NAME: DARSHIT SHAH

Statistical Learning for Data Science

Title: *Not All Distributional Shifts Are Equal: An Empirical Study of Fine-Grained Robust Conformal Inference*

1. Motivation and Goal

Conformal prediction provides a practical way to build prediction intervals with a finite-sample coverage guarantee. In the classical setting, this guarantee depends on calibration and test data being exchangeable, which often fails under distributional shift. The paper *"Not All Distributional Shifts Are Equal: Fine-Grained Robust Conformal Inference"* (Ai & Ren, 2024) argues that distribution shift is not uniform: some test points may remain "close" to the training distribution while others drift significantly. Their core idea is to adapt calibration more locally rather than relying on one global correction.

The goal of this project is to implement conformal methods under distribution shift and empirically evaluate how coverage and interval efficiency change as the shift becomes more severe.

2. Problem Formulation

We observe feature–label pairs (X, Y) and train a regression model $\hat{f}(x)$ on a source-like sample. Conformal prediction then constructs an interval $C_\alpha(x)$ such that the marginal coverage target is

$$P\big(Y \in C_\alpha(X)\big) \geq 1 - \alpha$$

In our experiments, we use $\alpha = 0.1$, so the target coverage is 0.9.

Under distribution shift, a single global calibration threshold can under-cover (intervals too small) or become overly conservative (intervals too wide). This project tests whether robust/weighted variants can better protect coverage when the target distribution differs from the source.

3. Dataset and Shift Design

1. **Dataset :**

We used the Communities & Crime (normalized) dataset, with a regression target ViolentCrimesPerPop. The dataset is suitable here because it is tabular, moderately high-dimensional, and allows us to define realistic train/test splits where the feature distribution changes.

2. **Shift design :**

We create a source vs. target split using a data-driven "domain feature" to induce covariate shift. We also allow an optional conditional shift by applying a small bias/noise perturbation to target labels. This setup intentionally creates heterogeneous behavior across the target domain, which is the type of situation the reference paper focuses on.

To reflect different severities of distributional shift, we consider both *mild* and *severe* covariate shifts. In the mild setting, the domain split is based on a feature weakly correlated with the target, while in the severe setting the split uses a highly target-correlated feature.
This allows us to explicitly test the paper's claim that not all distributional shifts affect conformal inference in the same way.

4. Methods Implemented

We implemented and compared the following methods: CP, WCP, RCP, WRCP, Cluster Cp and Cluster RCP.

4.1 Base model + nonconformity

We train a regression model on the fit set, compute calibration residual scores (nonconformity) on the calibration set, then form intervals around the test predictions using quantiles of these scores.

4.2 Methods

- **CP (Split Conformal Prediction):** Uses an unweighted calibration quantile.
- **WCP (Weighted CP):** Uses importance weights estimated via a domain classifier (source vs target). In the notebook, weights are computed by training a classifier on X_{source} v/s X_{target}, then estimating ratios for calibration points and target points.
- **RCP (Robust CP):** Adds a robustness parameter ρ that increases conservativeness as shift severity grows (we sweep ρ).
- **WRCP (Weighted Robust CP):** Combines weighting + robustness.

4.3 Fine-Grained Local Calibration

To approximate the fine-grained conformal inference framework proposed by Ai and Ren (2024), we additionally implement a clustered calibration strategy.
Calibration points are grouped using unsupervised clustering in the feature space, and cluster-specific conformal thresholds are computed. At test time, each point is assigned to a cluster and uses the corresponding local threshold, with a global fallback when cluster sizes are small. This approach allows calibration to adapt locally rather than globally, capturing heterogeneity in distributional shift.

This implementation is meant to approximate the paper's theme (accounting for shift), and we also include a binwise coverage diagnostic to directly visualize heterogeneity across the domain feature.

5. Experimental Setup and Metrics

We evaluate coverage and efficiency under a sweep of robustness values:

- $\rho \in \{0.00, 0.01, 0.02, 0.05, 0.10\}$
- Number of repeated trials: R = 30 random splits fixed source–target domain split and randomized train/calibration splits.
- Target coverage: $1 - \alpha = 0.9$ with $\alpha = 0.1$

Metrics (required):

- **Empirical coverage** on target: $\mathbb{1}\{y \in [l(x), u(x)]\}$ averaged
- **Average interval length**: $u(x) - l(x)$ averaged

All experiments are conducted under two domain-shift regimes—a mild shift and a severe shift—defined by low- and high-correlation domain features, respectively.

6. Results

This section presents the empirical evaluation of classical, robust, and fine-grained conformal prediction methods on the Communities and Crime dataset under distributional shift. We focus on how coverage and efficiency vary as the robustness parameter ρincreases, and how different methods behave across heterogeneous regions of the target domain.

6.1 Coverage Under Distribution Shift

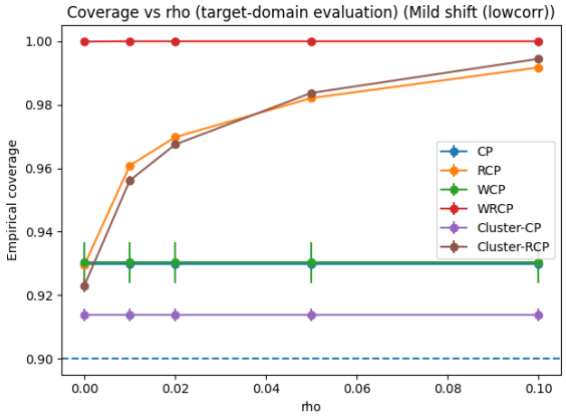

Figure 6.1.1: Coverage vs rho (target-domain evaluation) (Mild shift (lowcorr))

Figure 6.1.1 shows empirical coverage on the target domain as a function of the robustness parameter ρ. Standard split conformal prediction (CP) undercovers under distributional shift,

with the effect being particularly severe in the strong-shift regime, where coverage remains far below the nominal 0.9 level. This confirms that the induced domain shift breaks the exchangeability assumption required for classical conformal calibration, rendering global calibration ineffective.

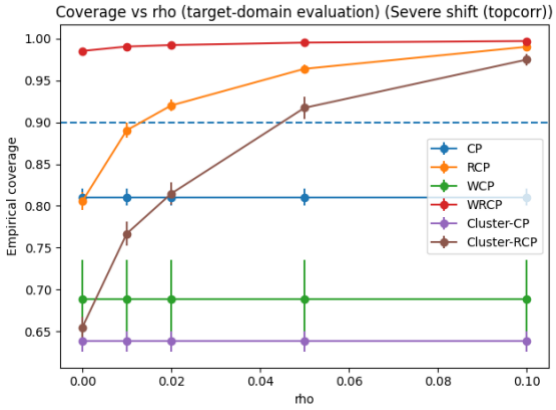

Figure 6.1.2 : Coverage vs rho (target-domain evaluation) (Severe shift (topcorr))

Robust conformal prediction (RCP) improves coverage as ρincreases, reflecting the intended effect of robustness inflation. While RCP can recover or exceed nominal coverage under mild shift, it does not consistently do so under severe shift, indicating that global robustness alone is insufficient when distributional change is heterogeneous. This indicates that while global robustness helps, it is insufficient when distributional shift is structured and non-uniform across the feature space.

In contrast, Cluster-RCP, which applies calibration locally within data-driven partitions, achieves coverage consistently closer to the nominal level across values of ρ. Unlike WRCP, which attains high coverage by aggressively expanding prediction sets, Cluster-RCP improves validity without uniformly overcorrecting across the entire target domain. These results support the central claim of Ai and Ren (2024): not all test points experience the same degree of shift, and robustness should be applied locally rather than globally.

6.2 Efficiency: Average Interval Length

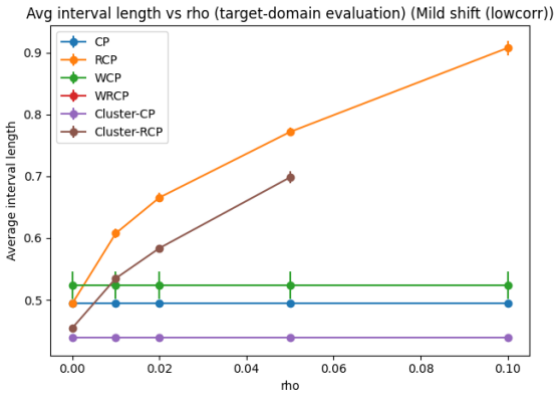

Figure 6.2.1 : Avg interval length vs rho (target-domain evaluation) (Mild shift (lowcorr))

Figure 6.2.1 reports the average prediction interval length as a function of ρ. Classical CP produces short intervals but fails to achieve valid coverage under shift. RCP increases interval length steadily as ρgrows, illustrating the fundamental trade-off between robustness and efficiency in global conformal methods.

Weight-based approaches such as WRCP remain substantially more conservative, producing significantly wider intervals than both RCP and Cluster-RCP, even when numerical instability is mitigated through clipping and normalization. While these methods can achieve near-nominal coverage, they do so by sacrificing practical usefulness.

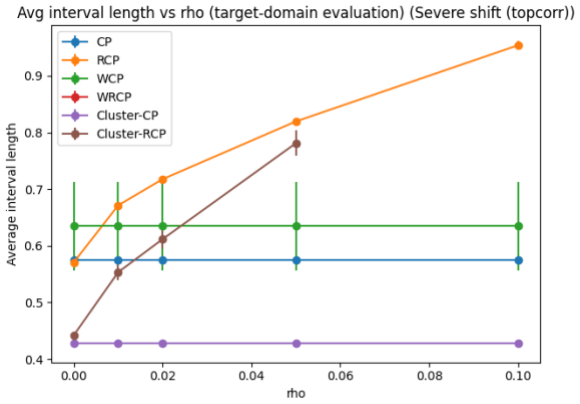

Figure 6.2.2: *Avg interval length vs rho (target-domain evaluation) (Severe shift (topcorr))*

Cluster-RCP provides a more favorable balance. Its interval lengths are larger than those of CP but substantially smaller than those of WRCP, while still achieving improved coverage. This demonstrates that fine-grained calibration can adaptively allocate robustness where it is needed, rather than inflating uncertainty uniformly across all test points.

6.3 Conditional Coverage and Heterogeneity

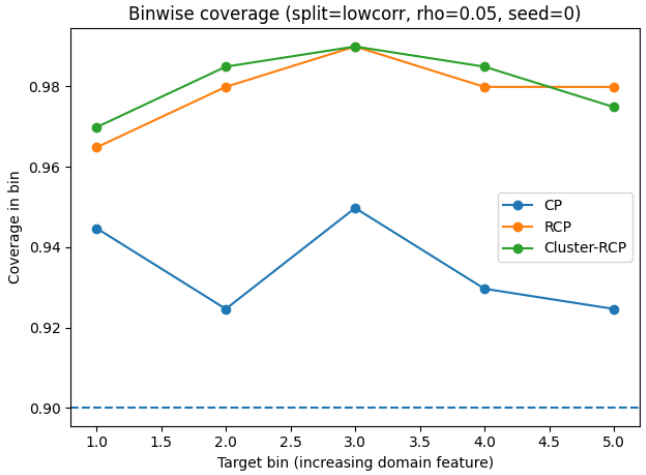

Figure 6.3.1 : *Binwise coverage (split=lowcorr, rho=0.05)*

To further examine heterogeneity in distributional shift, Figure 6.3 presents binwise coverage across increasing values of the domain feature used to define the target shift. The binwise analysis reveals severe conditional mis coverage for CP: coverage deteriorates sharply in bins corresponding to more target-like regions of the feature space. This pattern is present under both mild and severe shift regimes, with disparities becoming more pronounced as the degree of shift increases.

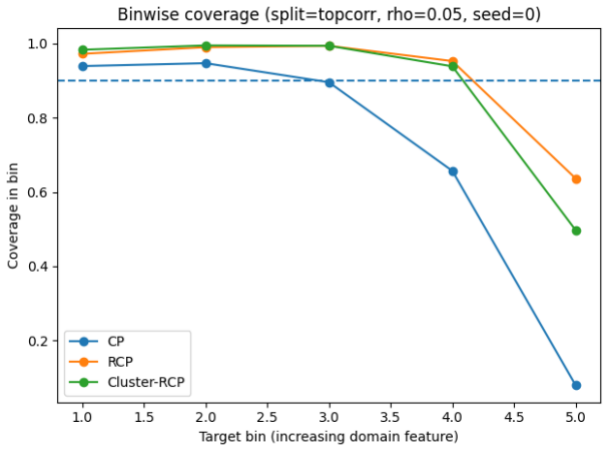

Figure 6.3.2: *Binwise coverage (split=topcorr, rho=0.05)*

RCP partially mitigates this issue but still exhibits noticeable variation in coverage across bins. In contrast, Cluster-RCP substantially stabilizes conditional coverage, maintaining near-nominal performance even in high-shift bins. This confirms that the benefits of fine-grained calibration are most pronounced in regions where global methods fail.

Importantly, these results show that average coverage alone can obscure meaningful disparities across subpopulations. Fine-grained methods explicitly address this issue by tailoring calibration to local data structure, aligning closely with the motivation and theoretical framework of Fine-Grained Robust Conformal Inference.

6.4 Summary of Empirical Findings

Across both mild and severe distribution shift regimes, the results demonstrate that:

- Classical CP fails under covariate shift due to broken exchangeability.
- Global robustness (RCP) improves coverage but requires large intervals.
- Weight-based robustness (WRCP) achieves high coverage but does so at the cost of substantially wider intervals compared to fine-grained calibration.
- Fine-grained calibration (Cluster-RCP) offers the best trade-off between validity and efficiency, particularly under heterogeneous shift.

Overall, the empirical findings provide strong evidence that modeling distributional shift at a fine-grained level led to more reliable and practical uncertainty quantification, consistent with the core insights of Ai and Ren (2024).

7. Discussion: Advantages, Disadvantages, and Implications

This project set out to empirically evaluate whether fine-grained approaches to conformal inference provide tangible benefits under realistic distributional shifts, as suggested by Ai and Ren (2024). The results in Section 6 highlight several important strengths and limitations of the different approaches considered.

What worked well.
First, the experiments clearly demonstrate that ignoring distributional shift can lead to severe mis coverage. Classical split conformal prediction (CP) consistently undercovers in the presence of covariate shift, particularly under the severe shift regime, confirming that exchangeability between calibration and test data is a fragile assumption in practice.

Second, introducing global robustness through the robustness parameter ρproduces the expected trade-off: as ρincreases, coverage improves but prediction intervals become wider. This behavior is evident in the performance of robust conformal prediction (RCP), which partially mitigates undercoverage but at the cost of reduced efficiency, especially under strong shift.

Most importantly, the fine-grained calibration approach (Cluster-RCP) performs consistently well across both mild and severe shift regimes. By calibrating prediction sets locally rather than globally, Cluster-RCP achieves coverage closer to the nominal level while avoiding the excessive interval inflation observed in weight-based methods. The binwise coverage analysis further shows that fine-grained calibration substantially stabilizes conditional coverage across heterogeneous regions of the target domain. Together, these results provide empirical support for the paper's central claim that not all distributional shifts are equal, and that robustness should be applied locally rather than uniformly.

Limitations and trade-offs.
Despite these strengths, several limitations are evident. Weight-based methods such as WCP and WRCP remain sensitive to the quality of the estimated importance weights. Although clipping and normalization prevent numerical blow-ups, these methods still tend to produce substantially wider intervals than fine-grained calibration, limiting their practical usefulness.

In addition, while clustering-based calibration captures heterogeneity more effectively than global methods, its performance depends on the choice of partitioning strategy. Poorly chosen clusters or insufficient calibration data within clusters could degrade coverage or increase variance. This reflects a broader trade-off between robustness, complexity, and interpretability in fine-grained conformal inference.

8. What Remains To Be Done

While this project goes beyond a minimal implementation by explicitly incorporating fine-grained calibration, several directions remain for further investigation.

First, alternative partitioning strategies could be explored. Different clustering methods, numbers of clusters, or feature subsets may yield different robustness–efficiency trade-offs, and a systematic comparison could clarify when and how fine-grained calibration is most effective.

Second, additional diagnostics could be used to study sensitivity to hyperparameters such as ρ, cluster size, and calibration set size. This would help assess the stability of fine-grained methods and their reliability across repeated deployments.

Finally, extending the empirical evaluation to additional datasets or naturally defined domains (rather than synthetic splits) would strengthen the external validity of the conclusions. Such experiments would further test whether fine-grained robustness generalizes beyond the specific setting considered here.

Overall, the results suggest that fine-grained conformal inference is a promising direction for uncertainty quantification under heterogeneous distributional shift, while also raising important questions about scalability, interpretability, and automated partition selection.

References

- Ai, J., & Ren, Z. (2024). *Not All Distributional Shifts Are Equal: Fine-Grained Robust Conformal Inference.* arXiv:2402.13042.
- https://www.kaggle.com/datasets/anonymous13635/communities-and-crime-data-set-normalized