Advanced Topic A09: *Conformal Prediction*

Data Science with Python – BSc Advanced Lectures

Joerg Osterrieder

10 Minutes

Motivation: Honest Uncertainty Without Distributional Assumptions

Most ML models produce point predictions – conformal prediction adds calibrated sets

- A neural network outputs a single value (or class probability); how confident should we be?
- Bayesian intervals require distributional assumptions; bootstrap is asymptotically valid only
- **Conformal prediction** produces prediction sets with a *guaranteed* coverage probability: $\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha$
- The guarantee is distribution-free: it holds for any model, any data distribution, any sample size
- Finance: a 95% conformal VaR interval means the true loss falls outside it at most 5% of the time – by construction, not by assumption

Conformal prediction: the only distribution-free, model-agnostic coverage guarantee in ML

The Core Idea: Nonconformity Scores

Measure how unusual a new point is relative to calibration data

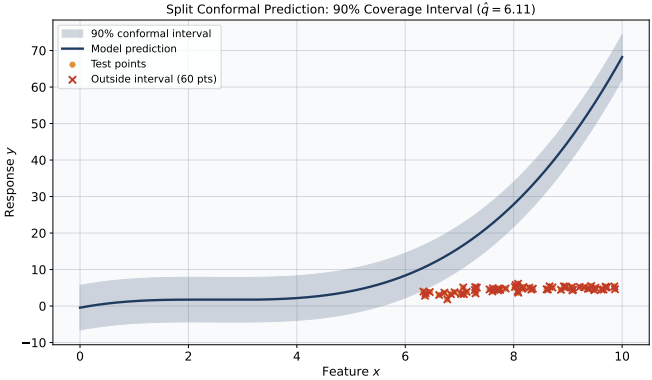
- Choose a **nonconformity score** $s(x, y)$: large when (x, y) does not fit the model well
- Examples: $|y - \hat{f}(x)|$ for regression; $1 - \hat{p}(y|x)$ for classification
- Compute scores on a held-out **calibration set** (not used for training)
- For a new test point, the prediction set contains all y values that are "conforming" relative to the calibration distribution
- Mathematically: include y if $s(x_{\text{test}}, y)$ would not be unusually large among the calibration scores

The nonconformity score is the key design choice: any model output can be converted into one

Split Conformal Prediction

The simplest algorithm: calibrate on a held-out split

1. Train model $\hat{f}$ on the training set (excluding calibration)
2. Compute nonconformity scores $s_i = |y_i - \hat{f}(x_i)|$ for all n calibration points
3. For coverage $1-\alpha$: compute the $(1-\alpha)(1+1/n)$ quantile $\hat{q}$ of $\{s_1, \ldots, s_n\}$
4. Prediction interval for a new point: $\hat{C}(x) = [\hat{f}(x) - \hat{q},\ \hat{f}(x) + \hat{q}]$



Split conformal: one pass, no retraining, exact finite-sample coverage guarantee

Why the Coverage Guarantee Holds

Exchangeability is the only assumption

- Vovk et al. (2005): if the calibration points and the new test point are exchangeable (a weaker condition than i.i.d.), the coverage guarantee holds exactly
- Exchangeability: the joint distribution is symmetric under permutation of indices
- This holds for i.i.d. data but also for some forms of time-series and spatial data
- The key insight: the rank of s_{test} among $\{s_1, \ldots, s_n, s_{\text{test}}\}$ is uniformly distributed under exchangeability
- No Gaussian assumption, no large-sample approximation: it works for $n = 20$

Conformal coverage is exact in finite samples – no CLT, no asymptotic approximation needed

Conditional vs Marginal Coverage

Marginal coverage is not enough for fair and reliable predictions

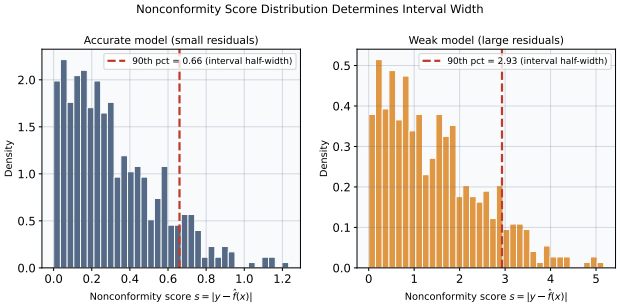
- **Marginal coverage**: $\mathbb{P}(Y \in \hat{C}(X)) \geq 1 - \alpha$ holds on average over all (X, Y) pairs – achievable distribution-free
- Problem: a model may have tight intervals for easy inputs and catastrophically wide intervals for hard ones; the average is still $1 - \alpha$ even if some subgroups are systematically under-covered
- **Conditional coverage**: $\mathbb{P}(Y \in \hat{C}(X) \mid X = x) \geq 1 - \alpha$ for every x; provably impossible without additional assumptions (Lei & Wasserman 2014)
- **Group conditional (Mondrian) conformal**: partition inputs into strata; apply split conformal within each stratum; achieves exact coverage per group
- Finance: marginal coverage may conceal that VaR intervals are systematically too narrow during high-volatility crises and too wide during calm periods

Mondrian conformal: stratify by regime or credit segment to approximate conditional coverage within groups

Nonconformity Score Distribution

The calibration scores determine how wide the intervals will be

- If the model is accurate: calibration scores are small and concentrated – tight intervals
- If the model is uncertain: calibration scores are large and dispersed – wide intervals
- The interval width is honest: a bad model produces wide intervals, not false precision
- Adaptive conformal prediction: use localised nonconformity scores that vary with x, producing narrower intervals where the model is confident

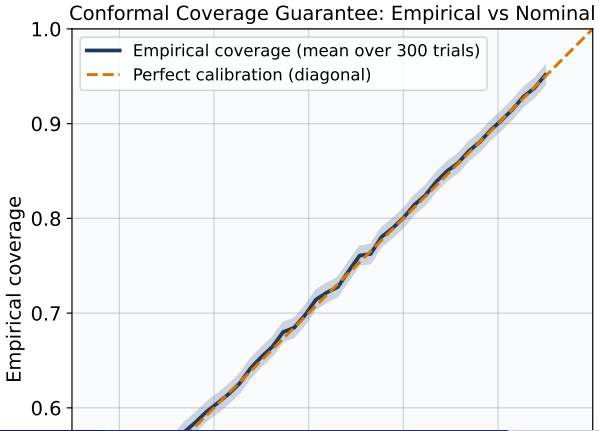


Wide conformal intervals are informative: they reveal where the model truly lacks confidence

Coverage Guarantee in Practice

Empirical coverage matches the nominal level across repeated experiments

- For any $\alpha \in (0, 1)$: split conformal produces intervals with *exactly* $1 - \alpha$ marginal coverage
- "Marginal" means: coverage holds on average over all test points (not necessarily conditional on x)
- Conditional coverage (tight intervals everywhere) requires stronger assumptions or adaptive methods
- On finite calibration sets: the guarantee is $\geq 1 - \alpha$ (slightly conservative by $1/n$ factor)



Conformal Classification: Prediction Sets

Instead of a single class, output a set of plausible classes

- Nonconformity score: $s(x, y) = 1 - \hat{p}(y \mid x)$ (lower probability = more nonconforming)
- Prediction set: all classes y with score below the calibration quantile $\hat{q}$: $\hat{C}(x) = \{y : \hat{p}(y \mid x) \geq 1 - \hat{q}\}$
- Under normal conditions: the set usually contains 1 class; under uncertainty, it contains more
- Set size is an automatic uncertainty signal: a large set flags ambiguous inputs
- Finance: conformal sets for credit rating changes include all plausible outcomes, not just the top-1 prediction

Conformal classification sets are coverage-guaranteed and interpretable as uncertainty signals

Conformal Anomaly Detection

A statistically valid test for novel or out-of-distribution inputs

- Standard anomaly detectors output raw scores with no probabilistic guarantee; the decision threshold is arbitrary
- **Conformal p-value** for a test point: the fraction of calibration scores at least as large as the test score $p = \big(|\{i : s_i \geq s_{\text{test}}\}| + 1\big)/(n+1)$
- Under the null (test point from the training distribution): p is uniformly distributed – classical one-sided hypothesis test
- **Multiple testing**: apply Benjamini-Hochberg to control the false discovery rate over a stream of test points
- Finance: detect unusual trade sequences (potential market manipulation), abrupt regime breaks, or data-feed anomalies with a provably controlled false-alarm rate

Conformal anomaly detection: the first statistically valid detector with a finite-sample false-positive-rate guarantee

Adaptive Conformal Prediction

Making intervals narrower where the model is confident

- Standard split conformal uses the same $\hat{q}$ for all inputs: intervals have constant width
- Adaptive methods: use a normalised nonconformity score $s(x, y) = |y - \hat{f}(x)|/\hat{\sigma}(x)$ where $\hat{\sigma}(x)$ is a local uncertainty estimate (from a second model or quantile regression)
- Conditional coverage approximation: intervals are tighter where the model is confident
- **CQR (Conformalized Quantile Regression):** uses quantile regression bounds as the base; conformal step corrects miscalibration of the quantile model
- Mondrian conformal: stratify by input region; exact conditional coverage within each stratum

CQR is the recommended adaptive method for regression: combines quantile regression + conformal

Cross-Conformal and Jackknife+

Better interval efficiency when labelled calibration data is limited

- Split conformal wastes data: the calibration split cannot be used for model training
- **Cross-conformal**: K-fold split; train K models, calibrate each on a different fold; aggregate nonconformity scores across folds – all data contributes to both training and calibration
- **Jackknife+** (Barber et al. 2021): leave-one-out residuals from a single model; provably valid marginal coverage without a separate calibration split; infeasible for large neural networks (requires n retraining runs)
- **CV+** (Barber et al. 2021): K-fold analogue of jackknife+; balances data efficiency and compute cost
- Finance: when labelled data is scarce (alternative data sources, rare credit default events), cross-conformal extracts more predictive value from each observation

Jackknife+ is optimal for small datasets; CV+ is the practical compromise for medium-scale models

Conformal Prediction for Time Series

Handling distribution shift and temporal dependence

- Standard conformal assumes exchangeability: violated under temporal autocorrelation and drift
- **Adaptive Conformal Inference** (Gibbs & Candes 2021): update $\hat{q}$ online using a simple step rule; provably achieves long-run coverage even under distribution shift
- **EnbPI** (Xu & Xie 2021): uses bootstrap residuals over a rolling window; valid for dependent time series under mild mixing conditions
- Finance: daily return prediction sets, updating the quantile as market regimes shift
- No need to detect regime changes: the adaptive mechanism adjusts automatically

Adaptive conformal inference is the right tool for financial time-series forecasting

Limitations of Conformal Prediction

Guarantees are marginal and depend on exchangeability

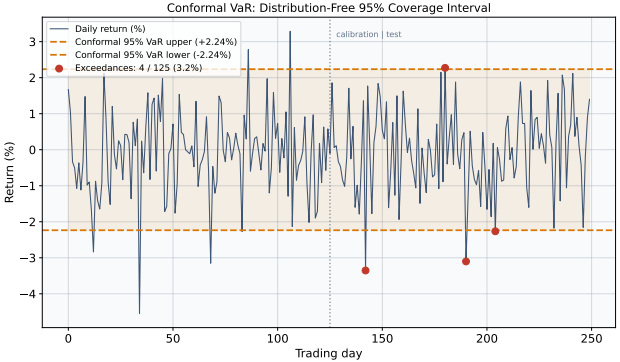
- Marginal coverage $\neq$ conditional coverage: intervals may be systematically too narrow for some subgroups even when the overall coverage is correct
- Exchangeability is not guaranteed for heavily autocorrelated financial time-series (e.g. high-frequency order flow)
- Computationally: full conformal (not split) requires refitting the model n times – only feasible for cheap models (kNN, ridge regression)
- The prediction set size is not controlled: a bad model produces intervals so wide they are uninformative – conformal does not fix model quality

Conformal prediction quantifies uncertainty honestly; it does not improve model accuracy

Finance Application: Value-at-Risk Forecasting

Distribution-free coverage guarantees for tail risk estimation

- VaR at $\alpha = 5\%$: the loss exceeded on at most 5% of days
- Traditional VaR: parametric (normal) or historical simulation; no coverage guarantee
- Conformal approach: use split conformal on standardised residuals of a GARCH model; the resulting interval is guaranteed to contain the true loss $\geq 95\%$ of the time
- Regulatory back-testing: Basel III requires ≤ 4 VaR exceedances per 250 days at 99%; conformal VaR satisfies this by construction



Code Pointer

Conformal prediction in Python

- **MAPIE** (`pip install mapie`): split conformal, CQR, and cross-conformal for regression and classification; scikit-learn compatible API
- **conformal-prediction** (Angelopoulos & Bates): reference implementations for classification, regression, and risk control
- **Nonconformist**: older library; implements inductive and transductive conformal prediction
- Minimal split conformal in 5 lines: sort calibration residuals, take $(1-\alpha)$ quantile, add and subtract from point prediction
- Finance: wrap any GARCH or ML return forecast with MAPIE's `MapieRegressor`

MAPIE is the standard library: sklearn-compatible, actively maintained, well-documented

Finance Application: Conformal Credit Scoring

Distribution-free prediction sets for regulatory-grade credit models

- Credit scoring: predict probability of default (PD); regulators require calibrated uncertainty estimates, not just point predictions
- Conformal approach: train a gradient boosting classifier; calibrate nonconformity scores on a held-out set stratified by credit segment
- **Mondrian conformal by credit grade**: guarantees $\geq 95\%$ coverage within each rating bucket, not just on average – prevents systematic under-coverage for high-risk obligors
- Output: a conformal prediction set of plausible credit ratings (e.g., {B, BB}) with a provably valid coverage rate per segment
- Basel III model validation: conformal back-testing provides a formal statistical test that the PD model is not systematically miscalibrated – replaces ad-hoc traffic-light tests

Mondrian conformal credit scoring: segment-specific coverage guarantees align with Basel internal model validation requirements

Summary and Open Questions

Three things to remember

- Conformal prediction produces prediction sets with a *guaranteed* marginal coverage $\mathbb{P}(Y \in \hat{C}(X)) \geq 1 - \alpha$ under exchangeability alone – no distributional assumptions
- Split conformal is one pass over a calibration set; intervals widen honestly when the model is uncertain; CQR gives adaptive width
- Finance: conformal VaR satisfies Basel back-testing requirements by construction; adaptive conformal handles time-series regime shifts

Open questions:

1. How do we achieve conditional (not just marginal) coverage without strong assumptions?
2. Can conformal prediction handle multi-step ahead forecasts with valid joint coverage?
3. How should conformal sets be communicated to non-expert financial users?

Further reading: Vovk et al. (2005) Algorithmic Learning; Angelopoulos & Bates (2023) tutorial