^{\text{test}}(x)}{dP_X^{\text{train}}(x)}$ denote the Radon-Nikodym derivative of the test and training covariate distributions. Then, the weighted conformal procedure adjusts the calibration step to account for w.

Therefore, the prediction interval becomes:

$$C(X_{n+1}) = [\hat{\mu}(X_{n+1}) - \hat{q}_{1-\alpha}^w, \hat{\mu}(X_{n+1}) + \hat{q}_{1-\alpha}^w]$$

where $\hat{q}_{1-\alpha}^w$ is the $(1-\alpha)$-quantile of the weighted empirical distribution of residuals

$$\sum_{i=1}^n p_i^w \delta_{V_i} + p_{n+1}^w \delta_\infty \tag{2.1}$$

with some weights p_i^w, proportional to w, that we will define formally in the next subsection.

This adjustment ensures that the quantile estimation prioritizes regions where P_X^{test} dominates.

Remark 2.1. The assumption that P_X^{test} is absolutely continuous with respect to P_X^{train} is necessary in order to use w. This usually doesn't pose any problem as most of the supports in real-world data would be equal between the training distribution and the testing one. In the following of this essay, all the training and testing supports will be equal (wether it is in continuous or discrete setups). To simplify, we assume in the following that this assumption is always verified.

Remark 2.2. One could consider the test distribution as the reference one and therefore weight all the calibration points to bring back exchangeability (with the inverse Radon-Nikodym derivative $\frac{dP_X^{\text{train}}(x)}{dP_X^{\text{test}}(x)}$). However, as we will see in the next subsection, this would lead to much more complicated weights p_i^w due to the fact that n points would have to be corrected instead of one.

2.3 Theoretical Guarantees

The validity of weighted conformal prediction rests on a generalized notion of **weighted exchangeability**.

Definition 2. ***Weighted Exchangeability***

A sequence of random variables $Z_1, \dots, Z_n$ is weighted exchangeable (with associated weight functions $w_1, \dots, w_n$) if the density of their joint distribution satisfies

$$f(z_1, \dots, z_n) = \prod_{i=1}^{n} w_i(z_i) \cdot g(z_1, \dots, z_n)$$

with g being any permutation invariant function (i.e. $g(z_1, \dots, z_n) = g(z_{\sigma(1)}, \dots, z_{\sigma(n)})$ for any permutation σ).

If the training and test distribution have densities f^{train} and f^{test} is easy to show that independent data points with a covariate shift for the test point X_{n+1} are weighted exchangeable. Indeed,

$$\begin{aligned} f^{\text{joint}}(z_1, \dots, z_{n+1}) &= f^{\text{test}}(z_{n+1}) \left(\prod_{i=1}^{n} f^{\text{train}}(z_i) \right) \quad \text{by independence} \\ &= w(z_{n+1}) \left(\prod_{i=1}^{n+1} f^{\text{train}}(z_i) \right) \quad \text{by definition of } w \end{aligned}$$

Explicitly, we get $\forall i \leq n, w_i = \tilde{1}$ and $w_{n+1} = w = \frac{dP_X^{\text{test}}(x)}{dP_X^{\text{train}}(x)}$.

Remark 2.3. In order to show to weighted exchangeability of $X_1, \dots, X_{n+1}$, we can no longer assume the training data to be only exchangeable (with the test point independent from all training points) and have to consider independent points.

We can now define the weights p_i^w used in **(2.1)** to compute the weighted quantile $q_{1-\alpha}$. In the general weighted exchangeability setting, these weights are defined as

$$p_i^w(z_1, \dots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}{\sum_{\sigma} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})} \quad \text{for all } i = 1, \dots, n+1$$

When the test point is the only shifted one, with the exchangeability weights derived above, we get for $i = 1, \dots, n+1$,

$$p_i^w = \frac{n! w(z_i)}{\sum_{j=1}^{n+1} \sum_{\sigma:\sigma(n+1)=j} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})} = \frac{n! w(z_i)}{\sum_{j=1}^{n+1} n! w(z_j)} = \frac{w(z_i)}{\sum_{j=1}^{n+1} w(z_j)}$$

In other words, if computing the prediction interval for $x \in \mathbb{R}$,

$$p_i^w = \frac{w(X_i)}{\sum_{j=1}^n w(X_j) + w(x)}, \quad \text{for } i = 1 \ldots, n+1$$

and

$$p_{n+1}^w = \frac{w(x)}{\sum_{j=1}^n w(X_j) + w(x)}$$

If the likelihood ratio w is known or accurately estimated, the procedure achieves marginal coverage. We show the following theorem, leveraging the fact that the weighted scores mimic exchangeability under P_X^{train}, in the context of **Full CP** (having **Split CP** as a special case as noted in **subsection 1.3**) where the prediction interval is now

$$C(X_{n+1}) = \{y \in \mathbb{R} : V_{y,n+1} \leq \text{Quantile}(1-\alpha; \sum_{i=1}^n p_i^w \delta_{V_i} + p_{n+1}^w \delta_\infty)\}$$

Remark 2.4. The quantile above refers to the $1-\alpha$ quantile of the distribution $\sum_{i=1}^n p_i^w \delta_{V_i} + p_{n+1}^w \delta_\infty$. Using the same notation for uniformly weighted quantiles, we would write

$$\text{Quantile}(1-\alpha; V_{1:n} \cup \infty) = \text{Quantile}(1-\alpha; \frac{1}{n+1}\sum_{i=1}^n \delta_{V_i} + \frac{1}{n+1}\delta_\infty)$$

Theorem 2. ***Weighted Conformal Coverage for Full CP*** *(from [Tibshirani et al., 2019])*

Under covariate shift, if $w(x) = \frac{dP_X^{test}}{dP_X^{train}}(x)$ is known, the weighted conformal prediction interval $C(X_{n+1})$ satisfies

$$P\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha.$$

In addition, if there is almost surely no ties between the score values $V_{Y_{n+1},1}, \ldots, V_{Y_{n+1},n+1}$, then $C(X_{n+1})$ also satisfies

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) \leq 1 - \alpha + \frac{1}{n+1}$$

Proof. We assume, for simplicity, that there is almost surely no ties between all score values. We consider, as in **remark 1.5**, the event E_z that $\{Z_1, \ldots, Z_{n+1}\} = \{z_1, \ldots, z_{n+1}\}$ and denote $v_i = S(z_i, z_{-i})$.
Then, for all $i = 1, \ldots, n+1$

$$\mathbb{P}(V_{n+1} = v_i | E_v) = \mathbb{P}(Z_{n+1} = z_i | E_v) = \frac{\sum_{\sigma:\sigma(n+1)=i} f^{\text{joint}}(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})}{\sum_\sigma f^{\text{joint}}(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})}$$

which, by weighted exchangeability with weights $w_i = \tilde{1}$ for $i \leq n$ and $w_{n+1} = w = \frac{dP_X^{\text{test}}(x)}{dP_X^{\text{train}}(x)}$ gives

$$\begin{aligned}
\mathbb{P}(V_{n+1} = v_i | E_v) &= \frac{\sum_{\sigma:\sigma(n+1)=i} w(z_{\sigma(n+1)}) g(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})}{\sum_\sigma w(z_{\sigma(n+1)}) g(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})} \\
&= \frac{\sum_{\sigma:\sigma(n+1)=i} w(z_i)}{\sum_\sigma w(z_{\sigma(n+1)})} \quad \text{by permutation invariance of } g \\
&= \frac{n! w(z_i)}{\sum_{j=1}^{n+1} n! w(z_j)} = p_i^w
\end{aligned}$$

meaning that

$$V_{n+1}|E_v \sim \sum_{i=1}^{n+1} p_i^w(z_1, \ldots, z_{n+1})\delta_{v_i}$$

Because we are under E_v and the same permutation maps Z_i and V_i to z_i and v_i, we can rewrite

$$V_{n+1}|E_v \sim \sum_{i=1}^{n+1} p_i^w(Z_1, \ldots, Z_{n+1})\delta_{V_i}$$

which leads to

$$\mathbb{P}\left(V_{y,n+1} \leq \text{Quantile}(1-\alpha; \sum_{i=1}^{n+1} p_i^w \delta_{V_i}) \;\middle|\; E_v\right) \geq 1-\alpha$$

and marginalizing,

$$\mathbb{P}\left(V_{y,n+1} \leq \text{Quantile}(1-\alpha; \sum_{i=1}^{n+1} p_i^w \delta_{V_i})\right) \geq 1-\alpha$$

We conclude by using the equivalence (which also applies with weighted quantiles) in the *proof of Theorem 1* to get

$$\mathbb{P}\left(V_{y,n+1} \leq \text{Quantile}(1-\alpha; \sum_{i=1}^{n} p_i^w \delta_{V_i} + p_{n+1}^w \delta_\infty)\right) \geq 1-\alpha$$

As in the *proof of Theorem 1*, having almost surely no ties, this probability is exactly equal to $\frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \leq 1-\alpha+\frac{1}{n+1}$. □

Even though weighted CP achieves marginal coverage, **Theorem 2** only proves so when the density ratio w is known, making it hardly applicable on real-life datasets.

2.4 Miscoverage Error when estimating w

We now propose a new lower bound showing how the density ratio estimation error impacts the weighted CP coverage.

Theorem 3. ***Worst Case Miscoverage for Likelihood Ratio Estimation***

Let $\hat{w}$ denote a density ratio estimator satisfying $\sup_{x\in\mathbb{R}} |\hat{w}(x) - w(x)| \leq \varepsilon$, let $D = \sum_{i=1}^{n+1} w(X_i)$ and $D' = \sum_{i=1}^{n+1} \hat{w}(X_i)$.

Under the condition that $D, D' \geq c > 0$, the CP interval $\hat{C}(X_{n+1})$ built with $\hat{w}$ achieves the following marginal coverage

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha - \frac{\varepsilon}{c}\left(n + \frac{(n+1)\sum_{i\leq n} w(X_i)}{c}\right)$$

Proof. Assuming that $\sup_{x\in\mathbb{R}} |\hat{w}(x) - w(x)| \leq \varepsilon$ and $D, D' \geq c > 0$, we get for all $i = 1, \ldots, n+1$

$$\begin{aligned}
|p_i^w - p_i^{\hat{w}}| &= \left|\frac{w(X_i)}{D} - \frac{\hat{w}(X_i)}{D'}\right| \\
&= \left|\frac{w(X_i)}{D} - \frac{w(X_i)}{D'}\right| + \left|\frac{w(X_i)}{D'} - \frac{\hat{w}(X_i)}{D'}\right| \quad \text{by triangle inequality} \\
&\leq \frac{w(X_i)|D - D'|}{DD'} + \frac{\varepsilon}{D'} \\
&\leq \frac{\varepsilon}{D'}\left(\frac{(n+1)w(X_i)}{D} + 1\right) \quad \text{as } |D - D'| \leq (n+1)\varepsilon \\
&\leq \frac{\varepsilon}{c}\left(\frac{(n+1)w(X_i)}{c} + 1\right) \quad \text{as } D, D' \geq c
\end{aligned}$$

Now considering F_{n+1} the cumulative distribution function of the weighted empirical distribution $\sum_{i=1}^n p_i^w \delta_{V_i} + p_{n+1}^w \delta_\infty$, and $\hat{F}_{n+1}$ its variant with $\hat{w}$, we have for $q \in \mathbb{R}$

$$\begin{aligned}
|F_{n+1}(q) - \hat{F}_{n+1}(q)| &\leq \sum_{i=1}^n |p_i^w - p_i^{\hat{w}}| \cdot \mathbb{1}_{V_i \leq q} \leq \sum_{i=1}^n |p_i^w - p_i^{\hat{w}}| \\
&\leq \frac{\varepsilon}{c}\left(\frac{(n+1)\sum_{i\leq n} w(X_i)}{c} + n\right) =: \Delta
\end{aligned}$$

Consequently, noting $\hat{q}_{1-\alpha}$ the $1-\alpha$ quantile of the estimated weighted empirical distribution, we get

$$\begin{aligned}
F_{n+1}(\hat{q}_{1-\alpha}) &\geq \hat{F}_{n+1}(\hat{q}_{1-\alpha}) - \Delta \\
&= 1 - \alpha - \Delta =: \beta
\end{aligned}$$

thus showing that $\hat{q}_{1-\alpha} \geq q_\beta$.

Finally,

$$\mathbb{P}\left(V_{y,n+1} \leq \hat{q}_{1-\alpha}\right) \geq \mathbb{P}\left(V_{y,n+1} \leq q_\beta\right) \geq \beta$$

□

Lets now try to approximate this bounds in usual settings: when n is high and the distribution shift is moderate. Having these assumptions, the Law of Large Numbers states that a c-value slightly lower

than $n \cdot \min\left(\mathbb{E}_{P_X^{\text{train}}}[w(X)]; \mathbb{E}_{P_X^{\text{train}}}[\hat{w}(X)]\right) \approx n \cdot \mathbb{E}_{P_X^{\text{train}}}[w(X)]$ satisfies with high probability $D, D' \geq c$ (we remove the test point drawn from P_X^{train} using $D \approx \sum_{i=1}^{n} w(X_i)$).

Therefore, under such settings, the miscoverage bound Δ can be approximated using

$$\frac{n\varepsilon}{c}\left(\frac{(n+1)\sum_{i\leq n} w(X_i)}{c} + 1\right) \approx \frac{2\varepsilon}{\mathbb{E}_{P_X^{\text{train}}}[w(X)]}$$

Having for example $\frac{1}{a} \leq \sup_{x\in\mathbb{R}} |w(x)|$, a shift under which the likelihood is at most divided by a, we get

$$\Delta \lesssim 2a\varepsilon$$

In practice, $D - D'$ is more likely to behave like a random walk of magnitude ε (with errors with different signs canceling each other). Therefore, it would remain to show that under certain regularity conditions we can get, with high probability, a new bound

$$\Delta' \approx \frac{\varepsilon}{c}\left(\frac{\sqrt{n+1}\sum_{i\leq n} w(X_i)}{c} + n\right)$$

The approximations above would then lead to

$$\Delta' \approx \frac{\varepsilon}{\mathbb{E}_{P_X^{\text{train}}}[w(X)]}(1 + \frac{1}{\sqrt{n}})$$

Remark 2.5. While having $\hat{q}_{1-\alpha} < q_{1-\alpha}$ violates the coverage guarantees, cases when $\hat{q}_{1-\alpha} > q_{1-\alpha}$ would lead to artificially wide intervals, eventually making the prediction intervals useless.

2.5 Estimating the Likelihood Ratio

In practice, $w(x)$ is often unknown and must be estimated. Let $\mathcal{D}_{\text{test}}^{\text{unlab}} = \{X_{n+1}, \ldots, X_{n+m}\}$ denote unlabeled test covariates.

Common approaches either indirectly estimate w through probabilistic classification or directly estimate it using non-parametric density estimation:

- **Logistic Regression**: First train a classifier to distinguish $\mathcal{D}_{\text{train}}$ from $\mathcal{D}_{\text{test}}^{\text{unlab}}$ and then use the odds ratio $\frac{P(\text{test}|X=x)}{P(\text{train}|X=x)}$ to estimate $w(x)$.
- **Density Ratio Estimation**: Directly model $w(x)$ via kernel methods or random forests.

As the accuracy of $\hat{w}(x)$ critically impacts performance (through coverage guarantees violation or interval widths inflation), large sets $\mathcal{D}_{\text{test}}^{\text{unlab}}$ are essential to mitigate estimation errors, especially in high-dimensional settings where the estimation can easily be biased or unstable.

Therefore, using CP under covariate shift requires multiple datasets: $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}}$ drawn from P^{train} (to fit the regression model(s) and compute the scores), $\mathcal{D}_{\text{test}}^{\text{unlab}}$ drawn from P_X^{train} to estimate w and eventually $\mathcal{D}_{\text{test}}$ drawn from P^{train} to check the coverage guarantee.

2.6 Effective Sample Size (ESS)

Prediction under covariate shift introduces the concept of **effective sample size (ESS)** (designed by [Gretton et al., 2009]), which quantifies the information loss caused by non-uniform likelihood ratio weights.

For identical training and calibration sizes weighted conformal prediction intervals exhibit greater variability in coverage rates compared to unweighted methods, mirroring the performance of unweighted conformal prediction when the training and calibration size is $n = \text{ESS}$.

The ESS heuristic is defined as:

$$\text{ESS} = \frac{\left(\sum_{i=1}^{n} w(X_i)\right)^2}{\sum_{i=1}^{n} w(X_i)^2} = \frac{\|w(X_{1:n})\|_1^2}{\|w(X_{1:n})\|_2^2}.$$

This formula measures how "concentrated" the weights are: while any constant w (including $w = 1$ when there is no distribution shift) gives $\text{ESS} = n$, highly non-uniform weights can lead to $\text{ESS} \ll n$, indicating significant efficiency loss.

Overall, the ESS framework underscores the critical balance between covariate shift correction and statistical efficiency, particularly in high-dimensional problems.

3 Conformal p-Values for Outlier Detection

This section focuses on the application of conformal prediction techniques to the task of outlier detection, where the goal is to identify which among new observations do not belong to the same distribution as a reference dataset. Building on the foundations established in **section 1**, we present conformal p-values as a principled approach to nonparametric outlier detection that provides finite-sample statistical guarantees.

3.1 Marginal Conformal p-values

In the context of outlier detection, we consider a dataset $\mathcal{D}$ containing independent and identically distributed points drawn from an unknown distribution in $\mathbb{R}^d$: P_X. Following the split conformal approach, we divide $\mathcal{D}$ into a training set $\mathcal{D}_{train}$ (with negative indexes for simplicity) and a calibration set $\mathcal{D}_{cal} = \{X_{n+i}\}_{i=1}^{n}$ of size n. The goal is to test whether each new test point in $D_{test} = \{X_{n+i}\}_{i=1}^{n_{test}}$ comes from the same distribution P_X (i.e., is an inlier) or from a different distribution (i.e., is an outlier).

Remark 3.1. In all this section we are only interested in the covariates, meaning that all the data from $\mathcal{D}$ and $\mathcal{D}_{\text{test}}$ can be unlabeled data. Consequently, the conformity scores are now only assessing the conformity of x with respect to the distribution P_X.

To perform this test, we first train a scoring function $\hat{s} : \mathbb{R}^d \to \mathbb{R}$ on $\mathcal{D}_{train}$ that assigns lower scores to potential outliers. The classical marginal conformal p-value for a test point x is then defined as:

$$\hat{u}^{(\text{marg})}(x) = \frac{1 + |\{i \in \mathcal{D}_{cal} : \hat{s}(X_i) \leq \hat{s}(x)\}|}{n+1} \in \left[\frac{1}{n+1}, 1\right]$$

This p-value represents the proportion, among the calibration points and the test point x, of points with scores less than or equal to the score of x. Intuitively, if x is an outlier, $\hat{s}(x)$ should be small, resulting in a small p-value.

The key property of these marginal conformal p-values is that they are marginally valid.

Theorem 4. ***Marginal Conformal p-values are Marginally Valid*** *(from [Bates et al., 2021])*

If X_{n+1} follows the same distribution P_X as the reference data, the marginal p-value constructed above is marginally valid in the sense that

$$\mathbb{P}(\hat{u}^{(marg)}(X_{n+1}) \leq t) \leq t$$

for any $t \in (0, 1)$.

Proof. We assume that $X_{n+1} \sim P_X$ and that, for simplicity, there is almost surely no ties in the scores $(\hat{s}(X_i))_{i=1}^{n+1}$. By i.i.d. assumption, we have when fixing S by conditioning on $\mathcal{D}_{\text{train}}$ that $\hat{u}^{(\text{marg})}(X_{n+1})|\mathcal{D}_{\text{train}}$ follows a uniform distribution on $\{\frac{1}{n+1}, \frac{2}{n+1}, \ldots, 1\}$ meaning that

$$\begin{aligned}\mathbb{P}(\hat{u}^{(\text{marg})}(X_{n+1}) \leq t|\mathcal{D}_{\text{train}}) &= \mathbb{P}(\hat{u}^{(\text{marg})}(X_{n+1}) \leq \lfloor t(n+1) \rfloor|\mathcal{D}_{\text{train}}) \\ &= \frac{\lfloor t(n+1) \rfloor}{n+1} \leq t \end{aligned} \tag{3.1}$$

Finally, after marginalizing

$$\mathbb{P}(\hat{u}^{(\text{marg})}(X_{n+1}) \leq t) \leq t$$

□

Remark 3.2. We get the same results without the assumption that there is almost surely no ties, simply replacing the equal sign in (3.1) by a $\leq$.

This guarantee holds marginally over both the randomness in $\mathcal{D}_{cal}$ and X_{n+1}. However, marginal validity may be insufficient in practice since it only ensures correct calibration on average over many random draws of the calibration data. For a practitioner working with a specific calibration dataset, stronger guarantees may be desirable.

3.2 Correction: Constructing Calibration-Conditional Valid p-values

To address the limitations of marginal p-values [Bates et al., 2021] introduce the concept of calibration-conditional valid (CCV) p-values, which satisfy a stronger property:

$$\mathbb{P}\left(\mathbb{P}[\hat{u}^{(\text{ccv})}(X_{n+1}) \leq t|\mathcal{D}] \leq t \text{ for all } t \in (0,1)\right) \geq 1-\delta$$

This means that with probability at least $1-\delta$ over the randomness in the calibration data D, the conditional probability of an inlier p-value being below any threshold t is at most t. This provides a much stronger guarantee for any single application.

The general strategy to construct CCV p-values is to apply an adjustment function h to the marginal p-values:

$$\hat{u}^{(\text{ccv})} = h \circ \hat{u}^{(\text{marg})}$$

where the adjustment function h is designed based on uniform bounds for the order statistics of uniform random variables.

Specifically, if $U_1, \ldots, U_n \overset{i.i.d.}{\sim} \text{Unif}([0,1])$ with order statistics $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(n)}$, and if one can find values $0 \leq b_1 \leq b_2 \leq \ldots \leq b_n \leq 1$ such that:

$$P(U_{(1)} \leq b_1, \ldots, U_{(n)} \leq b_n) \geq 1-\delta \qquad (\star)$$

Then the piece-wise constant function $h(t) = b_{\lceil (n+1)t \rceil}$ (where $b_{n+1} = 1$) for $t \in [\frac{1}{n+1}, 1]$ can be used to construct valid CCV p-values.

Theorem 5. ***Calibration Conditional Valid p-values*** *(from [Bates et al., 2021])*
Having values $0 \leq b_1 \leq b_2 \leq \ldots \leq b_n \leq 1$ satisfying ($\star$), the constructed p-values $\hat{u}^{(ccv)} = h \circ \hat{u}^{(marg)}$ are calibration conditional valid:

$$\mathbb{P}\left(\mathbb{P}[\hat{u}^{(ccv)}(X_{n+1}) \leq t|\mathcal{D}] \leq t \text{ for all } t \in (0,1)\right) \geq 1-\delta$$

Proof. We assume, for simplicity, that the distribution of the scores $\hat{s}(X_i)$ is continuous. Let $F^(t) := \mathbb{P}[\hat{s}(X_i) < t|\mathcal{D}_{\text{train}}]$ be the conditional CDF of the scores and define $U_i := F(\hat{s}(X_i))$ for $i = 1, \ldots, n$.

By construction, the U_is follow a uniform law on $[1, n]$. Additionally, when conditioning on $\mathcal{D}_{\text{train}}$, the score function $\hat{s}$ can be treated as fixed, and the independence of the calibration points consequently ensures the independence of the U_is.
Denoting $\varepsilon_n := \bigcap_{i=1}^n \{U_{(i)} \leq b_i\}$, we have by ($\star$) that $\mathbb{P}(\varepsilon_n) \geq 1-\delta$.

We now consider that ε_n holds and want to show that

$$\mathbb{P}[\hat{u}^{(\text{ccv})}(X_{n+1}) \leq t|\mathcal{D}] \leq t \text{ for all } t \in (0,1) \tag{3.2}$$

.

Because $\hat{u}^{(\text{ccv})}$ can only take values among the b_is, we have

$$\mathbb{P}[\hat{u}^{(\text{ccv})}(X_{n+1}) \leq t|\mathcal{D}] = \mathbb{P}[\hat{u}^{(\text{ccv})}(X_{n+1}) \leq b_j|\mathcal{D}]$$

where b_j is the highest b_i smaller than t.

If $b_j < 1$ (i.e. $j \leq n$), we have the equivalence

$$\hat{u}^{(\text{ccv})}(X_{n+1}) \leq b_j \iff \hat{u}^{(\text{marg})}(X_{n+1}) \leq \frac{j}{n+1} \iff \hat{s}(X_{n+1}) \leq s_{(j)}$$

where $s_{(j)}$ is the j^{th} largest score.

We get

$$\mathbb{P}[\hat{u}^{(\text{ccv})}(X_{n+1}) \leq t|\mathcal{D}] = \mathbb{P}[s(X_{n+1}) \leq s_{(j)}|\mathcal{D}] = F(s_{(j)}) = U_{(j)}$$

with the last equality being given by the monotonicity of F. Finally, (3.2) is satisfied as we are under ε_n and $b_j \leq t$.

If $b_j = 1$ (i.e. $j = n+1$), then (3.2) is immediately satisfied. □

3.3 Simes and DKWM Methods

Two primary methods for constructing the adjustment function h are the Simes adjustment and the Dvoretzky-Kiefer-Wolfowitz-Massart-Reeve (DKWM) adjustment.

3.3.1 DKWM Adjustment

The DKWM adjustment is based on the Dvoretzky-Kiefer-Wolfowitz-Massart-Reeve inequality for the uniform convergence of empirical cumulative distribution functions. This inequality establishes tight bounds on the probability that an empirical distribution function deviates significantly from the true underlying distribution.

Theorem 6. ***DKWM Inequality***

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed real-valued random variables with cumulative distribution function F, and let $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$ be the empirical distribution function.

Then for any $\varepsilon > 0$:

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}}(F_n(x) - F(x)) > \varepsilon\right) \leq e^{-2n\varepsilon^2}$$

For the two-sided version:

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}}|F_n(x) - F(x)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

Derivation of DKWM Bounds

In order to construct CCV p-values, we leverage the two-sided DKWM inequality to find values $0 \leq b_1 \leq b_2 \leq \ldots \leq b_n \leq 1$ satisfying $(\star)$.

We apply the two-sided DKWM inequality for uniform variables (on $[0,1]$) to get a confidence level $1-\delta$ on the complementary event:

$$\mathbb{P}\left(\sup_{t\in[0,1]} |t - F_n(t)| \leq \varepsilon\right) \geq 1-\delta \quad \text{where} \quad \varepsilon = \sqrt{\frac{\log(2/\delta)}{2n}}$$

implying that

$$\mathbb{P}\left(\sup_{i\leq n} F_n(b_i) \geq b_i - \varepsilon\right) \geq 1-\delta$$

Noting the event equality $\{U_{(i)} \leq t\} = \{F_n(t) \geq \frac{i}{n}\}$, we get that $(\star)$ is satisfied if $b_i - \varepsilon \geq |\frac{1}{n}|$ for all $i = 1, \ldots, n$.

Therefore, defining

$$b_i^d = \min\left\{\frac{i}{n} + \sqrt{\frac{\log(2/\delta)}{2n}}, 1\right\}, \quad i = 1, \ldots, n$$

we get that $b_1^d, \ldots, b_n^d$ satisfy $(\star)$.

3.3.2 Simes Adjustment

The Simes adjustment is based on the generalized Simes inequality from [Sarkar, 2008] and provides a particularly tight bound for small p-values, which is desirable in multiple testing contexts. For a given confidence level $1-\delta$ and parameter $k \leq n$, the bounds are defined as:

$$b_{n+1-i}^s = 1 - \delta^{1/k}\left(\frac{i \cdot \ldots \cdot (i-k+1)}{n \cdot \ldots \cdot (n-k+1)}\right)^{1/k}, \quad i = 1, \ldots, n$$

In practice, setting $k = n/2$ works well, making the Simes adjustment particularly effective for the smallest p-values while potentially being uninformative for larger ones.

For instance, [Bates et al., 2021] highlight that with $\delta = 0.1$ and $n = 1000$, the Simes adjustment would map the smallest possible marginal p-value of $1/(n+1)$ to approximately 0.0046 while the DKWM adjustment would map it to more than 0.1, making it too conservative for small p-values.

The difference in behavior makes the Simes adjustment more suitable for multiple testing in outlier detection, where accurately identifying the smallest p-values is crucial.

3.4 Multiple Testing with Benjamini-Hochberg

When testing multiple test points for outlier detection, it is important to account for multiplicity to control the overall error rate. The false discovery rate (FDR)-the expected proportion of false positives among all rejections-is a particularly relevant error metric in this context.

In the following, we denote $H_1, \ldots, H_m$ the null hypotheses that the corresponding test points are inliers, I_0 the true null hypotheses set (i.e. set of inliers) and $\pi_0 := \frac{|I_0|}{m}$.

Theorem 7. ***Benjamini-Hochberg procedure controls FDR***

For p-values $p_1, \dots, p_n$, rejecting $H_{(1)}, \dots, H_{(\hat{k})}$ where

$$\hat{k} = \max\Big\{ i : p_{(i)} \leq \frac{i\alpha}{m} \Big\}$$

maintains the FDR under level $\pi_0\alpha \leq \alpha$ if for all $i \in I_0$, p_i is independent of $\{p_j : j \neq i\}$.

Having non-independent p-values (on the set on inliers), we cannot directly apply the usual multiple-testing procedures that aims at controlling the FDR. However, a key result in [Bates et al., 2021] is that the marginal conformal p-values satisfy the positive regression dependence on a subset (PRDS) property on the set on inliers. PDRS being a special case of non-independent p-values that maintain Benjamini-Hochberg (BH) procedure guarantees, one can therefore apply BH with the conformal p-values.

Definition 3. ***Positive Regression Dependent on a Subset (PRDS) Random Vectors***

A random vector $X = (X_1, \dots, X_m)$ is PRDS on a subset $I_0 \subset \{1, \dots, m\}$ if

$$\mathbb{P}(X \in A | X_i = x) \quad \textit{is increasing in } x$$

for any $i \in I_0$ and increasing set A.

This property ensures that the Benjamini-Hochberg procedure applied to marginal conformal p-values controls the FDR at level $\pi_0\alpha \leq \alpha$, where π_0 is the proportion of true nulls (i.e., inliers) in the test set.

Remark 3.3. The BH procedure is particularly suited for setups where π_0 is close to one (i.e. most test point are inliers) but not otherwise as the control gap between $\pi_0\alpha$ and α is "lost" in the sense that we could have rejected more hypotheses while maintaining the FDR under α.

One can solve this issue applying Storey's correction from [Storey, 2002]: replacing $\hat{k}$ with $\max\{i : p_{(i)} \leq \frac{i\alpha\hat{\pi}_0}{m}\}$ where $\hat{\pi}_0$ is an estimation of π_0 (possibly very simple). While Storey's correction FDR control doesn't hold in general for PDRS p-values, [Bates et al., 2021] showed that it still holds in the specific context of marginal conformal p-values.

In summary, conformal p-values provide a powerful framework for nonparametric outlier detection with strong statistical guarantees. The choice between marginal and calibration-conditional p-values represents a trade-off between computational simplicity and strength of guarantees, with the latter providing stronger assurances at the cost of slightly more conservative inferences.

4 Outlier Detection under Distribution Shift

We now extend our framework to address outlier detection under distribution shift, with the framework of **section 2** where the test covariate distribution differs from the training distribution while the conditional distribution of outcomes given covariates remains unchanged.

Remark 4.1. Because we are performing outlier detection and the scores computation (both training and calibration parts) only require covariates X, we don't actually need the outputs Y and that the conditional distribution of outcomes given covariates remains unchanged. Indeed, we will consider in the following all the data as unlabeled data.

As seen in **subsection 2.3**, if we know (or estimate well-enough) the ratio of test to training covariate likelihoods, $w(x) = \frac{dP_X^{\text{test}}(x)}{dP_X^{\text{train}}(x)}$, we can still perform valid conformal inference by appropriately weighting the empirical distribution of conformity scores.

Under this framework, our goal remains to determine whether a new observation X_{test} is an outlier with respect to the test distribution P_X^{test}, while calibrating our procedure using the available training data that follows P_X^{train}.

We here consider that we have access to two datasets:

- a dataset $\mathcal{D}$ of points drawn from P_X^{train} that we split into $\mathcal{D}_{\text{train}}$ to train $\hat{s}$, $\mathcal{D}_{\text{cal}} = \{X_1, \ldots, X_n\}$ to compute the scores and, if w is not known, $\mathcal{D}_{\text{train},\, w}$ to train $\hat{w}$ (with some potential overlap between $\mathcal{D}_{\text{train},\, w}$ and $\mathcal{D}_{\text{train}}$).
- if w is not known, a dataset $\mathcal{D}_{\text{test},\, w}$ of points drawn P_X^{test} from to train $\hat{w}$

and we want to test our methods on the new test point X_{n+1}.

4.1 Weighted Conformal p-values for Outlier Detection

To account for distribution shift, we adapt the standard conformal p-value approach by incorporating the weights p_i^w derived in **subsection 2.3**.

Given a score function $\hat{s}$ and calibration data $\mathcal{D}_{\text{cal}} = \{X_1, \ldots, X_n\}$, we define the weighted conformal p-value for a test point x as

$$\hat{u}^{(w,\text{marg})}(x) = p_{n+1}^w + \sum_{i=1}^{n} p_i^w \cdot \mathbb{1}\{\hat{s}(X_i) \leq \hat{s}(x)\}$$

assigning greater importance to calibration points that are more likely under the test distribution. Intuitively, if $w(X_i)$ is large, the calibration point X_i is more likely under the test distribution than the training distribution, and therefore should have more influence on the p-value calculation.

Conversely, if $w(X_i)$ is small, the calibration point is less likely under the test distribution and should have less influence.

Unlike the standard conformal p-value, which gives equal weight to all calibration points, the weighted approach accounts for the relative likelihood of each point under the test distribution, ensuring that our outlier detection procedure remains valid despite the distribution shift.

Theorem 8. ***Weighted Marginal p-values are Marginally Valid***

If X_{n+1} follows the test distribution P_X^{test}, the weighted marginal p-value $\hat{u}^{(w,marg)}$ constructed above is marginally valid in the sense that

$$\mathbb{P}(\hat{u}^{(w,marg)}(X_{n+1}) \leq t) \leq t$$

for any $t \in (0,1)$.

Proof. We assume, for simplicity, that there is almost surely no ties between the test scores. The first part of the *proof of Theorem 2* gives us (simply replacing scores for $Z = (X, Y)$ by scores for X) that under the event E_s that $\{\hat{s}(X_1), \ldots, \hat{s}(X_{n+1}\} = \{s_1, \ldots, s_{n+1}\}$ with $s_1 \leq s_2 \leq \cdots \leq s_{n+1}$,

$$\hat{s}(X_{n+1})|E_v \sim \sum_{i=1}^{n+1} p_i^w \delta_{s_i} \tag{4.1}$$

where p_i^w is the weight corresponding to s_i

By definition of $\hat{u}^{(w,\text{marg})}$ we also have

$$\left\{\hat{u}^{(w,\text{marg})}(X_{n+1}) \leq t|E_v\right\} = \left\{\hat{u}^{(w,\text{marg})}(X_{n+1}) \leq \sum_{i=1}^{j} p_i^w \mid E_v\right\} = \left\{\hat{s}(X_{n+1}) \leq s_j \mid E_v\right\}$$

where j is the higher index such that $\sum_{i=1}^{j} p_i^w \leq t$.

Finally having

$$\mathbb{P}(\hat{u}^{(w,\text{marg})}(X_{n+1}) \leq t|E_v) = \mathbb{P}(\hat{s}(X_{n+1}) \leq s_j \mid E_v) \leq \sum_{i=1}^{j} p_i^w \leq t$$

by (4.1), which gives the result after marginalizing. □

Remark 4.2. In the context of Multiple Testing, one could proof that theses weighted marginal conformal p-values still satisfy the PRDS property, thus allowing for multiple testing with BH.

We now define $\hat{u}^{(w,\text{ccv})} := h \circ \hat{u}^{(\text{marg})}$ where h is defined as in **subsection 3.2** with bounds satisfying $(\star)$ (for example Simes or DKWM bounds).

The *proof of Theorem 5* still holds, meaning that p-values $\hat{u}^{(w,\text{ccv})}$ are calibration conditional valid:

$$\mathbb{P}\left(\mathbb{P}[\hat{u}^{(w,\text{ccv})}(X_{n+1}) \leq t|\mathcal{D}] \leq t \text{ for all } t \in (0,1)\right) \geq 1 - \delta \quad \text{if } X_{n+1} \sim P_X^{\text{test}}$$

4.2 Numerical Experiments

This section presents our experimental evaluation of conformal prediction methods applied to financial risk analysis using SEC-mandated corporate disclosures. We employ both standard and weighted conformal techniques to address the prevalent distribution shift in financial data over time, with particular emphasis on how regulatory changes impact prediction performance.

4.2.1 The 10-K Dataset

For our empirical evaluation, we utilize the dataset from [Kogan et al., 2009] consisting of annual publicly available financial reports (Form 10-K) filed with the Securities and Exchange Commission (SEC). This dataset contains 26,806 reports from publicly traded companies spanning the years 1996-2006. Each report is paired with measurements of stock return volatility for both the twelve months prior to the report (v_{-12}) and the twelve months following the report (v_{+12}).

The volatility is calculated as the standard deviation of daily stock returns over a twelve-month period, and we work, as in [Kogan et al., 2009], in the logarithmic domain as it is standard in finance. The distribution of log-volatility across companies exhibits an approximately normal distribution, but both the mean and variance evolve over time, reflecting changing economic conditions and regulatory environments.

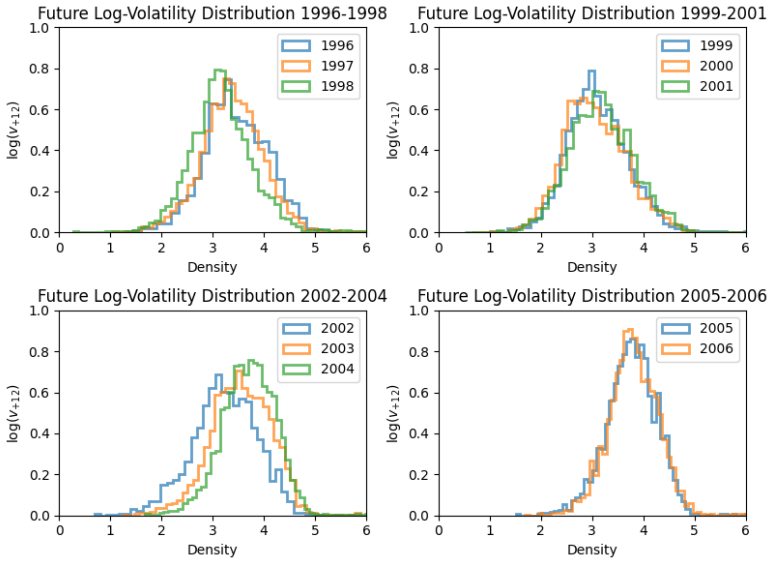


Figure 4: Distributions of $\log(v_{+12})$ across years

The 10-K reports, particularly Section 7 ("management's discussion and analysis of financial conditions and results of operations"), contain forward-looking content that may signal future financial risk making it, along with the publicly available v_{-12}, suitable for the prediction of v_{+12}.

For each year $y = 2001, \ldots, 2006$, we will consider as training data the data from the 5 previous years $y-5, \ldots, y-1$ and will use a proportion of the data from year y to estimate w the likelihood ratio. In practice, one wouldn't have to wait the end of the year to predict the future volatilizes as the estimation w only requires the covariates, making it applicable for real-life predictions.

Remark 4.3. As the training is made over the last five years, we could technically upgrade our weights by taking into account every distribution shift in the training set but this would much more computationally expensive.

For our feature representation, we adopt [Kogan et al., 2009]'s LOG1P approach with unigrams (meaning that the LOG1P frequency of each word in the vocabulary becomes a feature), which showed good performance in volatility prediction while not increasing the number of covariates compared to bigrams methods that consider pairs of consecutive words.

This representation is defined as $h_j(d) = \log(1 + \text{freq}(x_j; d))$, where $\text{freq}(x_j; d)$ denotes the number of occurrences of the x_j, the jth word in the vocabulary, in document d.

Specifically, in order to reduce the dimensionality, we will only consider the p^{th} more frequent words among all documents. We will assign to this parameter p different values between 0 and 50 to explore how the features dimension impacts the CP guarantees.

A notable characteristic of this dataset is the substantial increase in document length following the passage of the Sarbanes-Oxley Act of 2002, which imposed revised standards on financial reporting. The average document length nearly doubled from approximately 6,000 words in 2001 to over 12,000 words by 2005. This regulatory change creates a natural experiment for testing conformal prediction methods under distribution shift.

4.2.2 Comparison of CP Intervals

We now compare three setups of features size: a baseline setup where the only feature is $\log(v_{-12})$, an intermediate setup where the 10 most frequent words are added to $\log(v_{-12})$ and one where the 20 most frequent words are added to $\log(v_{-12})$.

After testing for Support Vector Regression (SVR) and LASSO-type models as experimented on the same dataset by [Kogan et al., 2009] and [Meinshausen & Bühlmann, 2015], we kept the Ordinary Least Square model for all the scores training because it performed as well on average and for its computational efficiency. For the density model $\hat{w}$, we use a Logistic Regression classifier as detailed in the beginning of **subsection 2.5**.

In all the setups defined above we split $\mathcal{D}$ the dataset of the previous five years in two equal size sets $\mathcal{D}_{\text{train}}$ (used to train both S and $\hat{w}$) and $\mathcal{D}_{\text{cal}}$ (used to compute the scores).

We also split the dataset of the test year y into two equal size sets $\mathcal{D}_{w,\text{test}}$ (to train $\hat{w}$) and $\mathcal{D}_{\text{test}}$ (used to test the validity of the intervals).

Figure 5 compares weighted and unweighted conformal prediction methods using the log past volatility as the sole feature. The results clearly showcase the advantages of weighted conformal prediction with residuals, which maintains the target 90% coverage during the first year while the non-weighted approach fails to accomplish. Additionally, the weighted residual method consistently sustains the 90% coverage threshold (or gets sufficiently close to it in 2003) across subsequent years while simultaneously producing shorter prediction intervals, representing a significant efficiency improvement in uncertainty quantification. For the rescaled variant, the outcome presents a more nuanced picture. The weighted rescaled method systematically increases coverage percentages but sometimes unnecessarily as observed in 2004 and 2005, thus resulting in wider average interval lengths. Nevertheless, it still delivers superior

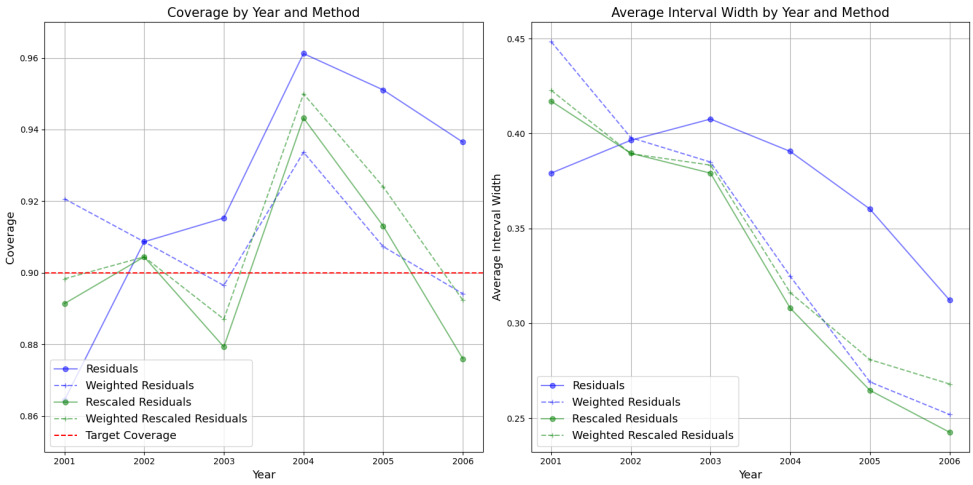


Figure 5: Distributions of $\log(v_{+12})$ across years

performance for all other analyzed years compared to its non-weighted counterpart.

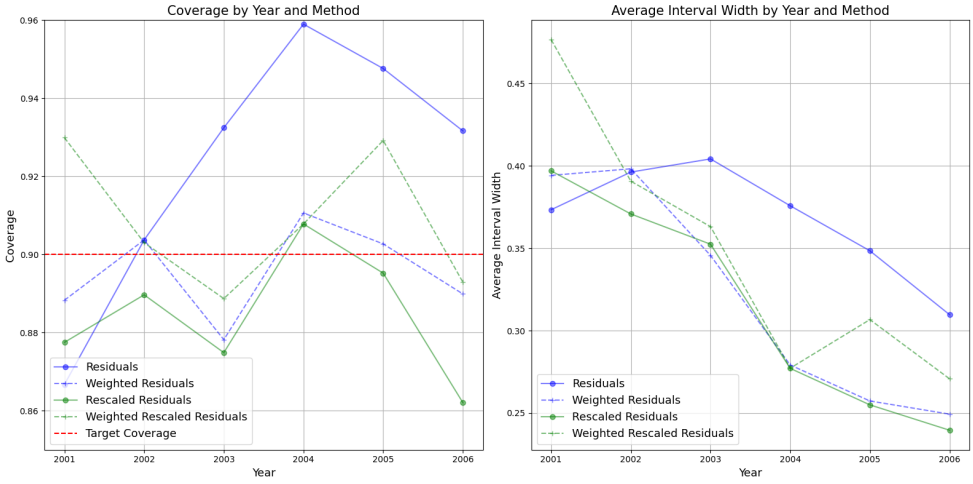


Figure 6: Distributions of $\log(v_{+12})$ across years

Figure 6, which presents results after adding ten features, reveals patterns very consistent with the single-feature scenario. The weighted residuals approach still performs well by substantially reducing interval lengths while maintaining coverage percentages very close to the desired 90% threshold. The rescaled method also exhibits identical behavioral patterns to those observed in Figure 5, with weighted rescaled versions producing higher coverage at the expense of increased interval lengths. Notably, the higher-dimensional feature space appears beneficial despite the additional computational complexity of estimating likelihood ratios in higher dimension. This finding is particularly interesting given that incorporating the ten additional covariates only reduces the regression Mean Squared Error by a modest 1.7%, indicating that these features have a disproportionately larger impact on conformal prediction interval quality than on point estimation accuracy.

When expanding to twenty features, the increased dimensionality creates substantial challenges for

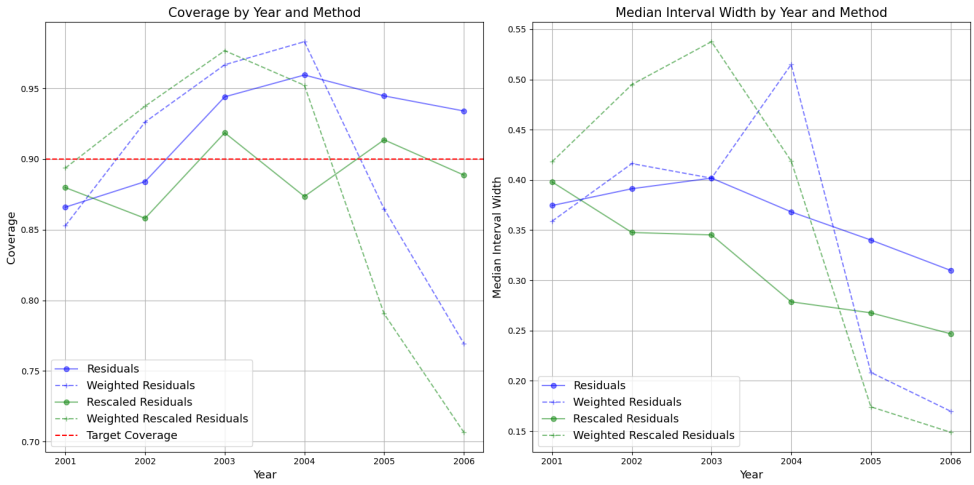


Figure 7: Distributions of $\log(v_{+12})$ across years

weighted conformal prediction methods. The performance deteriorates dramatically, necessitating the reporting of median interval lengths since a significant proportion of intervals (ranging from 0.6% in 2006 to a concerning 36% in 2002) reached infinite length. Even when considering only the finite intervals through median length measurement, the weighted methods demonstrate poor efficiency when achieving the target 90% coverage rate. This degradation likely stems from the difficulty in accurately estimating likelihood ratios in high-dimensional spaces. In fact, as dimensionality increases, the estimation of the ratio between the test and training distributions becomes increasingly unstable, leading to extreme weights that compromise the reliability of the resulting prediction intervals.

We also provide an estimation of the ESS in different setups. As shown in the last column of **Table 1**, it is very likely that the ESS estimation performs very poorly (due to $\hat{w}$ inaccuracy) in high dimension as the weighted CP intervals still manage to achieve the 90% coverage with a good average interval width when 10 features are added.

Year	**Baseline**	**Added 1 feature**	**Added 2 features**	**Added 5 features**
2001	76.13%	70.39%	49.36%	6.53%
2002	99.96%	93.80%	78.89%	8.70%
2003	94.77%	88.62%	84.01%	18.32%
2004	63.80%	62.59%	59.45%	23.72%
2005	56.01%	55.24%	54.02%	43.21%
2006	66.05%	65.87%	64.17%	58.11%

Table 1: Estimated ESS across years for the 10-K dataset

It appears that with appropriate feature dimensionality, weighted methods can significantly outperform traditional conformal prediction by maintaining target coverage while reducing interval lengths. However, there exists a clear tipping point beyond which the curse of dimensionality severely impairs weighted conformal prediction, as evidenced by the poor performance with twenty additional features. This suggests that practitioners should carefully balance the benefits of additional features against the increasing difficulty of accurate likelihood ratio estimation in higher dimensions. It also highlights the importance of feature selection methods when dealing with distribution shifts in high-dimension settings.

4.2.3 Outlier Testing

In hypothesis testing, we usually consider two primary metrics to evaluate the performance of a test: type I and type II errors. A type I error occurs when we incorrectly reject a true null hypothesis (a false positive), while a type II error happens when we fail to reject a false null hypothesis (a false negative). These error types represent fundamental trade-offs in statistical decision-making. **Theorem 8** guarantees marginal validity for weighted marginal conformal p-values, ensuring that p-values $\hat{u}^{(w,\text{marg})}$ should, on average over multiple calibration sets $\mathcal{D}_{\text{cal}}$, control the type I error rate to remain below the specified rejection level β. On the other side, the weighted CCV p-values should also maintain the type I error below β in at least a proportion $1-\delta$ of the calibration sets $\mathcal{D}_{\text{cal}}$. Following the methodology from [Bates et al., 2021], we employed an Isolation Forest approach to compute the score $\hat{s}$.

Remark 4.4. In order to generate some outliers, we randomly changed 30% of the features of some points into outliers with respect to the corresponding feature distribution, using the 1.5 Inter-Quantile Range (IQR) rule designed in [Tukey, 1977]: assigning the feature above $q_{75\%} + 1.5 \cdot IQR$ or below $q_{25\%} - 1.5 \cdot IQR$ where $IQR = q_{75\%} - q_{25\%}$.

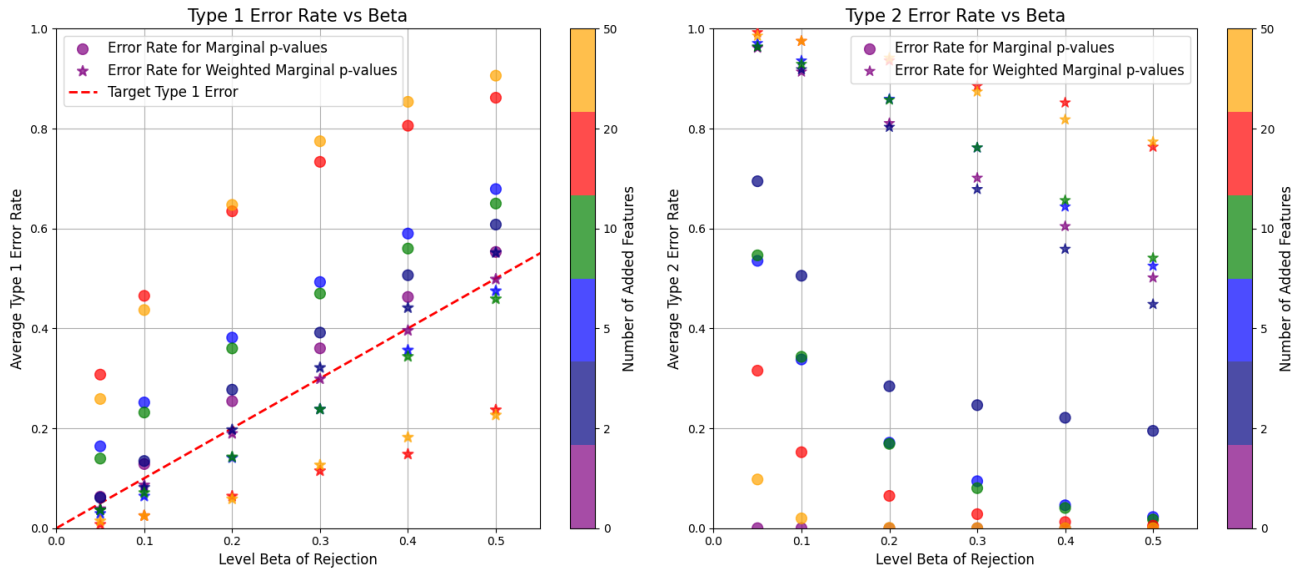


Figure 8: Distributions of $\log(v_{+12})$ across years

We plotted the error rates as a function of the rejection level β for our comparative analysis, with Figure 8 showing type I error rates and type II error rates across different testing configurations. These experiments compared weighted marginal and standard marginal methods under various setups with different numbers of added features. The results demonstrate that weighted marginal methods consistently achieve the target type I error control, remaining at or below the specified β level in most experimental scenarios. In contrast, the non-weighted methods consistently fail to maintain proper type I error control, exceeding the target threshold. However, this improved type I error control comes at a substantial cost: weighted methods exhibit significantly larger type II error rates, indicating that they fail to reject many outliers. This performance characteristic suggests that weighted marginal methods have limited practical applicability except in contexts where highly conservative decision-making is paramount and maintaining a low type I error rate takes absolute priority over statistical power.

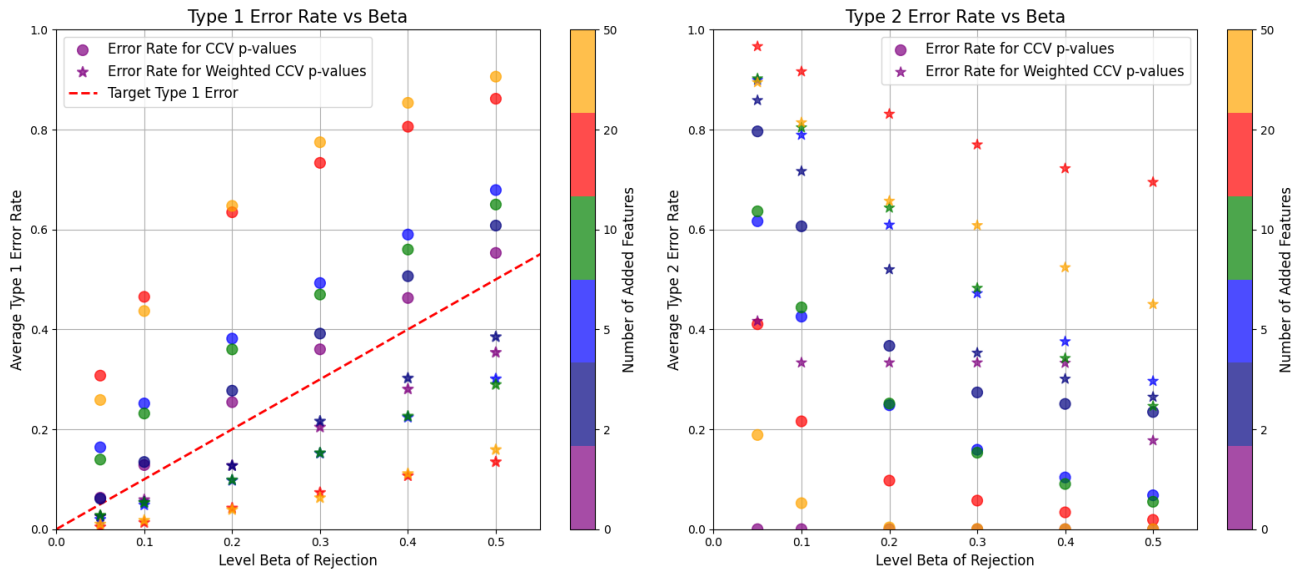


Figure 9: Distributions of $\log(v_{+12})$ across years

Figures 9 present analogous error analyses for Conditional Conformal Validation (CCV) and Weighted CCV methods. Similar to the pattern observed with marginal p-values, the weighted CCV approaches successfully achieve the target type I error control while their non-weighted counterparts fail to maintain error rates below the specified threshold. The weighted CCV methods also show inflated type II error rates compared to standard CCV, reflecting a reduced ability to identify true anomalies. However, it's notable that the difference in type II error rates between weighted and non-weighted variants is less pronounced in the CCV framework than in the marginal one. This suggests that the weighted CCV approaches might offer a more balanced trade-off between the two error types compared to weighted marginal methods.

Overall, while weighted methods (both marginal and CCV) almost invariably provide better control of type I errors, their substantially higher type II error rates make them less competitive in many practical applications where detecting true anomalies is equally important. The reduced rejection power means these methods often fail to identify genuine outliers, limiting their utility in scenarios requiring balanced performance. An additional observation from our experiments is that the type I error appears to increase with feature size across all methods, suggesting that high-dimensional data presents greater challenges for maintaining error control.

The poor type II error performance of weighted methods is very likely caused by inaccurate likelihood ratio estimation, particularly in higher dimensions where such estimation becomes even more challenging. Notably, even in the weighted scenarios without added features, the likelihood estimation remains inherently difficult – a limitation clearly visible in the type II error rates patterns across both marginal and CCV method figures. Furthermore, increased dimensionality degrades the accuracy of conformity scores themselves, as the curse of dimensionality impacts the precision of distance-based metrics and distributional assumptions. This dual effect (compromised likelihood estimation and reduced score reliability) collectively diminishes the competitiveness of conformal methods in high-dimensional settings, exacerbating the trade-off between type I error control and detection power.

Conclusion

This work proposes two contributions to conformal prediction under distribution shift and presents an empirical evaluation of both existing and newly developed methods on a real-world financial dataset exhibiting temporal distribution shifts. First, we derived a new lower bound that quantifies how density ratio estimation errors impact weighted conformal prediction coverage. Second, we introduced a framework for constructing valid p-values under distribution shift by incorporating likelihood ratio weighting into the calibration process, establishing a new method for nonparametric outlier detection under distribution shift.

Our experiments on the SEC 10-K financial dataset revealed some insights into the practical applications of these methods. Notably, weighted conformal prediction demonstrated slight advantages in maintaining target coverage while reducing interval widths in low to moderate dimensions. However, we observed a clear dimensionality threshold beyond which weighted methods deteriorate dramatically, with intervals frequently reaching infinite length when using 20 features. This deterioration highlights a fundamental limitation caused by the curse of dimensionality in likelihood ratio estimation.

In the context of outlier detection, our experiments uncovered an important trade-off: weighted methods consistently achieve better type I error control but suffer from substantially higher type II error rates. This asymmetric performance suggests these approaches could only be suited for applications where minimizing false alarms takes absolute priority over detection power.

Future research should focus on selecting more robust likelihood ratio estimation techniques for high-dimensional data, potentially through dimensionality reduction or regularization approaches. Additionally, exploring methods to balance the type I/type II error trade-off in outlier detection under shift represents an important direction, as does extending the weighted CP framework to handle more complex forms of distribution shift beyond covariate shift.

By linking theoretical results with practical applications, this work aims to offer useful insights and guidance for practitioners addressing the challenges of uncertainty quantification in non-stationary environments.

References

[Lei et al., 2018] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, *Distribution-free predictive inference for regression*, Journal of the American Statistical Association, vol. 113, no. 523, pp. 1094– 1111, 2018.

[Romano et al., 2019] Yaniv Romano and Evan Patterson and Emmanuel J. Candès, *Conformalized Quantile Regression*, `https://arxiv.org/abs/1905.03222`, 2019

[Koenker & Bassett, 1978] Koenker, Roger, and Gilbert Bassett. "Regression Quantiles." Econometrica, vol. 46, no. 1, 1978, pp. 33–50. `https://doi.org/10.2307/1913643.`

[Tibshirani et al., 2019] R. J. Tibshirani, R. Foygel Barber, E. Candès, and A. Ramdas, *Conformal prediction under covariate shift*, in Advances in Neural Information Processing Systems 32, 2019, pp. 2530–2540.

[Gretton et al., 2009] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Scholkopf. *Covariate shift by kernel mean matching.* In Dataset Shift in Machine Learning, chapter 8, pages 131–160. MIT Press, 2009.

[Bates et al., 2021] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, Matteo Sesia, *Testing for Outliers with Conformal p-values*, `https://arxiv.org/abs/2104.08279`, 2021

[Kogan et al., 2009] Shimon Kogan, Dimitry Levin, Bryan Routledge, Jacob Sagi, Noah Smith, *Predicting Risk from Financial Reports with Regression*, DOI 10.3115/1620754.1620794, pp. 272-280, 2009.

[Meinshausen & Bühlmann, 2015] Meinshausen, Nicolai and Bühlmann, Peter, *Maximin effects in inhomogeneous large-scale data*, The Annals of Statistics, volume 43, 2015.

[Tukey, 1977] Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

[Storey, 2002] Storey, John D. *A Direct Approach to False Discovery Rates.* Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64, no. 3 (2002): 479–98. `http://www.jstor.org/stable/3088784`.

[Sarkar, 2008] S. K. Sarkar. *Generalizing Simes' test and Hochberg's stepup procedure.* Annals of Statistics 36.1 (2008), pp. 337-363.

[Topics in Stats Theory] Richard J. Samworth. *Topics in Statistical Theory*, Part III, University of Cambridge, 2024.

[Modern Stats Methods] Rajen D. Shah, Sergio Bacallado. *Modern Statistical Methods*, Part III, University of Cambridge, 2024.