Tightening Hyperrectangular Conformal Prediction Regions for Multi-Target Regression under Non-exchangeability

Filip Schlembach FILIP@SCHLEMBACH.BE
Dennis Mertens Velasquez DENNIS.MERTENS@MAASTRICHTUNIVERSITY.NL
Evgueni Smirnov SMIRNOV@MAASTRICHTUNIVERSITY.NL
Mark H. M. Winands M.WINANDS@MAASTRICHTUNIVERSITY.NL
Department of Advanced Computing Sciences, Maastricht University, Maastricht, The Netherlands

Editors: Andy Song, Bo Han and Sarah Erfani

Abstract

Under exchangeability, conformal prediction equips multi-target regressors with prediction regions that contain the true label with at least a prescribed probability. Constructing these regions as axis-aligned hyperrectangles provides one prediction interval per output dimension, making the uncertainty estimates easy to interpret. Joint coverage is achieved by controlling the probability that at least one interval fails to contain the true value of its corresponding output, for example through Bonferroni correction. When exchangeability is violated, as in many time series, nominal coverage is no longer guaranteed, but reweighting the nonconformity scores allows the resulting coverage gap to be bounded. The weighted Bonferroni baseline nevertheless produces regions that are larger than necessary. It allocates the global significance level uniformly, ignoring differences in marginal score distributions, and it does not exploit dependence between scores across dimensions. To address these limitations, we extend the correction methods `DCP-MT` and `Max-Rank` through reweighting, obtaining `WDCP-MT` and `WMax-Rank`, respectively. `WDCP-MT` adapts the dimension-wise significance levels to the marginal score distributions, while `WMax-Rank` exploits positive dependence, under which the intervals tend to fail together. For each method, we derive a finite-sample upper bound on the coverage gap that vanishes under exchangeability. Experiments on synthetic and real-world time series reveal complementary strengths. `WMax-Rank` produces smaller regions than the weighted Bonferroni baseline in almost all settings, with particularly strong gains when the scores are positively correlated, whereas `WDCP-MT` does so when the scores are uncorrelated and their marginal distributions differ in shape.

Keywords: Inductive conformal prediction; multi-target regression; conformal time series forecasting; multiple hypothesis testing.

1. Introduction

Conformal prediction provides distribution-free uncertainty quantification through prediction regions with finite-sample coverage guarantees under exchangeability (Barber et al., 2023; Vovk et al., 2022). For multi-target regression, axis-aligned hyperrectangles offer an interpretable representation of joint uncertainty (Molnar, 2025), since they are Cartesian products of prediction intervals, one per output dimension. This structure has motivated recent interest in hyperrectangular prediction regions for multi-target regression and time series (Fan and Sesia, 2025; Schlembach et al., 2025b).

Achieving joint coverage for such regions requires controlling the probability that at least one output falls outside its interval. Under exchangeability, methods providing this

© 2026 F. Schlembach, D. Mertens Velasquez, E. Smirnov & M.H.M. Winands.

control fall into two broad classes. The first class is based on Boole's inequality. It includes Bonferroni correction, which distributes the global significance level uniformly across dimensions (Vovk, 2013), and `DCP-MT`, which allocates it adaptively to reduce region volume (Schlembach et al., 2025b). The second class consists of dependence-aware methods, including `Max-Rank` (Timans et al., 2025) and copula-based approaches (Messoudi et al., 2021). These methods exploit dependence between dimension-wise nonconformity scores and can produce smaller regions when outputs tend to fall outside their intervals together.

These coverage guarantees may fail when exchangeability is violated (Barber et al., 2023; Stocker et al., 2025). The reweighting framework of Barber et al. (2023) addresses this challenge by providing an explicit upper bound on the resulting coverage gap. Schlembach et al. (2025a) combine this framework with Bonferroni correction to obtain non-exchangeable multi-target conformal prediction (`nmtCP`) for multistep-ahead forecasting. Although reweighting accommodates non-exchangeability, `nmtCP` still produces regions that are larger than necessary, for two reasons. First, uniform significance-level allocation ignores differences in marginal score distributions, and second, Boole's inequality leaves cross-dimensional dependence unused.

We address these limitations through the following contributions.

- We introduce the weighted correction methods `WDCP-MT` and `WMax-Rank` by incorporating reweighting into `DCP-MT` and `Max-Rank`, respectively. Following Schlembach et al. (2025b) and Timans et al. (2025), we expect `WDCP-MT` to offer its greatest gains when marginal score distributions differ in shape and the scores are weakly correlated, and `WMax-Rank` when they are positively correlated.
- We derive a coverage-gap bound for weighted correction methods based on Boole's inequality. The result covers both fixed allocations and allocations selected on a separate calibration subset, as in `WDCP-MT`.
- We derive a separate coverage-gap bound for `WMax-Rank` through a comparison with its transductive counterpart.
- We provide implementations of both methods and evaluate their coverage and region size on synthetic and real-world time series. The experiments support the expected operating regimes and identify the settings in which each method improves informational efficiency over `nmtCP`.

The remainder of this paper is organised as follows. Section 2 reviews multi-target conformal prediction and the reweighting framework for non-exchangeable data. Sections 3 and 4 introduce `WDCP-MT` and `WMax-Rank`, respectively, and derive their coverage-gap bounds. Section 5 presents the experimental setup and results, followed by a discussion of their theoretical and practical implications in Section 6. Section 7 concludes with directions for future research. Complete proofs, implementation details, and additional experimental results are provided in the supplementary materials.

2. Background

This section provides the background on multi-target regression, hyperrectangular inductive conformal prediction, and conformal prediction for non-exchangeable data.

2.1. Uncertainty Quantification for Multi-target Regression

Let $\mathcal{X} \subseteq \mathbb{R}^p$ be the *input space* and $\mathcal{Y} \subseteq \mathbb{R}^m$ be the *output space*. Suppose an underlying multi-target regression model $\hat{\mathbf{f}} : \mathcal{X} \to \mathcal{Y}$ has already been learned and is used to predict m output variables from p input variables. In addition, let $\mathcal{D}_{cal} = \{(\mathbf{X}^{(i)}, \mathbf{Y}^{(i)})\}_{i=1}^{n}$ be a calibration data set consisting of n labelled instances $(\mathbf{X}^{(i)}, \mathbf{Y}^{(i)})$ not used for the training of $\hat{\mathbf{f}}$. $\mathbf{X}^{(i)} = (X_1^{(i)}, \ldots, X_p^{(i)})$ and $\mathbf{Y}^{(i)} = (Y_1^{(i)}, \ldots, Y_m^{(i)})$ are random vectors with values from $\mathcal{X}$ and $\mathcal{Y}$, respectively. Given a new test instance $\mathbf{X}^{(n+1)}$, the goal is to quantify the uncertainty of prediction $\hat{\mathbf{Y}}^{(n+1)} = \hat{\mathbf{f}}(\mathbf{X}^{(n+1)})$ with respect to the unknown true label $\mathbf{Y}^{(n+1)}$.

2.2. Inductive Conformal Prediction and Exchangeable Data

Inductive Conformal Prediction (ICP) (Vovk et al., 2022) provides distribution-free uncertainty quantification using prediction regions. Its finite-sample validity guarantee states that prediction region $\Gamma^{\epsilon}(\mathbf{X}^{(n+1)}) \subseteq \mathcal{Y}$ for test instance $\mathbf{X}^{(n+1)}$ satisfies

$$\mathsf{P}\left(\mathbf{Y}^{(n+1)} \in \Gamma^{\epsilon}\left(\mathbf{X}^{(n+1)}\right)\right) \geq 1 - \epsilon, \tag{1}$$

i.e., the true m-dimensional label $\mathbf{Y}^{(n+1)}$ is included in $\Gamma^{\epsilon}(\mathbf{X}^{(n+1)})$ with probability at least $1 - \epsilon$, where ϵ is the significance level.

The validity guarantee from Equation (1) holds for ICP under the assumption that the instances in the calibration data set $\mathcal{D}_{cal}$ and test instance $(\mathbf{X}^{(n+1)}, \mathbf{Y}^{(n+1)})$ are *exchangeable* (Barber et al., 2023). Formally, this means that for every permutation π of $\{1, \ldots, n+1\}$

$$\left((\mathbf{X}^{(i)}, \mathbf{Y}^{(i)})\right)_{i=1}^{n+1} \overset{d}{=} \left((\mathbf{X}^{(\pi(i))}, \mathbf{Y}^{(\pi(i))})\right)_{i=1}^{n+1} \tag{2}$$

holds, where $\overset{d}{=}$ denotes equality in distribution.

In this paper, ICP is used in a *multi-test setting* (Schlembach et al., 2025b), which means that the decision to include an m-dimensional label $\mathbf{y} \in \mathcal{Y}$ in the prediction region $\Gamma^{\epsilon}(\mathbf{X}^{(n+1)})$ is based on m dimension-wise conformity tests. This yields a dimension-wise prediction region $\Gamma_j^{\epsilon_j}(\mathbf{X}^{(n+1)}) \subseteq \mathcal{Y}_j$ for each dimension $j = 1, \ldots, m$. Thus, a candidate label $\mathbf{y} = (y_1, \ldots, y_m) \in \mathcal{Y}$ is included in the (global) prediction region $\Gamma^{\epsilon}(\mathbf{X}^{(n+1)})$ iff $y_j \in \Gamma_j^{\epsilon_j}(\mathbf{X}^{(n+1)})$ for each dimension j. This is equivalent to

$$\Gamma^{\epsilon}(\mathbf{X}^{(n+1)}) = \bigtimes_{j=1}^{m} \Gamma_j^{\epsilon_j}(\mathbf{X}^{(n+1)}). \tag{3}$$

To construct dimension-wise prediction regions $\Gamma_j^{\epsilon_j}(\mathbf{x}) \subseteq \mathcal{Y}_j$, ICP employs a dimension-wise nonconformity score function $\mathsf{S}_{\mathtt{dim}}$ using the underlying model $\hat{\mathbf{f}}$. Its component $\mathsf{S}_{\mathtt{dim},j}$ assesses how unusual the labelled pair $(\mathbf{X}^{(i)}, Y_j^{(i)})$ in the calibration set $\mathcal{D}_{cal}$ is relative to the j-th component $\hat{f}_j$ of the model $\hat{\mathbf{f}}$. Formally, $\mathsf{S}_{\mathtt{dim}}$ is a measurable mapping $\mathsf{S}_{\mathtt{dim}} : (\mathcal{F} \times \mathcal{X} \times \mathcal{Y}) \longrightarrow \bar{\mathbb{R}}^m$, where $\mathcal{F}$ is the domain of all models trained on a proper training set $\mathcal{D}_{tr}$, and $\mathsf{S}_{\mathtt{dim},j}$ depends on the label only through y_j (Vovk et al., 2022; Schlembach et al., 2026). Omitting the model, we write $\mathsf{S}_{\mathtt{dim},j}(\mathbf{x}, y_j)$.

ICP computes the dimension-wise nonconformity scores $B_j^{(i)} = \mathsf{S}_{\mathtt{dim},j}(\mathbf{X}^{(i)}, Y_j^{(i)})$ for the n instances in $\mathcal{D}_{cal}$. Then, for a new test instance $\mathbf{X}^{(n+1)}$, it constructs the dimension-wise prediction region in dimension j,

$$\Gamma_j^{\epsilon_j}\left(\mathbf{X}^{(n+1)}\right) = \left\{y_j \in \mathcal{Y}_j : \mathsf{S}_{\mathtt{dim},j}\left(\mathbf{X}^{(n+1)}, y_j\right) \leq q_j^{(1-\epsilon_j)}\right\}, \tag{4}$$

where $q_j^{(1-\epsilon_j)} = \mathsf{Q}_{1-\epsilon_j}(\{B_j^{(i)}\}_{i=1}^n)$ is the $(1-\epsilon_j)$-quantile of the calibration nonconformity scores in dimension j. For n scores, $\mathsf{Q}_{1-\epsilon}$ returns the $\lceil (1-\epsilon)(n+1)\rceil$th smallest one, or $+\infty$ if $\lceil (1-\epsilon)(n+1)\rceil > n$. The final prediction region $\Gamma^{\epsilon}(\mathbf{X}^{(n+1)})$ is computed according to Equation (3).

If the absolute residuals $\mathsf{S}_{\mathtt{dim},j}(\mathbf{x}, y_j) = |y_j - \hat{f}_j(\mathbf{x})|$ are chosen as the nonconformity score function in dimension j then the dimension-wise prediction region $\Gamma_j^{\epsilon_j}(\mathbf{X}^{(n+1)})$ simplifies to the prediction interval $\Gamma_j^{\epsilon_j}(\mathbf{X}^{(n+1)}) = \left[\hat{f}_j(\mathbf{X}^{(n+1)}) - q_j^{(1-\epsilon_j)},\ \hat{f}_j(\mathbf{X}^{(n+1)}) + q_j^{(1-\epsilon_j)}\right]$. We note that if dimension-wise prediction regions $\Gamma_j^{\epsilon_j}(\mathbf{X}^{(n+1)})$ are intervals, the prediction region $\Gamma^{\epsilon}(\mathbf{X}^{(n+1)})$ is a *hyperrectangle* in $\mathcal{Y} \subseteq \mathbb{R}^m$. Its sides are perpendicular to the coordinate axes of the output space which benefits interpretability.

To ensure the conformal validity guarantee from Equation (1) holds for the (global) prediction region $\Gamma^{\epsilon}(\mathbf{X}^{(n+1)})$, ICP needs a correction method to combine the dimension-wise prediction regions $\Gamma_j^{\epsilon_j}(\mathbf{X}^{(n+1)})$. Existing correction methods are divided into two classes.

The first class consists of *correction methods based on Boole's inequality*, which choose the dimension-wise significance levels ϵ_j such that

$$\sum_{j=1}^{m} \epsilon_j \leq \epsilon. \tag{5}$$

• *Bonferroni correction* (Vovk, 2013) sets $\epsilon_j = \epsilon/m$ for all output dimensions $j = 1, \ldots, m$. When used with ICP, it results in a fast inductive conformal predictor (`Bonferroni CP`) that generally produces overly large prediction regions.

• *Dynamic Bonferroni correction* (Schlembach et al., 2025b) generalises the Bonferroni correction by choosing the dimension-wise significance levels ϵ_j adaptively under the constraint in Equation (5) so as to minimise the volume of the global prediction region $\Gamma^{\epsilon}(\mathbf{X}^{(n+1)})$. The resulting dynamic conformal predictor for multi-target regression (`DCP-MT`) is shown to produce smaller prediction regions when the nonconformity scores have different marginal distributions and/or are only weakly correlated across dimensions.

The second class consists of *dependence-aware correction methods* that do not rely on Equation (5), but instead account more directly for dependence across dimensions.

• *Max-Rank correction* was proposed by Timans et al. (2025) for settings in which the nonconformity scores are positively correlated across output dimensions. For each output dimension j, it first defines the rank[1] $R_{\mathtt{ICP},j}^{(i)} = \mathrm{rank}(B_j^{(i)}, \{B_j^{(k)}\}_{k=1}^n)$, and then the maximum-rank $A_{\mathtt{ICP},\mathtt{MR}}^{(i)} = \max_{j=1}^m R_{\mathtt{ICP},j}^{(i)}$ for each calibration instance $(\mathbf{X}^{(i)}, \mathbf{Y}^{(i)})$. The $(1-\epsilon)$-quantile $q_{\mathtt{ICP}}^{(1-\epsilon)} = \mathsf{Q}_{1-\epsilon}(\{A_{\mathtt{ICP},\mathtt{MR}}^{(i)}\}_{i=1}^n)$ of the maximum-rank scores induces m dimension-wise prediction regions $\Gamma_{\mathtt{ICP},j}^{\epsilon_j}(\mathbf{X}^{(n+1)})$, where the effective dimension-wise significance levels

1. $\mathrm{rank}\left(v, \left\{b^{(k)}\right\}\right) = \left|\left\{k : b^{(k)} \leq v\right\}\right|$ is the number of values in $\left\{b^{(k)}\right\}$ at or below v.

$\epsilon_j = 1 - (q_{\texttt{ICP}}^{(1-\epsilon)} + 1)/(n+1), \quad j = 1, \ldots, m$ are identical across dimensions (Schlembach et al., 2026). The resulting global prediction region $\Gamma^{\epsilon}_{\texttt{ICP},\texttt{MR}}(\mathbf{X}^{(n+1)})$ is then obtained from Equation (3). Timans et al. (2025) and Schlembach et al. (2026) prove that this prediction region is valid even though the induced dimension-wise significance levels do not need to satisfy Equation (5).

Another dependence-aware correction method is based on copulas (Messoudi et al., 2021). Since `Max-Rank` can be viewed as a fast implementation of inductive conformal predictors based on the empirical copula, we do not discuss it further in this paper.

2.3. Non-exchangeability and Conformal Prediction

If the exchangeability assumption from Equation (2) is violated, applying ICP poses a challenge since the validity guarantee from Equation (1) no longer holds. Time series forecasting, for example, is a task that commonly involves non-exchangeable data.

Several techniques have been proposed to compensate for the lack of exchangeability. Stocker et al. (2025) classify them broadly into four overlapping strategies. The *adaptive* strategy introduced by Gibbs and Candès (2021) adapts the significance level ϵ over time to achieve the desired error rate. The *blocking* strategy clusters consecutive instances into "blocks". While the instances within a block are not exchangeable, blocking strategies assume that the blocks are (Chernozhukov et al., 2018; Stankevičiūtė et al., 2021). The *refreshing* strategy used by Xu and Xie (2021) continuously adds new instances to the calibration set as soon as the true label becomes available and removes the oldest instance from the calibration set for each new one added. Finally, the *reweighting* strategy, which we will focus on in this paper, attempts to approximate the distribution of the nonconformity scores at any instance by assigning weights $w^{(i)}$ to the nonconformity scores computed over the calibration set.

Consider scalar nonconformity scores $A^{(1)}, \ldots, A^{(n)}$ of the calibration instances and $A^{(n+1)}$ of the test instance, such as the dimension-wise scores $B_j^{(i)}$ of a single dimension j (Section 3) or the maximum ranks of `Max-Rank` (Section 4). Let $w^{(1)}, \ldots, w^{(n+1)}$ be a set of non-negative weights assigned to the n calibration scores and to the point mass at infinity, such that $w^{(n+1)} \geq w^{(i)}, \ \forall i \in \{1, \ldots, n\}$. The *reweighting* strategy, introduced by Tibshirani et al. (2019) for covariate shift and by Barber et al. (2023) for arbitrary violations of exchangeability, then works by employing a weighted quantile function

$$q^{(1-\epsilon)} = \mathsf{Q}_{1-\epsilon}\left(\sum_{i=1}^{n} \tilde{w}^{(i)} \cdot \delta_{A^{(i)}} + \tilde{w}^{(n+1)} \cdot \delta_{+\infty}\right) \tag{6}$$

where $\tilde{w}^{(i)} = w^{(i)} / \sum_{k=1}^{n+1} w^{(k)}$ are normalised weights, and $\delta_{A^{(i)}}$ is the Dirac point mass at $A^{(i)}$. When using this weighted quantile function, Barber et al. (2023) proved that the *coverage gap*, i.e., the excess fraction of true labels not covered by the prediction region due to the loss of exchangeability, satisfies

$$(1-\epsilon) - \mathsf{P}\left(\mathbf{Y}^{(n+1)} \in \Gamma^{\epsilon}\left(\mathbf{X}^{(n+1)}\right)\right) \leq \sum_{i=1}^{n} \tilde{w}^{(i)} \mathsf{d}_{\mathsf{TV}}\left(\mathbf{A}, \mathbf{A}^{\rightleftarrows i}\right), \tag{7}$$

where $\mathsf{d_{TV}}$ is the total variation distance between the distributions of its arguments, $\mathbf{A} = (A^{(1)}, \ldots, A^{(n+1)})$ is the vector of random variables of the nonconformity scores over the calibration set $\mathcal{D}_{cal}$ and the test instance, and $\mathbf{A}^{\rightleftarrows i} = (A^{(1)}, \ldots, A^{(i-1)}, A^{(n+1)}, A^{(i+1)}, \ldots, A^{(n)}, A^{(i)})$ is the same vector with the ith and the $(n+1)$st elements swapped. The bound holds for any fixed underlying model, so its probabilities and total variation distances are conditional on the proper training set $\mathcal{D}_{tr}$ on which $\hat{\mathbf{f}}$ is fitted.

Schlembach et al. (2025a) combine the reweighting strategy of Barber et al. (2023) with the Bonferroni correction. The resulting method named non-exchangeable multi-target conformal prediction (`nmtCP`) obtains hyperrectangular prediction regions for multistep-ahead multi-target time series. An upper bound on the coverage gap analogous to Equation (7) for `nmtCP` is derived. The experiments show that, although `nmtCP` adapts to distributional changes, it is often too conservative and produces prediction regions that are larger than necessary.

To obtain tighter hyperrectangular prediction regions, we combine the reweighting strategy of Barber et al. (2023) with `DCP-MT` and `Max-Rank` in Sections 3 and 4.

3. Weighted Dynamic Conformal Prediction for Multi-Target Regression

Schlembach et al. (2025b) show that `DCP-MT` can produce smaller prediction regions than `Bonferroni CP` (Vovk, 2013) under exchangeability. We extend it to the non-exchangeable setting with the reweighting strategy of Barber et al. (2023) and call the result *Weighted Dynamic Conformal Prediction for Multi-Target Regression* (`WDCP-MT`).

3.1. Implementation of Weighted DCP-MT

The `WDCP-MT` method is given in Algorithm 1. Selecting the dimension-wise significance levels ϵ_j on the same scores used to form the prediction region biases the region towards being too small. `WDCP-MT` therefore randomly partitions the calibration scores into a selection set E, used to select the dimension-wise significance levels, and a disjoint set O, used to form the prediction region. `WDCP-MT` renormalises the weights separately over each set. The levels ϵ_j thus depend on the scores only through E, as (A4) of Theorem 1 requires. The partition is drawn uniformly at random rather than deterministically (e.g., by index parity) because a systematic split is vulnerable to aliasing. If the scores retain a periodic component corresponding to the sampling interval of a systematic split, the two halves no longer share the same marginal distribution, biasing the weighted quantile.

For each output dimension j, `WDCP-MT` sorts the scores in E and computes the candidate dimension-wise significance level $\epsilon_j^{(k)}$ associated with each rank k. It then selects one level $\epsilon_j = \epsilon_j^{(k_j)}$ per dimension that minimises the sum of the logarithms of the dimension-wise thresholds $q_j^{(1-\epsilon_j)}$ estimated from the scores in E, subject to $\sum_j \epsilon_j \leq \epsilon$, which the validity requires, and to $\epsilon_j \geq \tilde{w}_O^{(n+1)}$, because smaller levels would yield unbounded intervals on O. If no such allocation exists, $\Gamma^{\epsilon}(\mathbf{x}^{(n+1)})$ is set to $\mathcal{Y}$. With the dimension-wise significance levels fixed, the weighted quantile function is applied to the scores in O at level $1-\epsilon_j$ to obtain the threshold $q_j^{(1-\epsilon_j)}$ for each dimension j. Together with the point predictions $\hat{f}_j(\mathbf{x}^{(n+1)})$, these thresholds determine the dimension-wise prediction intervals $\Gamma_j^{\epsilon_j}(\mathbf{x}^{(n+1)})$

Algorithm 1: `WDCP-MT`

Input: test instance $\mathbf{x}^{(n+1)}$; predictions $\hat{f}_j(\mathbf{x}^{(n+1)})$ for $j = 1, \ldots, m$; nonconformity scores $\{\beta_j^{(i)}\}_{i=1}^{n}$; weights $\{w^{(i)}\}_{i=1}^{n+1}$; global significance level ϵ

Output: prediction region $\Gamma^{\epsilon}(\mathbf{x}^{(n+1)})$

`// balanced random split`

$E \leftarrow$ uniformly random subset of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$; $O \leftarrow \{1, \ldots, n\} \setminus E$;

$\tilde{w}_S^{(i)} \leftarrow w^{(i)} / (w^{(n+1)} + \sum_{k \in S} w^{(k)})$ for each $S \in \{E, O\}$, $i \in S \cup \{n+1\}$;

`// Stage 1: select the dimension-wise significance levels on` E

for $j = 1, \ldots, m$ **do**

  Sort $\{\beta_j^{(i)}\}_{i \in E}$ in ascending order using permutation π_j;

  for $k = 1, \ldots, |E|$ **do**

    $\epsilon_j^{(k)} \leftarrow 1 - \sum_{l=1}^{k} \tilde{w}_E^{(\pi_j^{(l)})}$;

  end

end

Select ranks $k_1, \ldots, k_m$ that minimise $\sum_{j=1}^{m} \log \max\big(\beta_j^{(\pi_j^{(k_j)})}, 10^{-12}\big)$ s.t. $\sum_{j=1}^{m} \epsilon_j^{(k_j)} \leq \epsilon$ and $\epsilon_j^{(k_j)} \geq \tilde{w}_O^{(n+1)}$ for all j, by dynamic programming (Algorithm S2 in the supplementary materials);

if *no such ranks exist* **then**

  return $\Gamma^{\epsilon}(\mathbf{x}^{(n+1)}) = \mathcal{Y}$;

else

  `// Stage 2: form the region from` O `at the selected` ϵ_j

  for $j = 1, \ldots, m$ **do**

    $\epsilon_j \leftarrow \epsilon_j^{(k_j)}$; $q_j^{(1-\epsilon_j)} \leftarrow \mathsf{Q}_{1-\epsilon_j}\Big(\sum_{i \in O} \tilde{w}_O^{(i)} \delta_{\beta_j^{(i)}} + \tilde{w}_O^{(n+1)} \delta_{+\infty}\Big)$;

    $\Gamma_j^{\epsilon_j}(\mathbf{x}^{(n+1)}) \leftarrow \Big\{ y_j \in \mathcal{Y}_j : \mathsf{S}_{\mathtt{dim},j}(\mathbf{x}^{(n+1)}, y_j) \leq q_j^{(1-\epsilon_j)} \Big\}$;

  end

  return $\Gamma^{\epsilon}(\mathbf{x}^{(n+1)}) = \bigtimes_{j=1}^{m} \Gamma_j^{\epsilon_j}(\mathbf{x}^{(n+1)})$;

end

and the prediction region $\Gamma^{\epsilon}(\mathbf{x}^{(n+1)}) = \bigtimes_{j=1}^{m} \Gamma_j^{\epsilon_j}(\mathbf{x}^{(n+1)})$. The split makes the coverage-gap bound of Theorem 1 applicable but has two costs. First, it approximately halves the effective calibration set size. Second, the selected dimension-wise significance levels are optimal for the scores in E, where they never do worse than the Bonferroni correction (Proposition S10 in the supplementary materials), but not necessarily for those in O.

Schlembach et al. (2025b) present the selection of the dimension-wise significance levels ϵ_j as a budget allocation problem and solve it with a mixed-integer linear program, whose worst-case running time is exponential. `WDCP-MT` solves it as a multiple-choice knapsack problem by dynamic programming in polynomial time, up to a negligible loss of budget (Algorithm S2 in the supplementary materials). Complexity and runtimes for all methods are compared in Sections D.3 and D.4 of the supplementary materials. Both solvers need

the logarithm to make the objective additive over the dimensions. For absolute residuals, as in our experiments, the interval in dimension j has width $2q_j^{(1-\epsilon_j)}$, so the objective is the logarithm of the estimated volume up to a constant and has the same minimiser as the volume. The choice of solver affects only the runtime and, slightly, the size of the prediction region, not the coverage-gap bound of Theorem 1.

3.2. Coverage Gap of Weighted DCP-MT

Theorem 1 bounds the coverage gap of every multi-test conformal predictor whose dimension-wise significance levels satisfy Boole's inequality when the exchangeability assumption from Equation (2) is violated. It covers dimension-wise significance levels fixed in advance, as in `Bonferroni CP` and `nmtCP`, as well as dimension-wise significance levels selected on part of the calibration set, as in `WDCP-MT`. To treat both cases together, a split $\sigma = (E, O)$ partitions the calibration indices into a selection set E, on which the dimension-wise significance levels are selected, and a disjoint set O, from which the prediction region is formed, with $E \cup O = \{1, \dots, n\}$. If the dimension-wise significance levels are fixed in advance, $E = \emptyset$. Let $\mathbf{B}^{(i)} = (B_1^{(i)}, \dots, B_m^{(i)})$ be the dimension-wise nonconformity scores of instance i and $\mathbf{B}_E = (\mathbf{B}^{(i)})_{i \in E}$ the scores in E. For each dimension j, let $\mathbf{V}_j = (B_j^{(i)})_{i \in O \cup \{n+1\}}$ be the scores in O together with the test score, and let $\mathbf{V}_j^{\rightleftarrows i}$ be $\mathbf{V}_j$ with the entries of $i \in O$ and $n+1$ swapped. The weights normalised over O are $\tilde{w}_O^{(i)} = w^{(i)} / (w^{(n+1)} + \sum_{k \in O} w^{(k)})$ for $i \in O \cup \{n+1\}$.

Theorem 1 *Let $\Gamma^\epsilon(\mathbf{X}^{(n+1)})$ be the hyperrectangular prediction region from Equation (3), where $q_j^{(1-\epsilon_j)}$ in Equation (4) is the weighted quantile from Equation (6) of the scores in O with the weights $\tilde{w}_O^{(i)}$, and assume the following.*

(A1) $\hat{\mathbf{f}}$ is fitted on a proper training set $\mathcal{D}_{tr}$ disjoint from the calibration and test instances, and all probabilities and total variation distances are conditional on $\mathcal{D}_{tr}$.

(A2) The weights are fixed in advance and satisfy $w^{(n+1)} > 0$ and $w^{(i)} \leq w^{(n+1)}$ for $i = 1, \dots, n$, i.e., the test instance receives the largest weight.

(A3) The split σ is drawn independently of the scores (e.g., at random, in blocks, or $E = \emptyset$).

(A4) The dimension-wise significance levels ϵ_j are functions of σ and $\mathbf{B}_E$ only and satisfy $\sum_{j=1}^m \epsilon_j \leq \epsilon$. If no such allocation exists, $\Gamma^\epsilon(\mathbf{X}^{(n+1)}) = \mathcal{Y}$.

Then the coverage gap satisfies

$$(1 - \epsilon) - \mathsf{P}\left(\mathbf{Y}^{(n+1)} \in \Gamma^\epsilon\left(\mathbf{X}^{(n+1)}\right)\right) \leq \mathsf{E}_\sigma\left[\sum_{j=1}^m \sum_{i \in O} \tilde{w}_O^{(i)} \Delta_{ij}(\sigma)\right], \tag{8}$$

where $\Delta_{ij}(\sigma) = \mathsf{d_{TV}}\left((\mathbf{B}_E, \mathbf{V}_j), (\mathbf{B}_E, \mathbf{V}_j^{\rightleftarrows i})\right)$ is evaluated for the fixed index sets E and O of the split σ, since, by (A3), fixing the split does not change the distribution of the scores. The bound simplifies in the following three cases.

(i) Fixed dimension-wise significance levels *($E = \emptyset$, e.g., `Bonferroni CP` and `nmtCP`):* $\Delta_{ij} = \mathsf{d_{TV}}\left(\mathbf{B}_j, \mathbf{B}_j^{\rightleftarrows i}\right)$ *with* $\mathbf{B}_j = (B_j^{(i)})_{i=1}^{n+1}$, *and the bound equals*

$$\sum_{j=1}^{m} \frac{\sum_{i=1}^{n} w^{(i)} \mathsf{d_{TV}}\left(\mathbf{B}_j, \mathbf{B}_j^{\rightleftarrows i}\right)}{\sum_{k=1}^{n+1} w^{(k)}}. \tag{9}$$

(ii) Dimension-wise significance levels selected on E *(e.g., `WDCP-MT`):* $\mathsf{d_{TV}}\left(\mathbf{V}_j, \mathbf{V}_j^{\rightleftarrows i}\right) \leq \Delta_{ij}(\sigma) \leq \mathsf{d_{TV}}\left(\mathbf{B}_{1:m}, \mathbf{B}_{1:m}^{\rightleftarrows i}\right)$ *with* $\mathbf{B}_{1:m} = (\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(n+1)})$, *and the lower bound holds with equality if* $\mathbf{B}_E$ *is independent of* $\mathbf{V}_j$.

(iii) Exchangeable scores $\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(n+1)}$*:* $\Delta_{ij}(\sigma) = 0$, *so the coverage is at least* $1 - \epsilon$.

Proof *(Sketch.)* Given σ and $\mathbf{B}_E$, the dimension-wise significance levels are fixed by (A4), so in each dimension j the weighted split conformal bound of Barber et al. (2023, Theorem 2), which requires (A2), applies to $\mathbf{V}_j$ under its conditional distribution. Boole's inequality over the m dimensions and $\sum_j \epsilon_j \leq \epsilon$ then bound the conditional coverage gap by conditional total variation distances. Their expectation over $\mathbf{B}_E$ is $\Delta_{ij}(\sigma)$, because the swap does not affect $\mathbf{B}_E$. The supplementary materials prove Theorem S2, which contains Theorem 1. ■

Theorem 1 does not apply if the dimension-wise significance levels also depend on the scores in O, as in Schlembach et al. (2025b), or on the test score, because fixing $\mathbf{B}_E$ then does not fix the ϵ_j. Case (ii) shows the price of selecting the dimension-wise significance levels from data. Under serial dependence between the scores in E and O, the fit of the selection to noise in E carries over to O, and $\Delta_{ij}(\sigma)$ can exceed the distance on $\mathbf{V}_j$ alone.[2]

4. Weighted Max-Rank

Timans et al. (2025) show that `Max-Rank` can produce smaller prediction regions than the Bonferroni correction (Vovk, 2013), and Schlembach et al. (2026) show that its corrected implementation is valid under exchangeability. We extend `Max-Rank` to the non-exchangeable setting with the reweighting strategy of Barber et al. (2023) and call the result *Weighted Max-Rank* (`WMax-Rank`).

4.1. Implementation of Weighted Max-Rank

The `WMax-Rank` predictor is given in Algorithm S1 in the supplementary materials. The input includes test instance $\mathbf{x}^{(n+1)}$, the point predictions $\hat{f}_j(\mathbf{x}^{(n+1)})$ for output dimensions $j = 1, \ldots, m$, the calibration nonconformity scores $\left\{\beta_j^{(i)}\right\}_{i=1}^{n}$, the normalised weights $\left\{\tilde{w}^{(i)}\right\}_{i=1}^{n+1}$, and the global significance level ϵ. For each dimension j, `WMax-Rank` first sorts the dimension-wise nonconformity scores, so that $\pi_j^{(k)}$ is the index of the kth smallest score,

2. See Remark S5 in the supplementary materials.

and computes the corresponding rank $\rho^{(i)}_{\texttt{ICP},j}$ for each calibration instance $i = 1, \ldots, n$. It then computes the maximum rank $\alpha^{(i)}_{\texttt{ICP},\texttt{MR}} = \max_{j=1}^{m} \rho^{(i)}_{\texttt{ICP},j}$ for each calibration instance and obtains the common threshold $q^{(1-\epsilon)}_{\texttt{ICP}}$ from the weighted quantile function of Equation (6), with the maximum ranks as scores and the point mass $\tilde{w}^{(n+1)}$ at $+\infty$.

If the weighted quantile function returns $+\infty$ or the threshold $q^{(1-\epsilon)}_{\texttt{ICP}}$ equals n, `WMax-Rank` outputs $\Gamma^{\epsilon}_{\texttt{ICP},\texttt{WMR}}(\mathbf{x}^{(n+1)}) = \mathcal{Y}$. Otherwise, it bounds the dimension-wise nonconformity scores of each output dimension j by the $(q^{(1-\epsilon)}_{\texttt{ICP}}+1)$st smallest calibration score $\beta_j^{(\pi_j^{(q^{(1-\epsilon)}_{\texttt{ICP}}+1)})}$, as `Max-Rank` does (Schlembach et al., 2026). Together with the point predictions $\hat{f}_j(\mathbf{x}^{(n+1)})$, these thresholds determine the dimension-wise prediction intervals $\Gamma^{\epsilon_j}_j(\mathbf{x}^{(n+1)})$, and `WMax-Rank` outputs the hyperrectangular prediction region $\Gamma^{\epsilon}_{\texttt{ICP},\texttt{WMR}}(\mathbf{x}^{(n+1)}) = \bigtimes_{j=1}^{m} \Gamma^{\epsilon_j}_j(\mathbf{x}^{(n+1)})$. For positive weights, the thresholds correspond to the dimension-wise significance levels $\epsilon_j = 1 - \sum_{k=1}^{q^{(1-\epsilon)}_{\texttt{ICP}}+1} \tilde{w}^{(\pi_j^{(k)})}$, which, unlike those of `Max-Rank`, might differ across dimensions.

4.2. Coverage Gap of Weighted Max-Rank

Theorem 2 *Let*

$$\Gamma^{\epsilon}_{ICP,\,WMR}(\mathbf{X}^{(n+1)}) = \left\{ \mathbf{y} \in \mathcal{Y} : \mathsf{S}_{ICP,\,MR}(\mathbf{X}^{(n+1)}, \mathbf{y}) \leq q^{(1-\epsilon)}_{ICP} \right\} \tag{10}$$

be the ***WMax-Rank*** *prediction region using the weighted threshold $q^{(1-\epsilon)}_{ICP}$ of Section 4.1, where $\mathsf{S}_{ICP,\,MR}(\mathbf{X}^{(n+1)}, \mathbf{y}) = \max_{j=1}^{m} \operatorname{rank}(\mathsf{S}_{dim,j}(\mathbf{X}^{(n+1)}, y_j), \{B_j^{(k)}\}_{k=1}^{n})$ is the maximum rank of the test instance with candidate label $\mathbf{y}$, whose dimension-wise scores are ranked against those of the calibration instances. Assume (A1) and (A2) of Theorem 1. Then*

$$(1-\epsilon) - \mathsf{P}\left(\mathbf{Y}^{(n+1)} \in \Gamma^{\epsilon}_{ICP,\,WMR}\left(\mathbf{X}^{(n+1)}\right)\right) \leq \sum_{i=1}^{n} \tilde{w}^{(i)} \mathsf{d}_{\mathsf{TV}}\left(\mathbf{A}_{TCP,\,MR}, \mathbf{A}^{\rightleftarrows i}_{TCP,\,MR}\right), \tag{11}$$

where $\mathbf{A}_{TCP,\,MR} = (A^{(1)}_{TCP,\,MR}, \ldots, A^{(n+1)}_{TCP,\,MR})$ are the transductive *maximum ranks (Schlembach et al., 2026), $A^{(i)}_{TCP,\,MR} = \max_{j=1}^{m} \operatorname{rank}(B_j^{(i)}, \{B_j^{(k)}\}_{k=1}^{n+1})$, i.e., the maximum ranks of Section 2.2 that the calibration instances and the test instance would receive if the test instance, with its true label, were added to the calibration set, and $\mathbf{A}^{\rightleftarrows i}_{TCP,\,MR}$ is the same vector with the ith and the $(n+1)$st elements swapped.*

The proof in Section B of the supplementary materials shows that $\Gamma^{\epsilon}_{\texttt{ICP},\texttt{WMR}}(\mathbf{X}^{(n+1)})$ contains the prediction region of a transductive variant of `WMax-Rank`, which ranks the test instance together with the calibration instances and to which the full conformal bound of Barber et al. (2023, Theorem 2) applies directly.

The bound has the same form as Equation (7), with the transductive maximum ranks as scores. `WMax-Rank` never computes these transductive maximum ranks. Because they rank all $n+1$ instances in the same way, swapping the ith and the $(n+1)$st elements of $\mathbf{A}_{\texttt{TCP},\texttt{MR}}$ is the same as swapping instances i and $n+1$ before ranking. Therefore, if the calibration and test instances are exchangeable, so is $\mathbf{A}_{\texttt{TCP},\texttt{MR}}$, every total variation distance vanishes, and $\Gamma^{\epsilon}_{\texttt{ICP},\texttt{WMR}}(\mathbf{X}^{(n+1)})$ is valid for all weights that satisfy (A2). The inductive maximum ranks,

in contrast, rank the calibration instances without the test instance and the test instance against the calibration instances only, so their vector is not exchangeable even if the data are, and a bound in terms of them does not vanish in general. With equal weights, `WMax-Rank` reduces to `Max-Rank` as implemented by Schlembach et al. (2026).

5. Experiments

This section describes the data sets, the experimental setup, and the results we used to compare `nmtCP`, `WDCP-MT`, and `WMax-Rank`.

5.1. Data Sets

Synthetic Data Sets The synthetic data sets follow the simulation design of Barber et al. (2023); Schlembach et al. (2025a,b). Each data set consists of six time series, where two output features ($o_{t,5}$ and $o_{t,6}$) are noisy linear combinations of four periodic input features. We consider four settings constructed from two types of distributional shift crossed with two noise structures. The *changepoint* settings introduce an abrupt shift in the linear coefficients at instance 1632, whereas the *distribution drift* settings introduce a gradual shift between instances 1632 and 2448 in the test set. The *correlated noise* settings draw the noise term from a bivariate normal distribution with covariance 0.5, whereas the *uncorrelated noise* settings draw the two noise components independently from a uniform and a chi-squared distribution, respectively. Full specifications of the coefficients and noise terms, together with visualisations of all four data sets, are in Appendix E of the supplementary materials.

ELEC2 The ELEC2 (Harries, 1999) data set contains information on electricity demand and price in the Australian states of New South Wales and Victoria, measured every 30 minutes from May 1996 to December 1998. We use the first 20,000 time steps shown in Figure S5 in the supplementary materials. For its first 17,424 time steps, ELEC2 contains no data for Victoria and stores a constant value for the Victorian demand, one of our two forecast targets, instead. The proper training set for the underlying model $\hat{\mathbf{f}}$ contains the first 15,000 time steps, the calibration set the next 1,000 time steps and the test set the final 4,000 time steps.

Tétouan City The Tétouan City data set (Salam and Hibaoui, 2018) contains weather information and electricity consumption for three distribution networks in the city of Tétouan, measured every 10 minutes in 2017. We use the first 40,000 time steps and the three electricity-consumption features shown in Figure S6 in the supplementary materials. The proper training set for the underlying model $\hat{\mathbf{f}}$ contains the first 24,000 time steps, the calibration set the next 1,000 time steps, and the test set the final 15,000 time steps. The Tétouan City features exhibit a gradual distribution shift in the test set.

5.2. Experimental Setup

Each instance has the values of the targets at two consecutive time steps as its label, so that $m = 4$ for the synthetic data sets and ELEC2 and $m = 6$ for Tétouan City. We fit an underlying model $\hat{\mathbf{f}}$ to the proper training set using scikit-learn's linear regressor. Next, we calibrate the conformal predictors on the calibration set using absolute residuals

as the nonconformity measure. Finally, the test set is processed sequentially. For each test instance, $\hat{\mathbf{f}}$ produces a point forecast, after which `nmtCP`, `WDCP-MT`, and `WMax-Rank` construct the associated prediction region. Once the true label becomes available, we update the calibration set by adding the new instance and removing the oldest one. This setup therefore implements both the reweighting and the refreshing strategies described in Section 2.3. The code to reproduce the experiments is available at https://doi.org/10.5281/zenodo.23807278.

All experiments are repeated 10 times with different random seeds that determine the synthetic data and the splits of `WDCP-MT`. `WDCP-MT` draws a new split for each test instance. On the real-world data sets, the replications therefore differ only in the splits of `WDCP-MT`.

We use the exponential ($w^{(i)} = e^{\gamma(i-n)}$ with $\gamma = 0.007$), linear ($w^{(i)} = i/n$), and constant weight functions of Schlembach et al. (2025a), and give the point mass at $+\infty$ the largest weight, $w^{(n+1)} = \max_i w^{(i)}$ (Barber et al., 2023). With a faster decay, the prediction regions adapt more quickly to a distribution shift, but the effective size of the calibration set decreases. We report the mean error rate, i.e., the fraction of true labels outside the prediction region, the mean side length and the mean volume of the bounded prediction regions, and the fraction of unbounded ones.

5.3. Experimental Results

The test sets of the changepoint data sets and of ELEC2 consist of three sections, those of the drift data sets of five. In the first section, the data are drawn from the same distribution as the initial calibration set, and in the last section from a new distribution that has replaced the whole calibration window. In the sections in between, the distribution changes or has changed, but the calibration window still contains instances from before the change.

In the first and the last section, `WMax-Rank` has error rates close to ϵ, whereas `nmtCP` and `WDCP-MT` become conservative as ϵ grows, with error rates between 0.33 and 0.44 at $\epsilon = 0.5$ on the synthetic data sets. With decaying weights, the regions are unbounded for the smallest values of ϵ, for `WDCP-MT`, which forms its regions from half of the calibration set, up to $\epsilon = 0.05$ with exponential weights. `WMax-Rank` produces smaller regions than `nmtCP` in almost all settings. For the uncorrelated noise and $\epsilon \geq 0.2$, `WDCP-MT` almost always produces smaller regions than `nmtCP`, and with constant weights the smallest regions of all three methods in the first section and, for $\epsilon \leq 0.3$, in the last section. For the correlated noise and the real-world data sets, its regions are mostly of similar or larger volume than those of `nmtCP`. Figure 1 shows these results for the synthetic changepoint data set with $e \sim [\mathcal{U}(-1, 1), \chi^2_2]^\top$, one row per section. In the sections in between, the error rates rise, mostly above ϵ, most of all for `WMax-Rank`, which produces the smallest regions there, whereas the order of `nmtCP` and `WDCP-MT` depends on ϵ and the weight function. In the synthetic data sets, the correlation of the nonconformity scores across dimensions increases after the change (Figures S8, S10, S12, and S14 in the supplementary materials).

In the first section of the ELEC2 test set, the nonconformity scores of the constant Victorian demand are 0. `WDCP-MT` shifts almost its entire significance budget to the two informative dimensions, whereas `WMax-Rank` returns $\mathcal{Y}$ because all calibration instances share the maximum rank (Appendix F.5 of the supplementary materials).

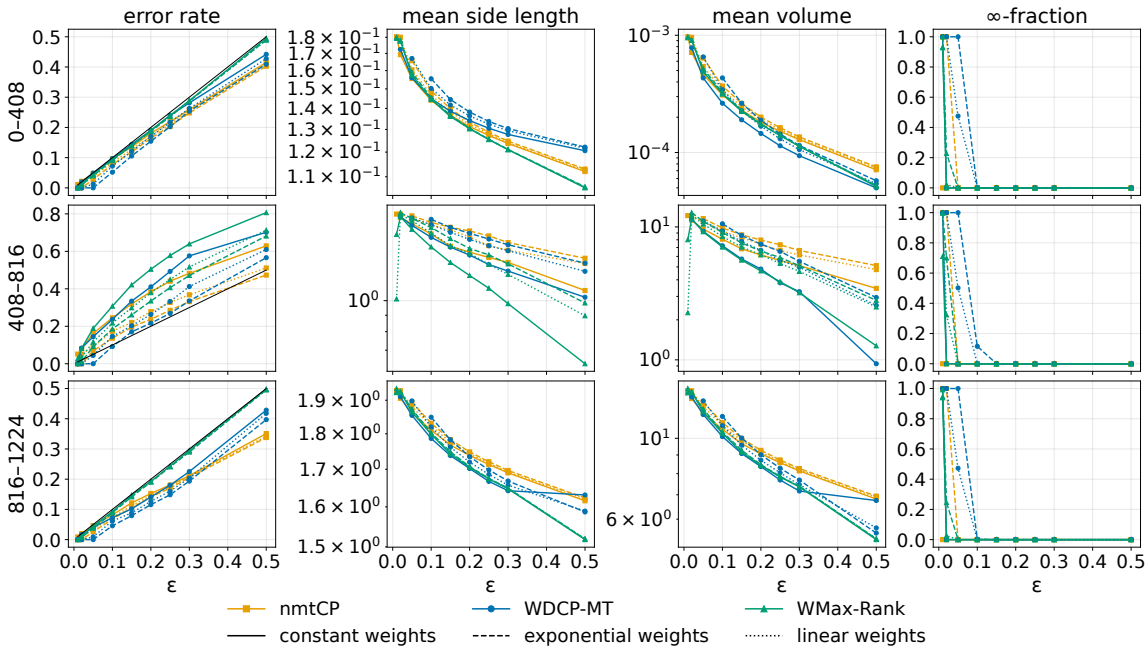


Figure 1: Mean error rate, mean hyperrectangle side length and mean hyperrectangle volume over the finite prediction regions, and fraction of prediction regions with at least one infinite bound (∞-fraction) as a function of the significance level ϵ for `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the synthetic changepoint data set with $e \sim [\mathcal{U}(-1, 1), \chi^2_2]^\top$, shown for three consecutive sections of the test set.

The test set of the Tétouan City data set (Salam and Hibaoui, 2018), displayed in Figure S6, shows no clear change in distribution. Figure S17 in the supplementary materials indicates that `WMax-Rank` is almost perfectly calibrated and produces the smallest prediction regions. `nmtCP` and `WDCP-MT` behave similarly, both having error rates that are far below the significance level ϵ. Here `WDCP-MT` produces larger regions than `nmtCP`.

Appendix F of the supplementary materials contains results for all data sets.

6. Discussion

The two coverage-gap bounds derived in Sections 3 and 4 differ in structure. Theorem 1 bounds the coverage gap of `WDCP-MT` by a sum of m dimension-wise sums of total variation distances, one per dimension and instance of O, so it grows with the number of dimensions and reflects the non-exchangeability of each dimension separately. Theorem 2 bounds the coverage gap of `WMax-Rank` by a single sum of total variation distances of the transductive maximum ranks $\mathbf{A}_{\texttt{TCP,MR}}$, one per calibration instance, irrespective of m. If the dimension-wise scores are comonotone, i.e., strictly increasing functions of one another, the maximum ranks are the ranks of any single dimension, and the bound of Theorem 2 is at most a

fraction $1/m$ of that of Theorem 1, case (i). In exchange, each distance in Theorem 2 depends on the joint distribution of the dimension-wise scores, so that a change in their dependence alone can increase the bound of Theorem 2 but not that of Theorem 1, case (i).

In the experiments, `nmtCP` is conservative for larger ϵ and produces larger regions than `WMax-Rank` in almost all settings, as observed by Schlembach et al. (2025a). Its uniform allocation $\epsilon_j = \epsilon/m$, covered by Theorem 1, case (i), exploits neither the dependence between the nonconformity scores nor differences in their marginal distributions, which the two methods introduced in this paper address.

`WDCP-MT` achieves its greatest gains in informational efficiency relative to `nmtCP` when the nonconformity scores are weakly correlated across output dimensions and their marginal distributions differ in shape. This behaviour is most clearly visible in the experiments on the synthetic data sets with uncorrelated noise. In this setting, the two marginal distributions are sufficiently different that a uniform allocation of the significance budget, as used by `nmtCP`, is suboptimal. `WDCP-MT` instead allocates the budget according to the estimated reduction in log-volume obtainable in each dimension, which depends on the shapes of the quantile functions of the scores. The result is a prediction region whose volume is closer to the minimum achievable under Boole's inequality. This gain is consistent with the analysis of Schlembach et al. (2025b). In addition, the proof of Theorem 1 rests on an upper bound of the coverage gap for each output dimension. It therefore allows for different weights to be used for different output dimensions, a flexibility `WMax-Rank` does not possess and which we do not explore in this paper.

`WMax-Rank` produces smaller regions than both `nmtCP` and `WDCP-MT` in almost all settings where the nonconformity scores are positively correlated across output dimensions. In these settings, `nmtCP` and `WDCP-MT`, bound by Boole's inequality, allocate a smaller error budget to each dimension than is jointly necessary to achieve the targeted error rate ϵ. This gain is not free since even with equal weights, the threshold of `Max-Rank` can be one order statistic larger than that of the Bonferroni correction (Schlembach et al., 2026, Section 6), and with tied scores `WMax-Rank` can return $\mathcal{Y}$ where `nmtCP` does not, as in the first section of ELEC2. The gain of `WMax-Rank` is largest on Tétouan City, where the cross-dimensional correlations are high throughout the test set.

All three methods exhibit elevated error rates in the section of the test set in which the distribution has begun to shift but the sliding calibration window has not yet been fully replaced with instances from the new distribution. We observe that methods that produce smaller prediction regions under stable conditions are more sensitive to calibration errors when the distribution changes. From the perspective of the coverage-gap bounds, the total variation distance terms in Theorems 1 and 2 are plausibly largest during this transition period, when the distribution of the calibration nonconformity scores diverges sharply from the distribution of the test nonconformity scores. The reweighting strategy partially compensates for this by downweighting older calibration instances, but it cannot fully recover validity until the calibration window reflects the new distribution.

7. Conclusion

This paper studies uncertainty quantification for multi-target regression under non-exchangeability. Starting from the reweighting framework of Barber et al. (2023), we propose

`WDCP-MT` and `WMax-Rank`, which extend the Boole-based and the dependence-aware correction methods for hyperrectangular conformal prediction regions, and derive a coverage-gap bound for each. The experiments support the expected operating regimes. `WDCP-MT` tightens the regions when the dimension-wise nonconformity scores are uncorrelated and their marginal distributions differ in shape. `WMax-Rank` yields smaller regions than `nmtCP` in almost all settings, most of all when the scores are positively correlated, and is well calibrated when the distribution is stable. All methods degrade while the distribution changes and the calibration window has not yet adapted. Tighter hyperrectangular prediction regions for non-exchangeable multi-target data can thus be obtained by matching the correction method to the dependence structure of the nonconformity scores.

Future work includes the choice of weights for the weighted quantile function. In addition, experiments with a larger number of output dimensions, more complex non-linear underlying models, and negatively correlated nonconformity scores can shine further light on the behaviour and stability of the methods presented in this paper. Finally, randomised tie-breaking for `WMax-Rank` may be considered.

Acknowledgments

We would like to thank Tianmin Xie and Professor Vladimir Vovk for their critical assessment of our previous work during the COPA 25 which led to this research, as well as Dr Misha Glazunov for his insightful comments on independence.
We acknowledge the assistance of Anthropic's Claude and OpenAI's ChatGPT in checking the theoretical results for correctness, implementing the dynamic programming algorithm, checking the manuscript for consistency, and polishing the writing style.

References

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 732–749, 2018.

Yunjie Fan and Matteo Sesia. Interpretable multivariate conformal prediction with fast transductive standardization, 2025. arXiv:2512.15383.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc., 2021.

Michael Bonnell Harries. SPLICE-2 comparative evaluation: Electricity pricing, 1999.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101, 2021.

Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Self-published, 3 edition, 2025.

Abdul Rahim Salam and Abdelaaziz El Hibaoui. Comparison of machine learning algorithms for the power consumption prediction: Case study of Tetouan city. In *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*, pages 1–5, 2018.

Filip Schlembach, Evgueni Smirnov, Irena Koprinska, and Mark H. M. Winands. Conformal multistep-ahead multivariate time-series forecasting. *Machine Learning*, 114(7):165, 2025a.

Filip Schlembach, Evgueni Smirnov, and Mark H. M. Winands. Dynamic conformal prediction for multi-target regression: Optimising informational efficiency under joint validity. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pages 193–213, 2025b.

Filip Schlembach, Evgueni Smirnov, and Mark H. M. Winands. Corrected Max-Rank for Finite-Sample Conformal Prediction. In *Proceedings of the Fifteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 329 of *Proceedings of Machine Learning Research*, pages 199–221. PMLR, September 2026. URL https://proceedings.mlr.press/v329/schlembach26a.html.

Kamilė Stankevičiūtė, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. In *Advances in Neural Information Processing Systems*, 2021.

M. Stocker, W. Małgorzewicz, M. Fontana, and S. Ben Taieb. A gentle introduction to conformal time series forecasting, 2025. arXiv:2511.13608.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, Christian A. Naesseth, and Eric Nalisnick. Max-Rank: Efficient multiple testing for conformal prediction. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3898–3906, 2025.

Vladimir Vovk. Transductive conformal predictors. In *Artificial Intelligence Applications and Innovations*, pages 348–360. Springer, 2013.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2022.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11559–11569, 2021.

Supplementary Materials for Tightening Hyperrectangular Conformal Prediction Regions for Multi-Target Regression under Non-exchangeability

Filip Schlembach FILIP@SCHLEMBACH.BE
Dennis Mertens Velasquez DENNIS.MERTENS@MAASTRICHTUNIVERSITY.NL
Evgueni Smirnov SMIRNOV@MAASTRICHTUNIVERSITY.NL
Mark H. M. Winands M.WINANDS@MAASTRICHTUNIVERSITY.NL
Department of Advanced Computing Sciences, Maastricht University, Maastricht, The Netherlands

Editors: Andy Song, Bo Han and Sarah Erfani

Appendix A. Proof of Theorem 1: Coverage Gap with Data-Dependent Dimension-Wise Significance Levels

Theorem 1 of the main paper bounds the coverage gap of hyperrectangular predictors whose dimension-wise significance levels $\epsilon_1, \ldots, \epsilon_m$ satisfy $\sum_{j=1}^{m} \epsilon_j \leq \epsilon$. If the dimension-wise significance levels are fixed in advance, as in `Bonferroni CP` and `nmtCP`, the result of Barber et al. (2023) applies to each dimension directly. The dimension-wise significance levels of `WDCP-MT`, in contrast, depend on the nonconformity scores because they are selected from the scores in the set E of a partition of the calibration set, and the prediction region is formed from the scores in the complementary set O. For non-exchangeable data, and time series in particular, the scores in E and O are dependent in general, so the disjointness of E and O alone does not establish that the result of Barber et al. (2023) applies to the scores in O. This section states the assumptions under which it does, proves Theorem 1 in a conditional and an unconditional form that covers both cases, with fixed dimension-wise significance levels as the special case $E = \emptyset$, and examines what the split costs.

A.1. Setting and Assumptions

For $i = 1, \ldots, n+1$ let $\mathbf{B}^{(i)} = (B_1^{(i)}, \ldots, B_m^{(i)})$ with $B_j^{(i)} = \mathsf{S}_{\mathtt{dim},j}(\mathbf{X}^{(i)}, Y_j^{(i)})$ be the vector of dimension-wise nonconformity scores of calibration instance $i \leq n$ and of the test instance $i = n+1$, let $\mathbf{B}_{1:m} = (\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(n+1)})$ collect the scores of all instances in all output dimensions, and let $\mathbf{B}_{1:m}^{\rightleftarrows i}$ denote $\mathbf{B}_{1:m}$ with $\mathbf{B}^{(i)}$ and $\mathbf{B}^{(n+1)}$ swapped. Let $\sigma = (E, O)$ be the split into the disjoint index sets E and O with $E \cup O = \{1, \ldots, n\}$. In `WDCP-MT`, σ is drawn uniformly at random with $|E| = \lfloor n/2 \rfloor$, but σ may also be deterministic, e.g. a blocked split. $E = \emptyset$ if the dimension-wise significance levels are fixed in advance. Let $\mathbf{B}_E = (\mathbf{B}^{(i)})_{i \in E}$ denote the scores in E in all m dimensions (an empty vector if $E = \emptyset$). For each dimension j let

$$\mathbf{V}_j = \left(B_j^{(i)}\right)_{i \in O \cup \{n+1\}} \tag{S1}$$

be the scores in O together with the test score, and let $\mathbf{V}_j^{\rightleftarrows i}$ denote $\mathbf{V}_j$ with the entries of $i \in O$ and $n+1$ swapped. The weights normalised over O are $\tilde{w}_O^{(i)} = w^{(i)}/(w^{(n+1)} +$

© 2026 F. Schlembach, D. Mertens Velasquez, E. Smirnov & M.H.M. Winands.

$\sum_{k\in O} w^{(k)}$) for $i \in O \cup \{n+1\}$, as in the `WDCP-MT` algorithm of the main paper. Unlike $\tilde{w}^{(i)} = w^{(i)} / \sum_{k=1}^{n+1} w^{(k)}$, they are normalised over the scores in O and the test instance only. We write $\mathcal{L}(\cdot)$ for the distribution (law) of a random vector and $\mathcal{L}(\cdot \mid \cdot)$ for its conditional distribution.

We assume the following.

(A1) The underlying model $\hat{\mathbf{f}}$ is fitted on a proper training set $\mathcal{D}_{tr}$ that is disjoint from the calibration and test instances, so that $\mathsf{S}_{\mathtt{dim}}(\hat{\mathbf{f}}, \cdot, \cdot)$ is fixed given $\mathcal{D}_{tr}$. All probabilities, distributions, and total variation distances are conditional on $\mathcal{D}_{tr}$. We suppress this in the notation.

(A2) The weights $w^{(1)}, \ldots, w^{(n+1)} \geq 0$ are fixed in advance and satisfy $w^{(n+1)} > 0$ and $w^{(i)} \leq w^{(n+1)}$ for $i = 1, \ldots, n$.

(A3) The split σ is drawn independently of $\mathbf{B}_{1:m}$. This holds for any random split that does not look at the scores and for any deterministic split.

(A4) The dimension-wise significance levels $\epsilon_j = \epsilon_j(\sigma, \mathbf{B}_E)$ are measurable functions of σ and $\mathbf{B}_E$ only, and $\sum_{j=1}^{m} \epsilon_j \leq \epsilon$ whenever the selection succeeds.[1] Otherwise, $\Gamma^{\epsilon}(\mathbf{X}^{(n+1)}) = \mathcal{Y}$.

These are the assumptions of Theorem 1 of the main paper. No assumption is made on the joint distribution of $\mathbf{B}_{1:m}$, and in particular none on the dependence between $\mathbf{B}_E$, the scores in O, and the test score. (A4) holds by construction of the `WDCP-MT` algorithm, and for any other rule that selects the dimension-wise significance levels from the scores in E only (e.g. a different objective or solver). For $E = \emptyset$, (A4) states that the dimension-wise significance levels are fixed in advance. In (A4), σ fixes which indices belong to E, and therefore which weights the selection of the dimension-wise significance levels uses, while $\mathbf{B}_E$ contains the score values at these indices. The proof relies on this assumption because the bound of Barber et al. (2023) requires the level of the weighted quantile to be fixed, and by (A4) the dimension-wise significance levels are fixed once the split and the scores in E are, so the bound can be applied conditionally on σ and $\mathbf{B}_E$. If ϵ_j also depended on the scores in O or on the test score, conditioning on $\mathbf{B}_E$ would not fix it, which is the oversight in Schlembach et al. (2025b).

(A1) does not require the training data to be independent of the calibration and test instances, which contiguous blocks of a time series would violate. Given $\mathcal{D}_{tr}$, the underlying model is a fixed function and the scores have some conditional distribution. Since the bound of Barber et al. (2023) holds for an arbitrary distribution, it holds for this one (cf. Schlembach et al., 2026, Section 2). Dependence between the training data and the calibration and test instances enters only through the total variation distances, which are now those of the conditional distributions given $\mathcal{D}_{tr}$. Averaged over $\mathcal{D}_{tr}$, they are at least the unconditional ones, by the same argument as in Step 4 of the proof of Theorem S2 below. For instance, if the model fits the calibration instances adjacent in time to the training block better than later instances, these calibration instances become distinguishable from the test instance and the bound grows. Decaying weights reduce their influence, and the sliding calibration window moves away from the training block.

1. I.e., for `WDCP-MT` whenever the allocation problem is feasible.

A.2. Weighted Split Conformal Bound

For a probability distribution μ on $\bar{\mathbb{R}}$ and $\tau \in [0,1]$, the τ-quantile is

$$\mathsf{Q}_\tau(\mu) = \inf\{t \in \bar{\mathbb{R}} : \mu([-\infty, t]) \geq \tau\}. \tag{S2}$$

In particular, $\mathsf{Q}_0(\mu) = -\infty$. For an index set $I \subseteq \{1,\dots,n\}$ and $\mu = \sum_{i\in I} \tilde{v}^{(i)}\,\delta_{A^{(i)}} + \tilde{v}^{(n+1)}\,\delta_{+\infty}$, $\mathsf{Q}_\tau(\mu)$ is the smallest $A^{(i)}$ at which the cumulative weight of the sorted scores reaches τ, and $+\infty$ if no score reaches it. With $I = \{1,\dots,n\}$ and equal weights $\tilde{v}^{(i)} = 1/(n+1)$, $\mathsf{Q}_{1-\epsilon}(\mu)$ is the $\lceil (1-\epsilon)(n+1)\rceil$th smallest of $A^{(1)},\dots,A^{(n)}$, with the conventions that the 0th smallest is $-\infty$ and the $(n+1)$st smallest is $+\infty$; this is the order statistic of unweighted split conformal prediction (Barber et al., 2023, Eq. (5)).

We use the result of Barber et al. (2023) in the following form.

Lemma S1 (Barber et al., 2023, Theorem 2) *Let $I \subseteq \{1,\dots,n\}$ be an index set, let $\mathbf{A} = (A^{(i)})_{i\in I\cup\{n+1\}}$ be real-valued random variables with an arbitrary joint distribution, let $(v^{(i)})_{i\in I\cup\{n+1\}}$ be fixed non-negative weights with $v^{(n+1)} > 0$ and $v^{(i)} \leq v^{(n+1)}$ for $i \in I$, let $\tilde{v}^{(i)} = v^{(i)}/(v^{(n+1)} + \sum_{k\in I} v^{(k)})$, and let $\epsilon \in [0,1]$ be fixed. Then*

$$\mathsf{P}\Big(A^{(n+1)} > \mathsf{Q}_{1-\epsilon}\Big(\sum_{i\in I} \tilde{v}^{(i)}\,\delta_{A^{(i)}} + \tilde{v}^{(n+1)}\,\delta_{+\infty}\Big)\Big) \leq \epsilon + \sum_{i\in I} \tilde{v}^{(i)}\,\mathsf{d_{TV}}\left(\mathcal{L}(\mathbf{A}), \mathcal{L}(\mathbf{A}^{\rightleftarrows i})\right), \tag{S3}$$

where $\mathbf{A}^{\rightleftarrows i}$ is $\mathbf{A}$ with the entries i and $n+1$ swapped.

Barber et al. (2023) state their result for the scores $\mathsf{S}(X_i, Y_i)$ of data points (X_i, Y_i) with an arbitrary joint distribution under a fixed score function S (their Appendix A). Lemma S1 follows by relabelling the indices in I as $1,\dots,|I|$ and the test index $n+1$ as $|I|+1$, and choosing the data points $(X_i, Y_i) = (x_0, A^{(i)})$ with a constant x_0 and the score function $\mathsf{S}(x,y) = y$, for which their score vector equals $\mathbf{A}$ and the score vector of the swapped data equals $\mathbf{A}^{\rightleftarrows i}$. They also state the result for $v^{(n+1)} = 1$ and $v^{(i)} \in [0,1]$. The general form follows because the weighted quantile depends on the weights only through $\tilde{v}$, so dividing all weights by $v^{(n+1)}$ changes neither side. The condition $v^{(i)} \leq v^{(n+1)}$ is used in step (22) of their proof. The endpoints $\epsilon \in \{0,1\}$ are trivial, since for $\epsilon = 1$ the quantile is $-\infty$ and the left-hand side is at most one, and for $\epsilon = 0$ the quantile is $+\infty$ because $\tilde{v}^{(n+1)} > 0$, so the left-hand side is zero.

A.3. Conditional Coverage Gap Bound

Theorem S2 (Theorem 1 of the main paper, full version) *Under (A1)–(A4), the prediction region $\Gamma^\epsilon(\mathbf{X}^{(n+1)})$, whose dimension-wise thresholds are the weighted quantiles of the scores in O with the weights $\tilde{w}_O^{(i)}$, satisfies the following.*

(i) Conditionally on the split and the scores in E,

$$\mathsf{P}\big(\mathbf{Y}^{(n+1)} \notin \Gamma^\epsilon(\mathbf{X}^{(n+1)}) \;\big|\; \sigma, \mathbf{B}_E\big) \leq \epsilon + \sum_{j=1}^{m} \sum_{i\in O} \tilde{w}_O^{(i)}\,\kappa_{ij}(\sigma, \mathbf{B}_E), \tag{S4}$$

where $\kappa_{ij}(\sigma, \mathbf{B}_E) = \mathsf{d_{TV}}\Big(\mathcal{L}(\mathbf{V}_j \mid \sigma, \mathbf{B}_E), \mathcal{L}(\mathbf{V}_j^{\rightleftarrows i} \mid \sigma, \mathbf{B}_E)\Big)$.

(ii) Conditionally on the split,

$$(1-\epsilon) - \mathsf{P}\big(\mathbf{Y}^{(n+1)} \in \Gamma^{\epsilon}(\mathbf{X}^{(n+1)}) \;\big|\; \sigma\big) \le \sum_{j=1}^{m} \sum_{i \in O} \tilde{w}_O^{(i)}\, \Delta_{ij}(\sigma), \tag{S5}$$

where

$$\Delta_{ij}(\sigma) = \mathsf{d_{TV}}\Big(\mathcal{L}(\mathbf{B}_E, \mathbf{V}_j \mid \sigma), \mathcal{L}(\mathbf{B}_E, \mathbf{V}_j^{\rightleftarrows i} \mid \sigma)\Big) = \mathsf{E}\big[\kappa_{ij}(\sigma, \mathbf{B}_E) \;\big|\; \sigma\big]. \tag{S6}$$

By (A3), the conditional laws given σ are the unconditional laws of the subvectors of $\mathbf{B}_{1:m}$ at the index sets E and O fixed by σ.

(iii) Unconditionally,

$$(1-\epsilon) - \mathsf{P}\big(\mathbf{Y}^{(n+1)} \in \Gamma^{\epsilon}(\mathbf{X}^{(n+1)})\big) \le \mathsf{E}_{\sigma}\Big[\sum_{j=1}^{m} \sum_{i \in O} \tilde{w}_O^{(i)}\, \Delta_{ij}(\sigma)\Big], \tag{S7}$$

which is the bound of Theorem 1 of the main paper.

(iv) For every $i \in O$ and j,

$$\mathsf{d_{TV}}\Big(\mathcal{L}(\mathbf{V}_j \mid \sigma), \mathcal{L}(\mathbf{V}_j^{\rightleftarrows i} \mid \sigma)\Big) \le \Delta_{ij}(\sigma) \le \mathsf{d_{TV}}\Big(\mathcal{L}(\mathbf{B}_{1:m}), \mathcal{L}(\mathbf{B}_{1:m}^{\rightleftarrows i})\Big). \tag{S8}$$

The lower bound holds with equality if $\mathbf{B}_E$ is independent of $\mathbf{V}_j$, and $\Delta_{ij}(\sigma) = 0$ if $\mathbf{B}^{(1)}, \dots, \mathbf{B}^{(n+1)}$ are exchangeable. If $E = \emptyset$, then $O = \{1, \dots, n\}$, $\mathbf{V}_j = \mathbf{B}_j = (B_j^{(i)})_{i=1}^{n+1}$, the lower bound holds with equality, and (S7) reads

$$(1-\epsilon) - \mathsf{P}\big(\mathbf{Y}^{(n+1)} \in \Gamma^{\epsilon}(\mathbf{X}^{(n+1)})\big) \le \sum_{j=1}^{m} \frac{\sum_{i=1}^{n} w^{(i)}\, \mathsf{d_{TV}}\Big(\mathcal{L}(\mathbf{B}_j), \mathcal{L}(\mathbf{B}_j^{\rightleftarrows i})\Big)}{\sum_{k=1}^{n+1} w^{(k)}}. \tag{S9}$$

Proof *Step 1: fixing the split.* By (A3), conditioning on σ does not change the law of $\mathbf{B}_{1:m}$. We therefore fix σ, so that E, O, and the weights $\tilde{w}_O$ are deterministic. If $E = \emptyset$, $\mathbf{B}_E$ is empty, and conditioning on it below can be ignored.

Step 2: one dimension, conditionally on E. Fix a value $\boldsymbol{\beta}_E$ of $\mathbf{B}_E$. By (A4), the dimension-wise significance levels $\epsilon_j = \epsilon_j(\sigma, \boldsymbol{\beta}_E)$ are then deterministic. If the selection fails, $\Gamma^{\epsilon}(\mathbf{X}^{(n+1)}) = \mathcal{Y}$ and the miscoverage probability is zero. Otherwise, by the definition of the dimension-wise prediction regions, $Y_j^{(n+1)} \notin \Gamma_j^{\epsilon_j}(\mathbf{X}^{(n+1)})$ if and only if

$$B_j^{(n+1)} > \mathsf{Q}_{1-\epsilon_j}\Big(\sum_{i \in O} \tilde{w}_O^{(i)}\, \delta_{B_j^{(i)}} + \tilde{w}_O^{(n+1)}\, \delta_{+\infty}\Big). \tag{S10}$$

This is the event of Lemma S1 with $I = O$, $\mathbf{A} = \mathbf{V}_j$, and weights $(w^{(i)})_{i \in O \cup \{n+1\}}$, which satisfy the conditions of the lemma by (A2), with ϵ_j in place of ϵ. Since the lemma holds for an arbitrary joint distribution, it holds in particular for the conditional law of $\mathbf{V}_j$ given σ and $\mathbf{B}_E = \boldsymbol{\beta}_E$, which gives

$$\mathsf{P}\big(Y_j^{(n+1)} \notin \Gamma_j^{\epsilon_j}(\mathbf{X}^{(n+1)}) \;\big|\; \sigma, \mathbf{B}_E = \boldsymbol{\beta}_E\big) \le \epsilon_j + \sum_{i \in O} \tilde{w}_O^{(i)}\, \kappa_{ij}(\sigma, \boldsymbol{\beta}_E). \tag{S11}$$

Step 3: all dimensions. Applying Boole's inequality to the conditional probability, summing the bounds of Step 2 over j, and using $\sum_j \epsilon_j \leq \epsilon$ from (A4) yields (S4).

Step 4: removing the conditioning. Since $i \in O$, the swap does not affect $\mathbf{B}_E$, so the laws $\mu_0 = \mathcal{L}(\mathbf{B}_E, \mathbf{V}_j \mid \sigma)$ and $\mu_i = \mathcal{L}(\mathbf{B}_E, \mathbf{V}_j^{\rightleftarrows i} \mid \sigma)$ have the same marginal $\lambda = \mathcal{L}(\mathbf{B}_E \mid \sigma)$ on the first component. For two such laws, the densities with respect to $\nu = (\mu_0 + \mu_i)/2$ factorise into the conditional densities given the first component, so that

$$\mathsf{d}_{\mathsf{TV}}(\mu_0, \mu_i) = \frac{1}{2} \int \left| \frac{d\mu_0}{d\nu} - \frac{d\mu_i}{d\nu} \right| d\nu = \int \mathsf{d}_{\mathsf{TV}}(\mu_0(\cdot \mid \boldsymbol{\beta}_E), \mu_i(\cdot \mid \boldsymbol{\beta}_E))\ \lambda(d\boldsymbol{\beta}_E), \tag{S12}$$

i.e., $\Delta_{ij}(\sigma) = \mathsf{E}[\kappa_{ij}(\sigma, \mathbf{B}_E) \mid \sigma]$. Taking the expectation of (S4) over $\mathbf{B}_E$ for the fixed split, and using that $\tilde{w}_O$ depends on σ only, therefore gives (S5). Taking the expectation over σ as well gives (S7).

Step 5: comparison. The total variation distance cannot increase under a measurable map applied to both arguments. Projecting $(\mathbf{B}_E, \mathbf{V}_j)$ onto $\mathbf{V}_j$ gives the lower bound in (iv), and $(\mathbf{B}_E, \mathbf{V}_j)$ and $(\mathbf{B}_E, \mathbf{V}_j^{\rightleftarrows i})$ are the same projection of $\mathbf{B}_{1:m}$ and $\mathbf{B}_{1:m}^{\rightleftarrows i}$, respectively, which gives the upper bound, because by (A3) the law of $\mathbf{B}_{1:m}$ given σ is its unconditional law. If $\mathbf{B}_E$ is independent of $\mathbf{V}_j$, it is also independent of $\mathbf{V}_j^{\rightleftarrows i}$, so μ_0 and μ_i are product laws with the common factor λ and their distance equals that of their second factors. This holds trivially for $E = \emptyset$, where in addition $\tilde{w}_O^{(i)} = w^{(i)} / \sum_{k=1}^{n+1} w^{(k)}$. Finally, under exchangeability $\mathbf{B}_{1:m}$ and $\mathbf{B}_{1:m}^{\rightleftarrows i}$ have the same law, so the upper bound is zero. ■

Remark S3 *Steps 2 and 3 apply Lemma S1 to each dimension separately, so Theorem S2 remains valid if each dimension j uses its own weights $w_j^{(1)}, \ldots, w_j^{(n+1)}$ satisfying (A2), with $\tilde{w}_O^{(i)}$ replaced by the correspondingly normalised $\tilde{w}_{j,O}^{(i)}$.*

Remark S4 *Theorem S2 uses the same device that (Barber et al., 2023, Section 4.5) propose for random weights. The data-dependent quantity, here the dimension-wise significance levels, is conditioned on, and the total variation distances are replaced by their conditional counterparts. Part (ii) shows that, after averaging over $\mathbf{B}_E$, these conditional distances are exactly the distances between the joint laws of $(\mathbf{B}_E, \mathbf{V}_j)$ before and after the swap, for the index sets fixed by the split.*

Remark S5 *The lower bound in (iv) is the distance on $\mathbf{V}_j$ alone, i.e., the distance one would obtain by treating the dimension-wise significance levels as fixed in advance and applying the result of Barber et al. (2023) to the scores in O naively. It is attained when the scores in E carry no information about the scores in O and the test score, e.g., when the scores are independent across time but not identically distributed. For time series with serial dependence this does not hold. The random split interleaves E and O in time, so the neighbours of an instance $i \in O$ typically lie in E, and $\Delta_{ij}(\sigma)$ can exceed the naive distance. Theorem S2 therefore states the bound without assuming independence between E and O.*

As an extreme example, suppose that the scores in O are copies of those in E, with the same weights. `WDCP-MT` selects the dimension-wise significance levels that minimise the volume estimated on E, and, as for a model evaluated on its training data, this estimate is optimistic, since among many feasible allocations the one with the smallest estimated volume

tends to be one whose estimate happens to be favourable. Since the quantiles computed from O equal those computed from E, the region inherits this optimism, exactly as if the levels were selected and the region formed on the same scores, as in Schlembach et al. (2025b). In this scenario, the split offers no protection, and a fresh test score exceeds the thresholds more often than ϵ_j. This requires both dimension-wise significance levels selected from the data and dependence between E and O. With levels fixed in advance there is nothing to overfit, and with E independent of O the fit to E does not carry over.

Theorem S2 does not certify coverage in this scenario. Given $\mathbf{B}_E$, the score of each instance $i \in O$ is known, because it is a copy of a score in E. After the swap, position i holds the test score instead, which for continuous scores differs from this copy almost surely. The two conditional laws can thus be told apart with certainty, so $\kappa_{ij} = 1$ and the bound exceeds one. The distance on $\mathbf{V}_j$ alone, in contrast, ignores $\mathbf{B}_E$. It can be zero, e.g. if the scores in E and the test score are independent and identically distributed, and would then wrongly certify coverage. Near-copies are realistic, because under strong short-range autocorrelation neighbouring instances have similar scores, and the random split typically assigns them to different halves.

A.4. What the Split Costs

Validity is retained, but the split has two costs.

Effective calibration set size. The region is formed from the $\lceil n/2 \rceil$ scores in O instead of all n calibration scores. The weighted quantile is therefore computed on a coarser grid, and the mass at infinity $\tilde{w}_O^{(n+1)} = w^{(n+1)}/(w^{(n+1)} + \sum_{k \in O} w^{(k)})$ roughly doubles compared to the unsplit calibration set. Any dimension-wise significance level $\epsilon_j < \tilde{w}_O^{(n+1)}$ yields an unbounded interval, so the smallest dimension-wise significance level that can produce a bounded region roughly doubles. This matters most for small ϵ and many output dimensions, where the dimension-wise significance levels are small.

Optimality of the dimension-wise significance levels. The allocation minimises the volume of the region estimated from the scores in E, while the region is formed from the scores in O. The selected levels are thus optimal for E, where the region is never larger than that of the Bonferroni correction (Proposition S10), but not necessarily for O, and the region formed from O is not guaranteed to be smaller than the region that the uniform allocation $\epsilon_j = \epsilon/m$ would produce on O.

Appendix B. Coverage Gap of `WMax-Rank`

In this Section we prove the upper bound on the miscoverage gap for the inductive variant of `WMax-Rank` stated in Theorem 2 and interpret this result. To do this, we first introduce the transductive variant of `WMax-Rank` and derive its coverage gap. We then show that the inductive variant of `WMax-Rank` contains the prediction region of the transductive variant and conclude that it must therefore inherit the transductive coverage gap bound. Finally, Remark S8 shows that the bound also holds for the region returned by Algorithm S1, which gives the pseudocode of `WMax-Rank` described in Section 4.1 of the main paper.

Algorithm S1: `WMax-Rank`

Input: test instance $\mathbf{x}^{(n+1)}$; predictions $\hat{f}_j(\mathbf{x}^{(n+1)})$ for $j = 1, \ldots, m$; nonconformity scores $\left\{\beta_j^{(i)}\right\}_{i=1}^{n}$; normalised weights $\left\{\tilde{w}^{(i)}\right\}_{i=1}^{n+1}$; global significance level ϵ

Output: hyperrectangular prediction region $\Gamma^{\epsilon}_{\texttt{ICP,WMR}}(\mathbf{x}^{(n+1)})$

for $j = 1, \ldots, m$ **do**
  Sort $\left\{\beta_j^{(i)}\right\}_{i=1}^{n}$ in ascending order using a permutation π_j;
  for $i = 1, \ldots, n$ **do**
    $\rho^{(i)}_{\texttt{ICP},j} \leftarrow \text{rank}\big(\beta_j^{(i)}, \{\beta_j^{(k)}\}_{k=1}^{n}\big)$;
  end
end
for $i = 1, \ldots, n$ **do**
  $\alpha^{(i)}_{\texttt{ICP,MR}} \leftarrow \max_{j=1}^{m} \rho^{(i)}_{\texttt{ICP},j}$;
end
$q^{(1-\epsilon)}_{\texttt{ICP}} \leftarrow \mathsf{Q}_{1-\epsilon}\Big(\sum_{i=1}^{n} \tilde{w}^{(i)}\delta_{\alpha^{(i)}_{\texttt{ICP,MR}}} + \tilde{w}^{(n+1)}\delta_{+\infty}\Big)$;
if $q^{(1-\epsilon)}_{ICP} = +\infty$ **or** $q^{(1-\epsilon)}_{ICP} = n$ **then**
  return $\Gamma^{\epsilon}_{\texttt{ICP,WMR}}(\mathbf{x}^{(n+1)}) = \mathcal{Y}$;
else
  for $j = 1, \ldots, m$ **do**
    $\Gamma_j^{\epsilon_j}(\mathbf{x}^{(n+1)}) \leftarrow \left\{ y_j \in \mathcal{Y}_j : \mathsf{S}_{\texttt{dim},j}(\mathbf{x}^{(n+1)}, y_j) \leq \beta_j^{\left(\pi_j^{\left(q^{(1-\epsilon)}_{\texttt{ICP}}+1\right)}\right)} \right\}$;
  end
  return $\Gamma^{\epsilon}_{\texttt{ICP,WMR}}(\mathbf{x}^{(n+1)}) = \bigtimes_{j=1}^{m} \Gamma_j^{\epsilon_j}(\mathbf{x}^{(n+1)})$;
end

We start by introducing the transductive `Max-Rank` nonconformity score function $\mathsf{S}_{\texttt{TCP,MR}}$ previously used in Schlembach et al. (2026)

$$\mathsf{S}_{\texttt{TCP,MR}}\left(\hat{\mathbf{f}}, \mathcal{D} \cup \left\{\left(\mathbf{x}^{(n+1)}, \mathbf{y}\right)\right\}, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right) = \max_{j=1}^{m}\left(\left\{\rho^{(i)}_{\texttt{TCP},j}\right\}\right) \tag{S13}$$

such that

$$\alpha^{(i)}_{\texttt{TCP,MR}} = \begin{cases} \mathsf{S}_{\texttt{TCP,MR}}\left(\hat{\mathbf{f}}, \mathcal{D} \cup \left\{\left(\mathbf{x}^{(n+1)}, \mathbf{y}\right)\right\}, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right) & \text{if } i \in \{1, \ldots, n\} \\ \mathsf{S}_{\texttt{TCP,MR}}\left(\hat{\mathbf{f}}, \mathcal{D} \cup \left\{\left(\mathbf{x}^{(n+1)}, \mathbf{y}\right)\right\}, \mathbf{x}^{(i)}, \mathbf{y}\right) & \text{if } i = n+1. \end{cases} \tag{S14}$$

with

$$\rho^{(i)}_{\texttt{TCP},j} = \text{rank}\left(\beta_j^{(i)}, \left\{\beta_j^{(k)}\right\}_{k=1}^{n+1}\right) \tag{S15}$$

and

$$\boldsymbol{\beta}^{(i)} = \begin{cases} \mathsf{S}_{\texttt{dim}}\left(\hat{\mathbf{f}}, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right) & \text{if } i \in \{1, \ldots, n\} \\ \mathsf{S}_{\texttt{dim}}\left(\hat{\mathbf{f}}, \mathbf{x}^{(i)}, \mathbf{y}\right) & \text{if } i = n+1. \end{cases} \tag{S16}$$

With the normalised weights $\tilde{w}^{(i)}$ of (A2), $\mathsf{S}_{\texttt{TCP,MR}}$ induces the weighted transductive prediction region

$$\Gamma^{\epsilon}_{\texttt{TCP,WMR}}\left(\mathbf{X}^{(n+1)}\right) = \left\{\mathbf{y} \in \mathcal{Y} : \mathsf{S}_{\texttt{TCP,MR}}\left(\mathbf{X}^{(n+1)}, \mathbf{y}\right) \leq q^{(1-\epsilon)}_{\texttt{TCP}}\left(\mathbf{y}\right)\right\} \tag{S17}$$

with the threshold

$$q^{(1-\epsilon)}_{\texttt{TCP}}\left(\mathbf{y}\right) = \mathsf{Q}_{1-\epsilon}\left(\sum_{i=1}^{n+1} \tilde{w}^{(i)} \cdot \delta_{A^{(i)}_{\texttt{TCP,MR}}}\right), \tag{S18}$$

which depends on $\mathbf{y}$ because the maximum ranks of Equation (S14) do. This is the full conformal region of Barber et al. (2023, Equation (12)), with the maximum ranks in place of the residuals. Barber et al. (2023, Appendix A) show that their Theorem 2 holds for such scores if they are computed by an algorithm that treats the $n+1$ instances symmetrically. $\mathsf{S}_{\texttt{TCP,MR}}$ does, because, by (A1), the model $\hat{\mathbf{f}}$ does not depend on the instances, and the maximum ranks depend on them only through the multiset of their dimension-wise scores. Swapping instances i and $n+1$ thus swaps the ith and the $(n+1)$st maximum ranks, and Theorem 2 of Barber et al. (2023) gives

$$(1-\epsilon) - \mathsf{P}\left(\mathbf{Y}^{(n+1)} \in \Gamma^{\epsilon}_{\texttt{TCP,WMR}}\left(\mathbf{X}^{(n+1)}\right)\right) \leq \sum_{i=1}^{n} \tilde{w}^{(i)} \mathsf{d}_{\mathsf{TV}}\left(\mathbf{A}_{\texttt{TCP,MR}}, \mathbf{A}^{\rightleftarrows i}_{\texttt{TCP,MR}}\right), \tag{S19}$$

where $\mathbf{A}_{\texttt{TCP,MR}} = (A^{(i)}_{\texttt{TCP,MR}})_{i=1}^{n+1}$ and $\mathbf{A}^{\rightleftarrows i}_{\texttt{TCP,MR}}$ is the same vector with the ith and the $(n+1)$st elements swapped. With equal weights, $\Gamma^{\epsilon}_{\texttt{TCP,WMR}}$ is the transductive `Max-Rank` region of Schlembach et al. (2026). Since its threshold depends on $\mathbf{y}$, it is, however, intractable to compute $\Gamma^{\epsilon}_{\texttt{TCP,WMR}}$ in general. Therefore, we use the inductive region $\Gamma^{\epsilon}_{\texttt{ICP,WMR}}$ of Theorem 2 instead, whose threshold, the weighted quantile of Barber et al. (2023, Equation (11)),

$$q^{(1-\epsilon)}_{\texttt{ICP}} = \mathsf{Q}_{1-\epsilon}\left(\sum_{i=1}^{n} \tilde{w}^{(i)} \cdot \delta_{A^{(i)}_{\texttt{ICP,MR}}} + \tilde{w}^{(n+1)} \cdot \delta_{+\infty}\right), \tag{S20}$$

does not depend on $\mathbf{y}$.

To show that the coverage gap of the inductive variant $\Gamma^{\epsilon}_{\texttt{ICP,WMR}}$ is bounded in the same way as the transductive one we compare the cumulative distribution functions (CDF) built from the transductive and inductive sets of maximum ranks next.

Lemma S6 *For any threshold $t \in \mathbb{R}$, let*

$$\mathsf{F}_{ICP,MR}(t) = \sum_{\substack{i \in \{1,\ldots,n\}: \\ A^{(i)}_{ICP,MR} \leq t}} \tilde{w}^{(i)} \quad \text{and} \quad \mathsf{F}_{TCP,MR}(t) = \sum_{\substack{i \in \{1,\ldots,n+1\}: \\ A^{(i)}_{TCP,MR} \leq t}} \tilde{w}^{(i)} \tag{S21}$$

be the total weight of the maximum ranks that do not exceed t in the ICP and TCP variants of `WMax-Rank`. Then

$$\mathsf{F}_{ICP,MR}(t-1) \leq \mathsf{F}_{TCP,MR}(t). \tag{S22}$$

Proof *Lemma 8 of Schlembach et al. (2026) tells us that $A^{(i)}_{TCP,MR} \leq A^{(i)}_{ICP,MR} + 1$ for $i \in \{1,\ldots,n\}$. Therefore, if $A^{(i)}_{ICP,MR} \leq t-1$, then $A^{(i)}_{TCP,MR} \leq t$, i.e., every calibration instance counted in $\mathsf{F}_{ICP,MR}(t-1)$ is also counted in $\mathsf{F}_{TCP,MR}(t)$, with the same weight $\tilde{w}^{(i)}$. $\mathsf{F}_{TCP,MR}(t)$ may additionally count the test instance $n+1$, whose weight $\tilde{w}^{(n+1)}$ is non-negative.* ■

$\mathsf{F}_{\texttt{ICP},\texttt{MR}}$ and $\mathsf{F}_{\texttt{TCP},\texttt{MR}}$ are the CDFs of the two distributions in Equations (S20) and (S18). The point mass of the ICP variant at $+\infty$ does not contribute at any finite t. Both distributions have finitely many point masses, so their CDFs are right-continuous step functions. This means that $\mathsf{F}_{\texttt{TCP},\texttt{MR}}(t)$ and $\mathsf{F}_{\texttt{ICP},\texttt{MR}}(t)$ are step functions that include the point mass at t itself. The weighted quantile is the smallest maximum rank at which the CDF reaches $1-\epsilon$. In particular, the CDF at the weighted quantile is at least $1-\epsilon$. Hence, for any $t \in \mathbb{R}$,

$$q_{\texttt{ICP}}^{(1-\epsilon)} \leq t \iff \mathsf{F}_{\texttt{ICP},\texttt{MR}}(t) \geq 1-\epsilon, \tag{S23}$$

and

$$q_{\texttt{TCP}}^{(1-\epsilon)}(\mathbf{y}) \leq t \iff \mathsf{F}_{\texttt{TCP},\texttt{MR}}(t) \geq 1-\epsilon. \tag{S24}$$

In words, the weighted quantile is at most t if and only if the maximum ranks that do not exceed t carry a total weight of at least $1-\epsilon$.[2]

We are now in a position to describe the relationship between the weighted quantile functions of the ICP and TCP variants of the maximum rank scores.

Lemma S7 *The weighted quantiles of the maximum ranks in the ICP and TCP variants of* ***WMax-Rank*** *are related by*

$$q_{TCP}^{(1-\epsilon)}(\mathbf{y}) \leq q_{ICP}^{(1-\epsilon)} + 1. \tag{S25}$$

Proof *If $\epsilon < \tilde{w}^{(n+1)}$, then $q_{ICP}^{(1-\epsilon)} = +\infty$ and Equation (S25) holds trivially. Otherwise, Equation (S23) with $t = q_{ICP}^{(1-\epsilon)}$ gives $\mathsf{F}_{ICP,MR}(q_{ICP}^{(1-\epsilon)}) \geq 1-\epsilon$. Applying Lemma S6 with $t = q_{ICP}^{(1-\epsilon)} + 1$ yields*

$$1-\epsilon \leq \mathsf{F}_{ICP,MR}\left(q_{ICP}^{(1-\epsilon)}\right) \leq \mathsf{F}_{TCP,MR}\left(q_{ICP}^{(1-\epsilon)} + 1\right), \tag{S26}$$

and Equation (S24) with $t = q_{ICP}^{(1-\epsilon)} + 1$ gives the claim. ■

Armed with Lemmas S6 and S7 we are now able to prove the upper bound of the coverage gap for `WMax-Rank` stated in Theorem 2.

Proof of Theorem 2 To prove Equation (11), it is sufficient to show that

$$\Gamma^{\epsilon}_{\texttt{TCP},\texttt{WMR}}\left(\mathbf{X}^{(n+1)}\right) \subseteq \Gamma^{\epsilon}_{\texttt{ICP},\texttt{WMR}}\left(\mathbf{X}^{(n+1)}\right), \tag{S27}$$

because then

$$\mathsf{P}\left(\mathbf{Y}^{(n+1)} \in \Gamma^{\epsilon}_{\texttt{TCP},\texttt{WMR}}\left(\mathbf{X}^{(n+1)}\right)\right) \leq \mathsf{P}\left(\mathbf{Y}^{(n+1)} \in \Gamma^{\epsilon}_{\texttt{ICP},\texttt{WMR}}\left(\mathbf{X}^{(n+1)}\right)\right), \tag{S28}$$

and Equation (S19) gives Equation (11). Replacing the regions with their respective definitions, Equation (S27) reads

2. For equal weights $\tilde{w}^{(i)} = 1/(n+1)$, we have $\mathsf{F}_{\texttt{ICP},\texttt{MR}}(t) = N_{\texttt{ICP},\texttt{MR}}(t)/(n+1)$ and $\mathsf{F}_{\texttt{TCP},\texttt{MR}}(t) = N_{\texttt{TCP},\texttt{MR}}(t)/(n+1)$, and Equations (S23) and (S24) reduce to Equations (34) and (35) of Schlembach et al. (2026) with $k = \lceil (1-\epsilon)(n+1) \rceil$.

$$\left\{\mathbf{y} \in \mathcal{Y} : \mathsf{S}_{\texttt{TCP,MR}}\left(\mathbf{X}^{(n+1)}, \mathbf{y}\right) \leq q_{\texttt{TCP}}^{(1-\epsilon)}\left(\mathbf{y}\right)\right\} \subseteq \left\{\mathbf{y} \in \mathcal{Y} : \mathsf{S}_{\texttt{ICP,MR}}\left(\mathbf{X}^{(n+1)}, \mathbf{y}\right) \leq q_{\texttt{ICP}}^{(1-\epsilon)}\right\}. \quad \text{(S29)}$$

The nonconformity score functions $\mathsf{S}_{\texttt{TCP,MR}}$ and $\mathsf{S}_{\texttt{ICP,MR}}$ are not affected by the weights $\tilde{w}^{(i)}$. Therefore, Lemma 11 of Schlembach et al. (2026) applies and

$$\mathsf{S}_{\texttt{TCP,MR}}\left(\mathbf{X}^{(n+1)}, \mathbf{y}\right) = \mathsf{S}_{\texttt{ICP,MR}}\left(\mathbf{X}^{(n+1)}, \mathbf{y}\right) + 1. \quad \text{(S30)}$$

Substituting Equation (S30) in Equation (S29) yields

$$\left\{\mathbf{y} \in \mathcal{Y} : \mathsf{S}_{\texttt{ICP,MR}}\left(\mathbf{X}^{(n+1)}, \mathbf{y}\right) + 1 \leq q_{\texttt{TCP}}^{(1-\epsilon)}\left(\mathbf{y}\right)\right\} \subseteq \left\{\mathbf{y} \in \mathcal{Y} : \mathsf{S}_{\texttt{ICP,MR}}\left(\mathbf{X}^{(n+1)}, \mathbf{y}\right) \leq q_{\texttt{ICP}}^{(1-\epsilon)}\right\}, \quad \text{(S31)}$$

which holds if

$$q_{\texttt{TCP}}^{(1-\epsilon)}\left(\mathbf{y}\right) - 1 \leq q_{\texttt{ICP}}^{(1-\epsilon)}. \quad \text{(S32)}$$

Lemma S7 proves this claim.

Note that the proof requires neither properties of the weights beyond (A2) nor distinct scores. ■

Remark S8 *Since $q_{ICP}^{(1-\epsilon)}$ does not depend on $\mathbf{y}$, Algorithm S1 computes the region of Theorem 2 with one threshold per dimension instead of testing individual labels. If $q = q_{ICP}^{(1-\epsilon)} < n$, a dimension-wise score has rank at most q if and only if it is smaller than the $(q+1)$st smallest calibration score $\beta_j^{(\pi_j^{(q+1)})}$. Otherwise, the region equals $\mathcal{Y}$. Line 15 of Algorithm S1 uses $\leq \beta_j^{(\pi_j^{(q+1)})}$ instead of $< \beta_j^{(\pi_j^{(q+1)})}$, so the region it returns, which we also denote by $\Gamma^{\epsilon}_{ICP,WMR}\left(\mathbf{X}^{(n+1)}\right)$, additionally contains the labels whose dimension-wise score equals $\beta_j^{(\pi_j^{(q+1)})}$ in some dimension. Therefore, its coverage is at least as large, and the bound of Theorem 2 holds for it as well. For the other methods, such as `nmtCP` and `WDCP-MT`, whose thresholds are quantiles of the scores themselves, closed intervals are exact. Algorithm S1 uses closed intervals, like the implementation of `Max-Rank` in Schlembach et al. (2026), so that all methods return intervals of the same form.*

Tied scores. Algorithm S1 ranks a score by the number of calibration scores of the same dimension at or below it, as in the proof above, so that tied scores share the largest of their ranks. The `python` listings of Timans et al. (2025) and Schlembach et al. (2026) instead compute ranks with `np.argsort(np.argsort(scores, axis=0), axis=0)`, which gives tied scores distinct ranks in the order of the sort. For distinct scores, which both papers assume, the two ranks coincide. With ties, as in the first section of ELEC2 (Section F.5), the listings assign tied calibration scores smaller ranks than the analysis covers, so that the threshold, and with it the prediction region, can be too small. `scipy.stats.rankdata(scores, method="max", axis=0)` would compute the ranks of Algorithm S1 also for tied scores.

Appendix C. Practical Implications of the Coverage-Gap Bounds

The total variation distances in Theorems 1, S2, and 2 cannot be computed from data, so the bounds do not certify a coverage level for a given application. This holds for all coverage-gap bounds in the framework of Barber et al. (2023). The bounds identify the conditions under which the methods are valid and the mechanisms by which validity degrades. This section summarises what follows from them for a practitioner.

What the bounds guarantee. If the nonconformity scores are close to exchangeable around the test instance, i.e., their distribution is stable over the weighted part of the calibration window and the dependence between nearby instances is weak, the total variation distances are small and the coverage of all three methods is close to $1-\epsilon$. For `WDCP-MT` this holds for every split.

The assumptions of the theorems also constrain the implementation. The underlying model must be fitted on a separate training block, the weights must be fixed in advance with the test instance receiving the largest weight, and for `WDCP-MT` the split must be drawn independently of the data and only the scores in E may enter the selection of the dimension-wise significance levels. Violating any of these voids the guarantees, whatever the data.

Two sources of a large coverage gap. The total variation distances grow for two reasons, which call for different remedies.

- *Distribution shift* (drift, changepoints): the test instance differs from older calibration instances. This affects all three methods and is addressed by decaying weights and the sliding calibration window.
- *Serial dependence*: swapping the test instance with a calibration instance changes the relation of both to their neighbours in time, which increases the distances for all three methods. `WDCP-MT` is additionally exposed to leakage, since under dependence between E and O the fit of the selected dimension-wise significance levels to noise in E carries over to O (Remark S5). This can be reduced by a split that leaves less dependence between E and O or avoided by a method without a separate selection step.

Choosing a method. Neither `nmtCP` nor `WMax-Rank` selects dimension-wise significance levels on part of the calibration set, so the leakage between E and O does not arise for them. `WDCP-MT` obtains its gains through this selection step, which also halves its calibration set and, for strongly autocorrelated scores with a random split, exposes it to leakage. In the experiments, it produces smaller regions than `nmtCP` only for uncorrelated scores whose marginal distributions differ in shape. This complements the operating regimes motivated in the main paper (weakly correlated scores whose marginal distributions differ in shape for `WDCP-MT`, positively correlated scores, i.e., overlapping miscoverage events across dimensions, for `WMax-Rank`), which are expectations supported by the experiments rather than consequences of the bounds.

What to monitor. Since the bounds cannot be evaluated, they should be complemented by empirical checks, which the online setting permits because the true labels become available over time.

- *Rolling empirical coverage*, which measures directly what the bounds only cap.
- *Autocorrelation of the nonconformity scores at short lags.* If it is strong, a random split is risky for `WDCP-MT`.
- *Sensitivity to the seed of the split* for `WDCP-MT`.

If the rolling coverage falls below $1-\epsilon$, adapting the significance level online (Gibbs and Candès, 2021) restores long-run coverage regardless of why the bound failed.

Appendix D. Implementation and Computational Complexity

`WDCP-MT` selects its dimension-wise significance levels by solving an allocation problem on the selection set E (Algorithm 1 of the main paper). Schlembach et al. (2025b) solve the unweighted counterpart with a mixed-integer linear program (MILP). `WDCP-MT` instead uses a dynamic program (DP) that solves the weighted problem up to an arbitrarily small loss of error budget, serves all global significance levels from a single table, and has a running time polynomial in $|E|$ and m. The experiments use the DP, and the MILP serves as the baseline in the comparison of complexity and runtime below. The DP selects the dimension-wise significance levels from σ and $\mathbf{B}_E$ only and satisfies $\sum_j \epsilon_j \le \epsilon$ exactly, so (A4) of Theorem 1 holds and the coverage-gap bound is unchanged. More generally, the objective and the solver only affect the efficiency of the region, not its validity (see the remark after (A4) in Section A). This section states the allocation problem, gives the DP, compares the computational complexity of all methods, and reports measured runtimes.

D.1. The Allocation Problem

Let π_j sort the scores $\{\beta_j^{(i)}\}_{i\in E}$ in ascending order, and let $\epsilon_j^{(k)} = 1 - \sum_{l=1}^{k} \tilde{w}_E^{(\pi_j^{(l)})}$ be the candidate dimension-wise significance level of rank $k \in \{1,\dots,|E|\}$, as in Algorithm 1. Selecting rank k in dimension j costs $c_j^{(k)} = \log\max\big(\beta_j^{(\pi_j^{(k)})}, 10^{-12}\big)$. `WDCP-MT` solves

$$V^\star(\epsilon) = \min_{k_1\in\mathcal{K}_1,\dots,k_m\in\mathcal{K}_m} \sum_{j=1}^{m} c_j^{(k_j)} \quad \text{s.t.} \quad \sum_{j=1}^{m} \epsilon_j^{(k_j)} \le \epsilon, \qquad \mathcal{K}_j = \big\{k : \epsilon_j^{(k)} \ge \tilde{w}_O^{(n+1)}\big\}, \tag{S33}$$

with $V^\star(\epsilon) = +\infty$ and $\Gamma^\epsilon(\mathbf{X}^{(n+1)}) = \mathcal{Y}$ if no allocation is feasible.

- *Objective.* For absolute residuals the interval of dimension j has width $2q_j^{(1-\epsilon_j)}$, so $\sum_j c_j^{(k_j)}$ is, up to the constant $m\log 2$, the logarithm of the volume of the region estimated on E. The logarithm is strictly increasing and therefore has the same minimiser as the volume. Using the logarithm only makes the objective additive over the dimensions, which both the MILP (linear objective) and the DP (additive recurrence) require. For other nonconformity scores, the objective is a log-volume only if the interval width is proportional to the threshold, and any other additive measure of size (e.g. the sum of the side lengths) could be used instead. All experiments use absolute residuals. Both solvers replace scores below 10^{-12} by 10^{-12} so that the logarithm is finite.

- *Excluded levels.* A level $\epsilon_j < \tilde{w}_O^{(n+1)}$ is feasible on E but yields an infinite threshold on O, because the finite scores in O carry a weight of only $1 - \tilde{w}_O^{(n+1)}$. The optimiser would select such levels often, since the largest score in E is the cheapest way to free budget for the other dimensions. Therefore, $\mathcal{K}_j$ excludes them. The bound $\tilde{w}_O^{(n+1)}$ depends only on the split and the weights, both independent of the scores, so (A4) still holds.

- *When the region is $\mathcal{Y}$.* The levels decrease in k, so the smallest selectable level of dimension j is $\underline{\epsilon}_j = \min_{k \in \mathcal{K}_j} \epsilon_j^{(k)} \geq \max\bigl(\tilde{w}_E^{(n+1)}, \tilde{w}_O^{(n+1)}\bigr)$, and `WDCP-MT` returns $\mathcal{Y}$ exactly when $\epsilon < \sum_j \underline{\epsilon}_j$. For constant weights and $|O| = |E|$ this threshold is $m/(|E|+1)$ in exact arithmetic, about twice that of `nmtCP` ($m/(n+1)$), because `WDCP-MT` normalises the weights over half of the calibration set. For $n = 408$ and $m = 4$, finite regions require $\epsilon \geq 0.020$ for `WDCP-MT` and $\epsilon \geq 0.010$ for `nmtCP`. In floating-point arithmetic, the level $1/(|E|+1)$ can fall just below $\tilde{w}_O^{(n+1)}$ and be excluded, which doubles the threshold. This happens on ELEC2 ($n = 500$), where `WDCP-MT` with constant weights returns $\mathcal{Y}$ for $\epsilon < 0.032$. Decaying weights raise the weight of the point mass at $+\infty$ and with it the threshold.

A multiple-choice knapsack problem. Problem (S33) is a multiple-choice knapsack problem (Kellerer et al., 2004). The dimensions are the classes, the admissible ranks are the items, exactly one item is chosen per class, item k of class j has the weight $\epsilon_j^{(k)}$ and the cost $c_j^{(k)}$, and the capacity is ϵ. Within each class the cost is non-decreasing and the weight non-increasing in k. A higher rank buys error budget with volume. For constant weights, $\epsilon_j^{(k)} = 1 - k/(|E|+1)$, the budget constraint becomes an integer constraint on $\sum_j k_j$, and a DP over integer rank budgets solves the problem exactly. With unequal weights the levels are real numbers that depend on the dimension-specific order π_j. Scaling them to integers is exact only if all weights are integer multiples of a common unit, which fails, e.g., for exponential weights, whose pairwise ratios are irrational. The DP below therefore rounds the levels to a grid whose resolution is chosen directly by the user.

D.2. Ceil-Rounded Dynamic Program

Fix the grid unit $\eta = 1/(L(|E|+1))$ with $L \in \mathbb{N}$ ($L = 100$ in the experiments), round the candidate levels up to integer multiples of η, and round the capacity down, which gives

$$\omega_j^{(k)} = \lceil \epsilon_j^{(k)}/\eta \rceil, \qquad C(\epsilon) = \lfloor \epsilon/\eta \rfloor. \tag{S34}$$

Rounding the cumulative level $\epsilon_j^{(k)}$ rather than the individual weights keeps the rounding error per dimension below η, independently of $|E|$. Small weights cause no numerical difficulty, because their contributions fall into the same grid cell. The rounded problem replaces the constraint of (S33) by $\sum_j \omega_j^{(k_j)} \leq C(\epsilon)$ and is solved by the recurrence

$$G_0(c) = 0, \qquad G_j(c) = \min_{k \in \mathcal{K}_j:\ \omega_j^{(k)} \leq c} \Bigl(G_{j-1}\bigl(c - \omega_j^{(k)}\bigr) + c_j^{(k)} \Bigr), \qquad c = 0, \ldots, C, \tag{S35}$$

with $\min \emptyset = +\infty$ and $C = C(\epsilon_{\max})$ for the largest requested global significance level $\epsilon_{\max}$. Because G_0 vanishes for every $c \geq 0$, $G_j(c)$ is the minimal cost of the first j dimensions

with a total rounded level of *at most* c. A single table therefore serves all n_ϵ requested global significance levels. For a requested global significance level ϵ, the minimal cost of the rounded problem is $G_m(C(\epsilon))$, and the allocation that attains it is recovered by backtracking the stored minimisers from $(m, C(\epsilon))$. The DP returns the exact levels $\epsilon_j^{(\hat{k}_j)}$, not their rounded values, and these are used for the thresholds on O. Algorithm 1 of the main paper selects its ranks with Algorithm S2.

Proposition S9 *Let $\hat{k} = (\hat{k}_1, \ldots, \hat{k}_m)$ be the allocation returned by (S35) for the global significance level ϵ, and $\hat{G} = G_m(C(\epsilon))$ its cost.*

(i) Validity. *If $\hat{G} < +\infty$, then $\sum_{j=1}^m \epsilon_j^{(\hat{k}_j)} \leq \epsilon$.*

(ii) Near-optimality. *$V^\star(\epsilon) \leq \hat{G} \leq V^\star(\epsilon - m\eta)$.*

(iii) Exactness. *If $\epsilon_j^{(k)}/\eta \in \mathbb{N}$ for all j and $k \in \mathcal{K}_j$, e.g. for constant weights, then $\hat{G} = V^\star(\epsilon)$.*

Proof (i) $\epsilon_j^{(k)} \leq \eta\, \omega_j^{(k)}$ for all j and k, so $\sum_j \epsilon_j^{(\hat{k}_j)} \leq \eta \sum_j \omega_j^{(\hat{k}_j)} \leq \eta\, C(\epsilon) \leq \epsilon$. (ii) The lower bound follows from (i), since $\hat{k}$ is feasible for (S33). For the upper bound, let k satisfy $\sum_j \epsilon_j^{(k_j)} \leq \epsilon - m\eta$. Since $\lceil x \rceil < x + 1$, $\sum_j \omega_j^{(k_j)} < \sum_j \epsilon_j^{(k_j)}/\eta + m \leq \epsilon/\eta$, and since the left-hand side is an integer, $\sum_j \omega_j^{(k_j)} \leq C(\epsilon)$. So k is feasible for the rounded problem, whose optimum $\hat{G}$ is therefore at most the cost of k. (iii) The rounding is the identity, so the two constraints coincide. For constant weights, $\epsilon_j^{(k)}/\eta = L(|E| + 1 - k)$. ■

Part (i) is what the coverage guarantee needs, and part (ii) bounds the price of the rounding. The region is at most as large, in the volume estimated on E, as the optimal region for the budget $\epsilon - m\eta$. With $L = 100$ this loss is $m/100$ of the budget of a single rank, $1/(|E| + 1)$.

Proposition S10 *Fix the split and the scores in E, let $\epsilon' \in (0, \epsilon]$, and for each dimension j let k_j^{B} be the smallest rank k with $\epsilon_j^{(k)} \leq \epsilon'/m$, i.e., the rank of the threshold that the Bonferroni correction at level ϵ' obtains from the scores in E with the weights $\tilde{w}_E$. If these ranks exist and $k_j^{\mathrm{B}} \in \mathcal{K}_j$ for all j, then $V^\star(\epsilon') \leq \sum_{j=1}^m c_j^{(k_j^{\mathrm{B}})}$. Hence an exact solver ($\epsilon' = \epsilon$) and the DP ($\epsilon' = \epsilon - m\eta$, or $\epsilon' = \epsilon$ under the condition of Proposition S9 (iii)) select an allocation whose cost is at most that of the Bonferroni correction applied to E.*

Proof $\sum_j \epsilon_j^{(k_j^{\mathrm{B}})} \leq \epsilon'$, so the ranks k_j^{B} are feasible for (S33) at level ϵ'. The statement for the DP follows from Proposition S9 (ii) and (iii). ■

For absolute residuals, the cost is the logarithm of the volume estimated from the scores in E up to a constant, so on E the selected region is never larger than that of the Bonferroni correction. This does not carry over to the regions reported in the experiments. `WDCP-MT` forms its region from the scores in O, about half of the calibration set, whereas `nmtCP` uses all of them, and the selected levels are optimal for E but not necessarily for O (Section A.4).

Algorithm S2: Selection of the dimension-wise significance levels of `WDCP-MT` by dynamic programming

Input: scores $\{\beta_j^{(i)}\}_{i\in E}$ for $j = 1, \dots, m$; normalised weights $\{\tilde{w}_E^{(i)}\}_{i\in E}$; mass $\tilde{w}_O^{(n+1)}$; global significance levels $\epsilon^1, \dots, \epsilon^{n_\epsilon}$; resolution L

Output: dimension-wise significance levels $(\epsilon_1, \dots, \epsilon_m)$ for each ϵ^h, or $\mathcal{Y}$

$\eta \leftarrow 1/(L(|E|+1))$; $C \leftarrow \max\{c \in \mathbb{N} : c\,\eta \le \max_h \epsilon^h\}$;
for $j = 1, \dots, m$ **do**
  Sort $\{\beta_j^{(i)}\}_{i\in E}$ in ascending order using permutation π_j;
  for $k = 1, \dots, |E|$ **do**
    $\epsilon_j^{(k)} \leftarrow 1 - \sum_{l=1}^{k} \tilde{w}_E^{(\pi_j^{(l)})}$; $c_j^{(k)} \leftarrow \log \max\big(\beta_j^{(\pi_j^{(k)})}, 10^{-12}\big)$;
    $\omega_j^{(k)} \leftarrow \lceil \epsilon_j^{(k)}/\eta \rceil$ if $\epsilon_j^{(k)} \ge \tilde{w}_O^{(n+1)}$, else $\omega_j^{(k)} \leftarrow +\infty$;
  end
end
`// table in ''at most` c`'' form, built once for all` ϵ^h
$G_0(c) \leftarrow 0$ for $c = 0, \dots, C$;
for $j = 1, \dots, m$ **do**
  for $c = 0, \dots, C$ **do**
    $G_j(c) \leftarrow \min_{k:\,\omega_j^{(k)} \le c}\big(G_{j-1}(c - \omega_j^{(k)}) + c_j^{(k)}\big)$, and store the minimiser $\theta_j(c)$;
  end
end
for $h = 1, \dots, n_\epsilon$ **do**
  $c \leftarrow \max\{c \in \mathbb{N} : c\,\eta \le \epsilon^h\}$;
  while $c \ge 0$ **and** $G_m(c) < +\infty$ **do**
    `// backtrack`
    $b \leftarrow c$; **for** $j = m, \dots, 1$ **do**
      $\hat{k}_j \leftarrow \theta_j(b)$; $b \leftarrow b - \omega_j^{(\hat{k}_j)}$;
    end
    if $\sum_j \epsilon_j^{(\hat{k}_j)} \le \epsilon^h$ **then return** $(\epsilon_1^{(\hat{k}_1)}, \dots, \epsilon_m^{(\hat{k}_m)})$ *for* ϵ^h;
    $c \leftarrow c - 1$; `// exact check failed (floating point)`
  end
  return $\mathcal{Y}$ *for* ϵ^h;
end

D.3. Computational Complexity

Table S1 compares the cost of one update step, i.e., of re-fitting a method after each test instance for all n_ϵ global significance levels. The sliding calibration window and, for `WDCP-MT`, the new split require this at every step. `nmtCP` computes one weighted quantile per dimension and level, and `WMax-Rank` ranks the scores, takes one weighted quantile of the maximum ranks, and converts it into m dimension-wise levels per global level. Both computational costs are dominated by sorting. `WDCP-MT` additionally solves the allocation problem.

Table S1: Cost of one update step of each method for all n_ϵ global significance levels, with n calibration scores, $|E| \approx n/2$, m output dimensions, the largest requested level $\epsilon_{\max}$, and the grid resolution L of the DP. The thresholds on O add $\mathcal{O}(m|O|\log|O| + n_\epsilon m \log|O|)$ to both `WDCP-MT` solvers.

Method	Time	Space
`nmtCP`	$\mathcal{O}(mn\log n + n_\epsilon m\log n)$	$\mathcal{O}(mn)$
`WMax-Rank`	$\mathcal{O}(mn\log n + n_\epsilon m)$	$\mathcal{O}(mn)$
`WDCP-MT`, MILP	$\mathcal{O}(m\|E\|\log\|E\|)$ plus n_ϵ MILPs with $m\|E\|$ binary variables each (worst case exponential)	$\mathcal{O}(m\|E\|)$ per MILP
`WDCP-MT`, DP	$\mathcal{O}(m\|E\|\log\|E\| + Lm\|E\|^2\epsilon_{\max} + n_\epsilon m)$	$\mathcal{O}(Lm\|E\|\epsilon_{\max})$

- *MILP.* One MILP per global level, with $m|E|$ binary variables and $m+1$ constraints. The running time of branch and bound is not polynomially bounded. In the worst case it explores up to $2^{m|E|}$ nodes, and exhaustive search over the $|E|^m$ allocations is exponential in m. This is not an artefact of the formulation. The general multiple-choice knapsack problem is NP-hard, since it contains the 0–1 knapsack problem (Kellerer et al., 2004).

- *DP.* The hardness is weak, i.e., it disappears for integer capacities of polynomial size, which is what the rounding achieves. Each of the m rows of (S35) has $C+1 = \lfloor L(|E|+1)\epsilon_{\max}\rfloor + 1$ cells, each minimising over at most $|E|$ items, so the table costs $\mathcal{O}(Lm|E|^2\epsilon_{\max})$ time, independently of the weight function, and every level then costs an $\mathcal{O}(m)$ backtrack. The stored minimisers take $\mathcal{O}(Lm|E|\epsilon_{\max})$ space. With 8-byte integers and $\epsilon_{\max} = 0.5$, this amounts to 328 kB for the synthetic data sets ($|E| = 204$, $m = 4$), 402 kB for ELEC2 ($|E| = 250$, $m = 4$), and 602 kB for Tétouan City ($|E| = 250$, $m = 6$).

D.4. Measured Runtime

All methods ran in one queue on the same 8-core virtual server. After every test instance the predictor is updated. The calibration window slides by one instance, the split of `WDCP-MT` is redrawn, and the dimension-wise significance levels are re-selected for all $n_\epsilon = 9$ global levels up to $\epsilon_{\max} = 0.5$. The MILP solves the levels one after another with HiGHS, and the DP builds one table with $L = 100$ and reads all levels from it. `WDCP-MT` with the MILP was run for two replications per data set and weight function and all other methods for ten. Replication r uses the same data and split for both solvers. Table S2 reports the mean wall time of the update step per test instance, pooled over the three weight functions and the replications.

The DP is 100 to 119 times faster than the MILP. On Tétouan City, one pass over the 7 487 test instances took 5.1 hours with the MILP and 3.3 minutes with the DP. Summed over all runs, the prediction and update steps took 49.1 hours for the 36 replications with the MILP and 2.6 hours for the 180 replications with the DP. The time of the DP grows with $|E|$ and m roughly as the term $Lm|E|^2\epsilon_{\max}$ of Table S1 predicts. We measure a factor

Table S2: Mean wall time of the update step per test instance (re-selecting the dimension-wise significance levels for all $n_\epsilon = 9$ global levels), pooled over the weight functions and replications, for n calibration instances and m output dimensions.

(n, m)	CP, χ^2 $(408, 4)$	CP, corr. $(408, 4)$	Drift, χ^2 $(408, 4)$	Drift, corr. $(408, 4)$	ELEC2 $(500, 4)$	Tétouan $(500, 6)$
`nmtCP`	0.332 ms	0.328 ms	0.330 ms	0.336 ms	0.374 ms	0.389 ms
`WMax-Rank`	0.324 ms	0.329 ms	0.329 ms	0.331 ms	0.376 ms	0.377 ms
`WDCP-MT`, DP	10.4 ms	10.9 ms	10.5 ms	10.8 ms	14.8 ms	24.3 ms
`WDCP-MT`, MILP	1.22 s	1.29 s	1.22 s	1.28 s	1.51 s	2.44 s
MILP / DP	117×	119×	116×	118×	102×	100×

of 1.4 and predict $(250/204)^2 \approx 1.5$ when going from $|E| = 204$ to $|E| = 250$. For the shift from $m = 4$ to $m = 6$ we measure 1.6 and predict 1.5. The DP remains 30 to 65 times slower than `nmtCP` and `WMax-Rank`, whose update steps take a third of a millisecond and barely depend on n and m, since sorting dominates their cost.

Agreement of the two solvers. On the two replications that both solvers completed, they selected identical dimension-wise significance levels at 96% of the test steps on average with decaying weights (at least 87% for every combination of data set, weight function, and ϵ, the lowest on Tétouan City) and at 99.9% with constant weights (at least 98.9%), apart from an exact tie. If the levels of an optimal allocation sum to exactly ϵ, which happens for constant weights and $\epsilon(|E| + 1) \in \mathbb{N}$, their floating-point sum can exceed ϵ, and the exact check of the DP (Algorithm S2) then returns the best allocation with a strictly smaller sum. Excluding the tie, the error rates over the whole test set differ by at most 0.0012. At the tie (constant weights, $\epsilon = 0.2$ on the synthetic data sets) the levels never coincide: the DP selects levels one rank lower, its error rate is 0.003 to 0.005 lower, and its finite regions are on average 0.9% to 1.7% larger. The DP returned $\mathcal{Y}$ where the MILP returned a finite region for 59 pairs of test step and ϵ, all at the smallest level of the grid with finite regions for decaying weights ($\epsilon = 0.05$ for linear and $\epsilon = 0.1$ for exponential weights). For 58 of them the levels selected by the MILP left less than $m\eta$ of the budget unused, the rounding loss of Proposition S9(ii), and for one the MILP exceeded the budget. The MILP exceeded the budget by more than 10^{-12} for 49 of 630 755 pairs with finite regions, by at most $9.9 \cdot 10^{-7}$, because HiGHS enforces the budget only up to its feasibility tolerance. The DP never did.

Appendix E. Data Sets

This section describes the synthetic data sets in detail and displays all data sets used for the experiments in the paper.

Synthetic Data Sets The synthetic data sets are inspired by the simulations performed by Barber et al. (2023); Schlembach et al. (2025a,b). They consist of six time series gener-

ated as

$$o_{t,1} = \sin\left(\frac{\pi t}{12}\right), \qquad o_{t,2} = \sin\left(\frac{\pi t}{84}\right),$$
$$o_{t,3} = -\operatorname{sign}\left(\sin\left(\frac{t}{12}\right)\right), \qquad o_{t,4} = \operatorname{sign}(120 - \operatorname{mod}(t, 168)),$$
$$o_{t,5} = [o_{t,1}, o_{t,2}, o_{t,3}, o_{t,4}]\,\phi_{t,5} + e_{t,5}, \qquad o_{t,6} = [o_{t,1}, o_{t,2}, o_{t,3}, o_{t,4}]\,\phi_{t,6} + e_{t,6}.$$

Features $o_{t,5}$ and $o_{t,6}$ are linear combinations of the first four features, where t indexes the time steps of the underlying series. The data sets differ in coefficients $\phi_{t,5}$, $\phi_{t,6}$, and in noise terms $e_{t,5}$ and $e_{t,6}$. Each instance $\mathbf{x}^{(i)}$ consists of two consecutive time steps of the first four features and the corresponding label $\mathbf{y}^{(i)}$ consists of two consecutive time steps from $o_{t,5}$ and $o_{t,6}$, i.e., instance i covers time steps $2i-1$ and $2i$. The main text and the figure captions index instances. The coefficient specifications and the full-data-set figures below are given in terms of the underlying time steps. Therefore, e.g., the changepoint at time step 3264 corresponds to instance 1632.

We consider four simulation settings.

Synthetic data set 1: changepoint with correlated noise. A changepoint is introduced at time step 3264 (instance 1632). The coefficients are

$$\phi_{t,5} = \begin{cases} [0.7, 0.3, 0, 0]^\top & \text{if } t \in \{1, \dots, 3264\} \\ [0.2, 0.8, 0, 0]^\top & \text{if } t \in \{3265, \dots, 4896\}, \end{cases} \tag{S36}$$

$$\phi_{t,6} = \begin{cases} [0, 0, 0.6, 0.4]^\top & \text{if } t \in \{1, \dots, 3264\} \\ [0, 0, 0.2, 0.8]^\top & \text{if } t \in \{3265, \dots, 4896\}. \end{cases} \tag{S37}$$

The noise term $[e_{t,5},\ e_{t,6}]^\top \sim \mathcal{N}([0,0]^\top, \boldsymbol{\sigma})$, where $\boldsymbol{\sigma}$ is a 2×2 matrix with ones on the main diagonal and 0.5 elsewhere, scaled by a factor of 0.1. The proper training set contains the first 816 instances, the calibration set the next 408 instances, and the test set the final 1224 instances. Figure S1 displays the entire data set. The changepoint is highlighted by a dotted line.

Synthetic data set 2: changepoint with uncorrelated noise. This data set is identical to the synthetic data set 1 except for the noise term. $e_{t,5} \sim \mathcal{U}(-1, 1)$, while $e_{t,6}$ is drawn from a chi-squared distribution with two degrees of freedom, multiplied by an independent random sign, and rescaled to $[-1, 1]$ by dividing by the maximum absolute value over the series. The two components are sampled independently and both are scaled by a factor of 0.1. For brevity we denote this noise profile by $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$ in the figure captions. Figure S2 shows the full data set, with the changepoint marked by a dotted line.

Synthetic data set 3: distribution drift with correlated noise. Distribution drift starts at time step 3264 (instance 1632) and ends at time step 4896 (instance 2448). The coefficients are

$$\phi_{t,5} = \begin{cases} [0.7, 0.3, 0, 0]^\top & \text{if } t \in \{1, \dots, 3264\} \\ [0.7, 0.3, 0, 0]^\top + \frac{t-3264}{1632}[-0.5, 0.5, 0, 0]^\top & \text{if } t \in \{3265, \dots, 4896\} \\ [0.2, 0.8, 0, 0]^\top & \text{if } t \in \{4897, \dots, 6528\}, \end{cases} \tag{S38}$$

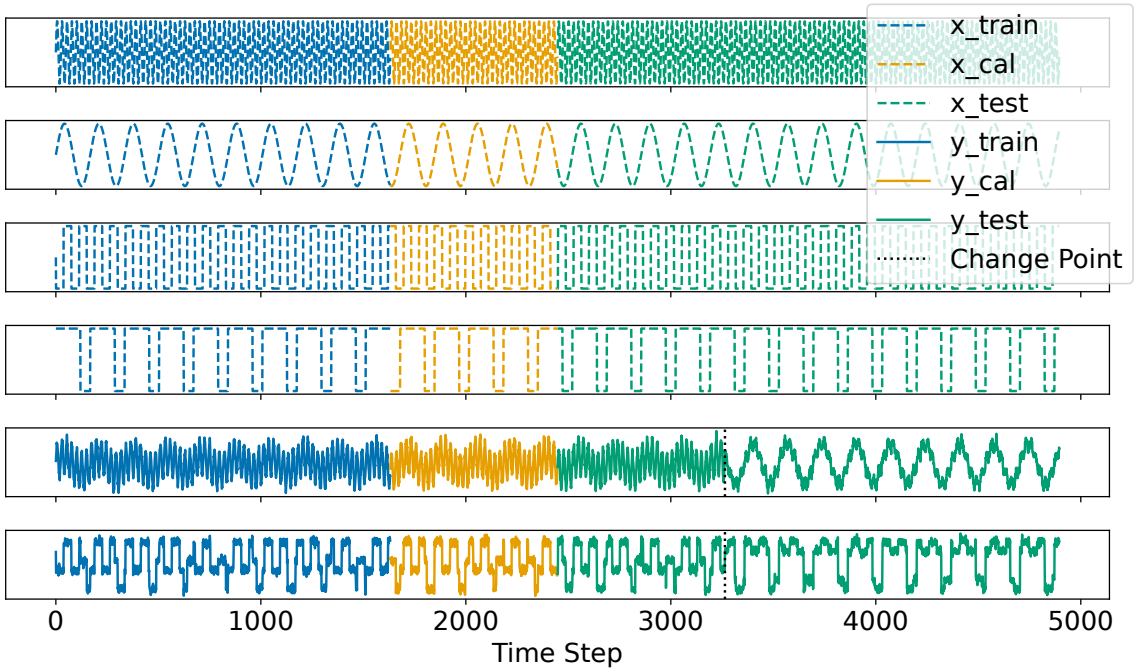


Figure S1: Synthetic data set 1: changepoint at instance 1632 with $e \sim \mathcal{N}([0,0]^\top, \boldsymbol{\sigma})$

$$\phi_{t,6} = \begin{cases} [0,0,0.6,0.4]^\top & \text{if } t \in \{1,\ldots,3264\} \\ [0,0,0.6,0.4]^\top + \frac{t-3264}{1632}[0,0,-0.4,0.4]^\top & \text{if } t \in \{3265,\ldots,4896\} \\ [0,0,0.2,0.8]^\top & \text{if } t \in \{4897,\ldots,6528\}. \end{cases} \tag{S39}$$

The noise term $[e_{t,5},\ e_{t,6}]^\top \sim \mathcal{N}([0,0]^\top, \boldsymbol{\sigma})$, where $\boldsymbol{\sigma}$ is a 2×2 matrix with ones on the main diagonal and 0.5 elsewhere, scaled by a factor of 0.1. The proper training set for the underlying model contains the first 816 instances, the calibration set the next 408 instances, and the test set the final 2040 instances. Figure S3 shows the full data set, with the start and end of the drift marked by dotted lines.

Synthetic data set 4: distribution drift with uncorrelated noise. This data set is identical to the synthetic data set 3 except for the noise term, which follows the uncorrelated noise profile of the synthetic data set 2. Figure S4 shows the full data set, with the start and end of the drift marked by dotted lines.

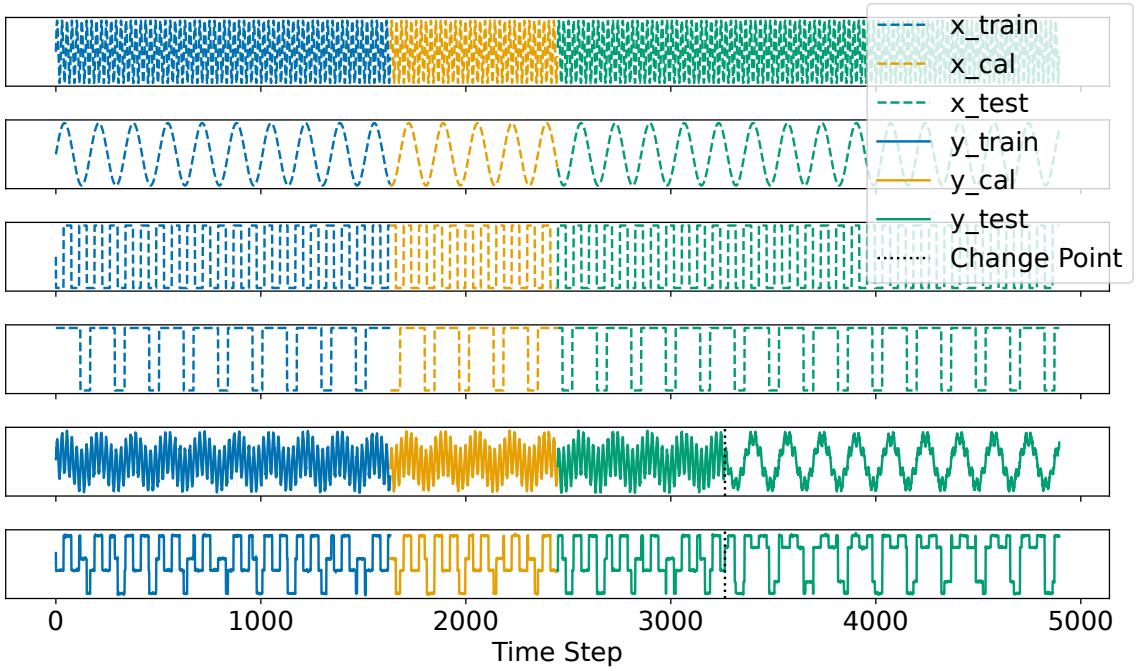


Figure S2: Synthetic data set 2: changepoint at instance 1632 with $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$

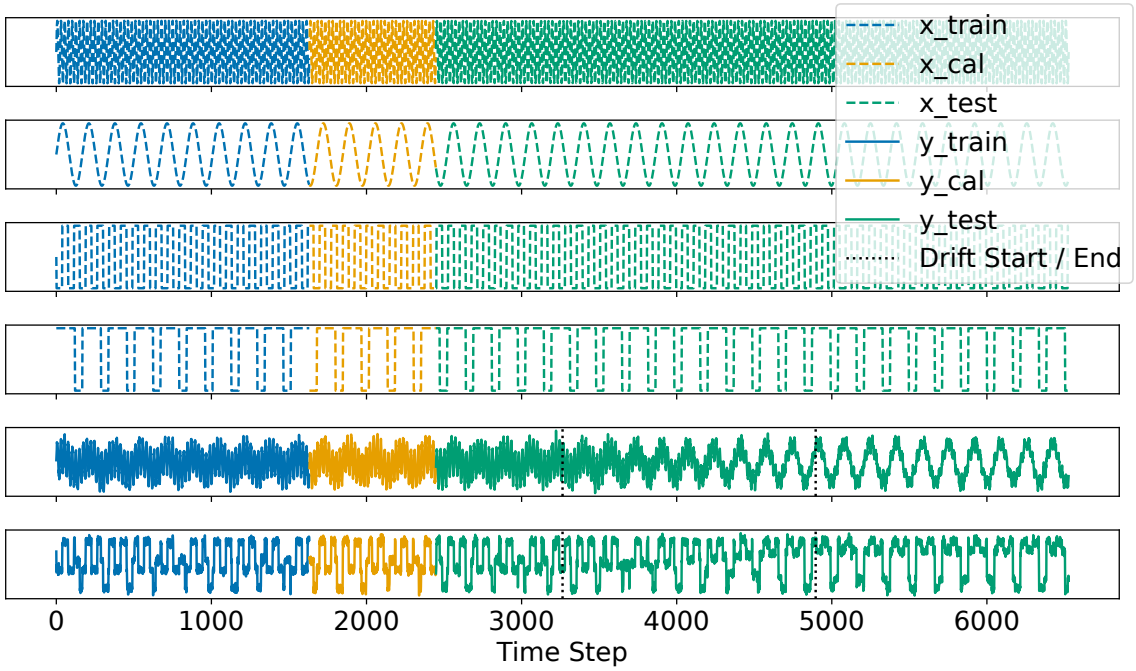


Figure S3: Synthetic data set 3: distribution drift starting at instance 1632 and ending at instance 2448 with $e \sim \mathcal{N}([0,0]^\top, \boldsymbol{\sigma})$

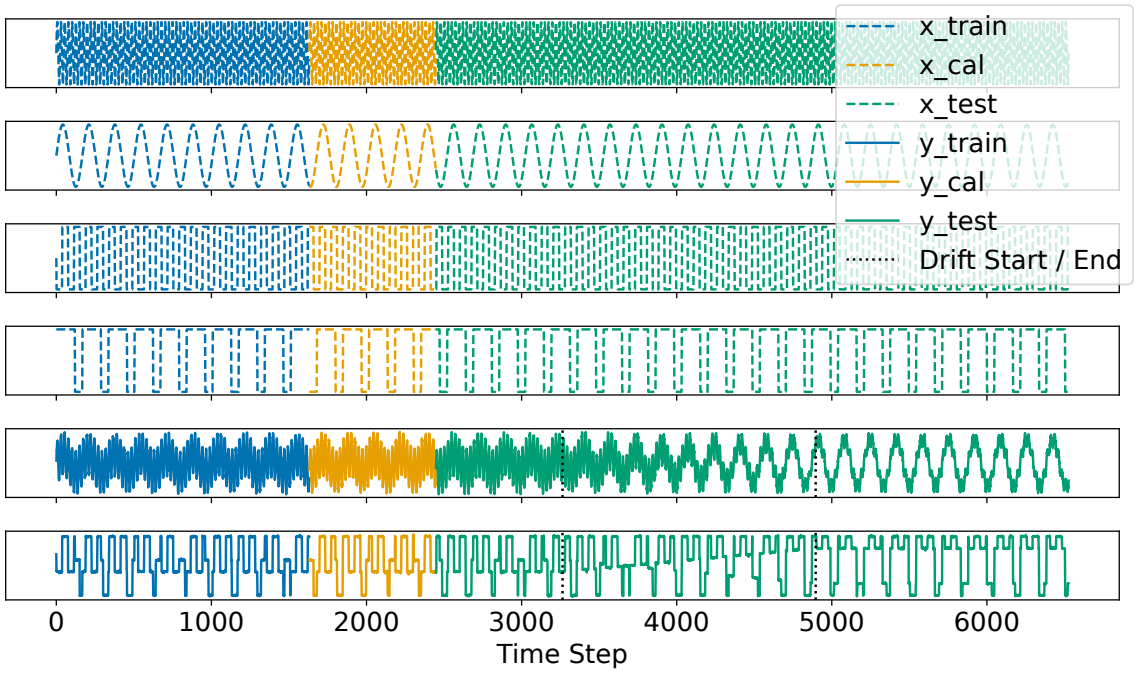


Figure S4: Synthetic data set 4: distribution drift starting at instance 1632 and ending at instance 2448 with $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$

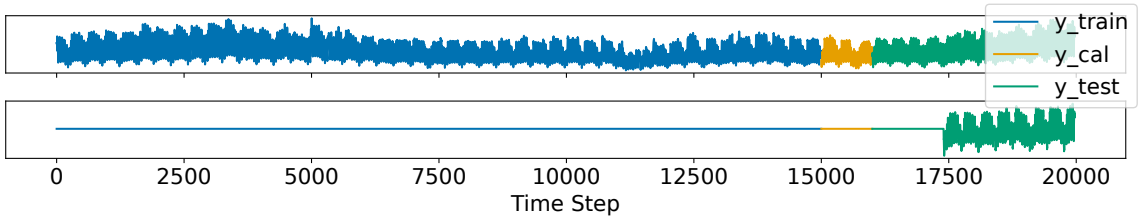


Figure S5: ELEC2 Data Set (Harries, 1999)

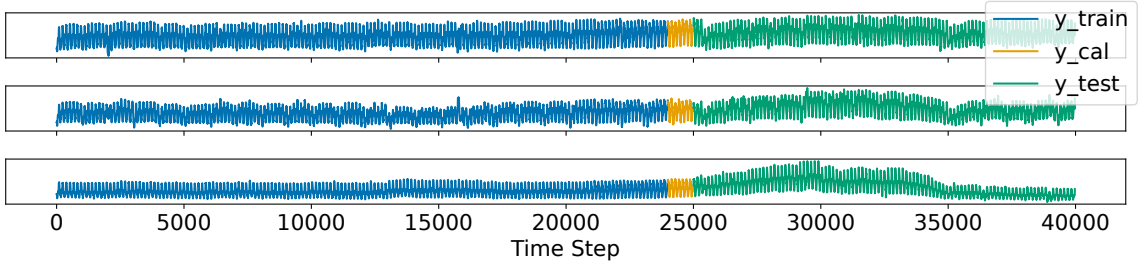


Figure S6: Tétouan City data set (Salam and Hibaoui, 2018)

Appendix F. Additional Experimental Results

F.1. Synthetic Change Point with Correlated Noise

Figure S7 displays the results from applying `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the Synthetic Change Point with Correlated Noise data set. The correlation of the nonconformity scores across dimensions is shown in Figure S8.

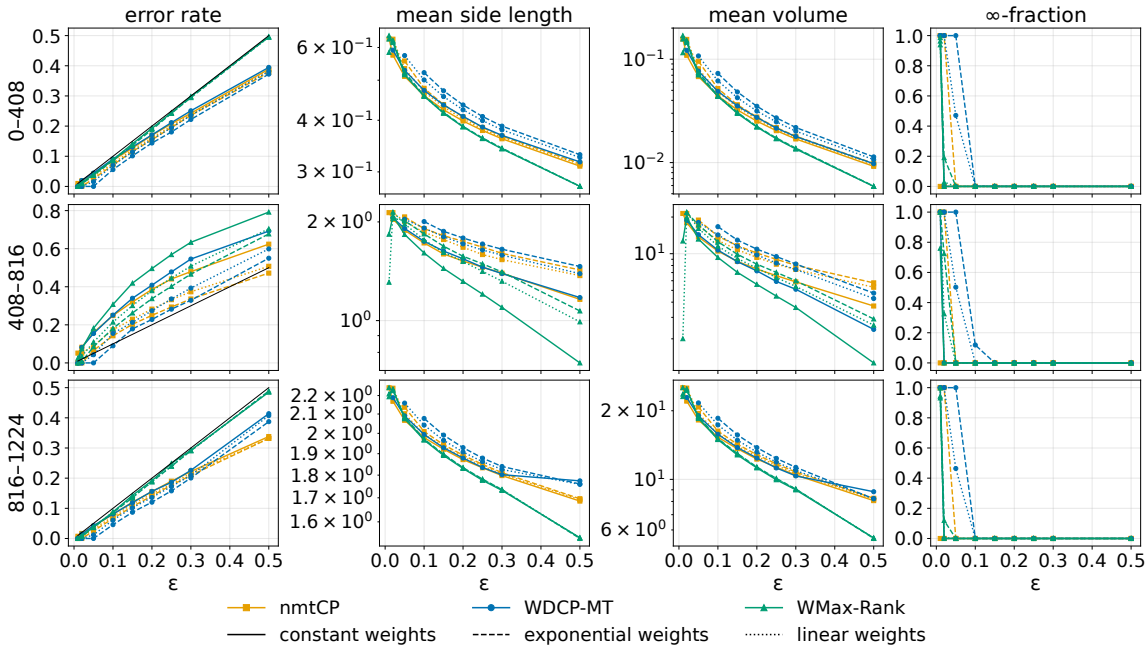


Figure S7: Mean error rate, mean hyperrectangle side length and mean hyperrectangle volume over the finite prediction regions, and fraction of prediction regions with at least one infinite bound (∞-fraction) as a function of the significance level ϵ for `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the synthetic changepoint data set with $e \sim \mathcal{N}([0,0]^\top, \boldsymbol{\sigma})$, shown for three consecutive sections of the test set.

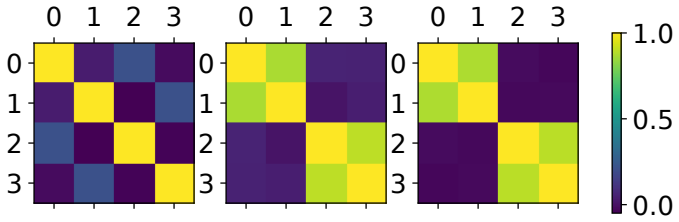


Figure S8: Correlation of the nonconformity scores across dimensions for the synthetic changepoint data set with $e \sim \mathcal{N}([0,0]^\top, \boldsymbol{\sigma})$ and the three sections of the test set.

F.2. Synthetic Change Point with Uncorrelated Noise

Figure S9 displays the results from applying `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the Synthetic Change Point with Uncorrelated Noise data set. The correlation of the nonconformity scores across dimensions is shown in Figure S10.

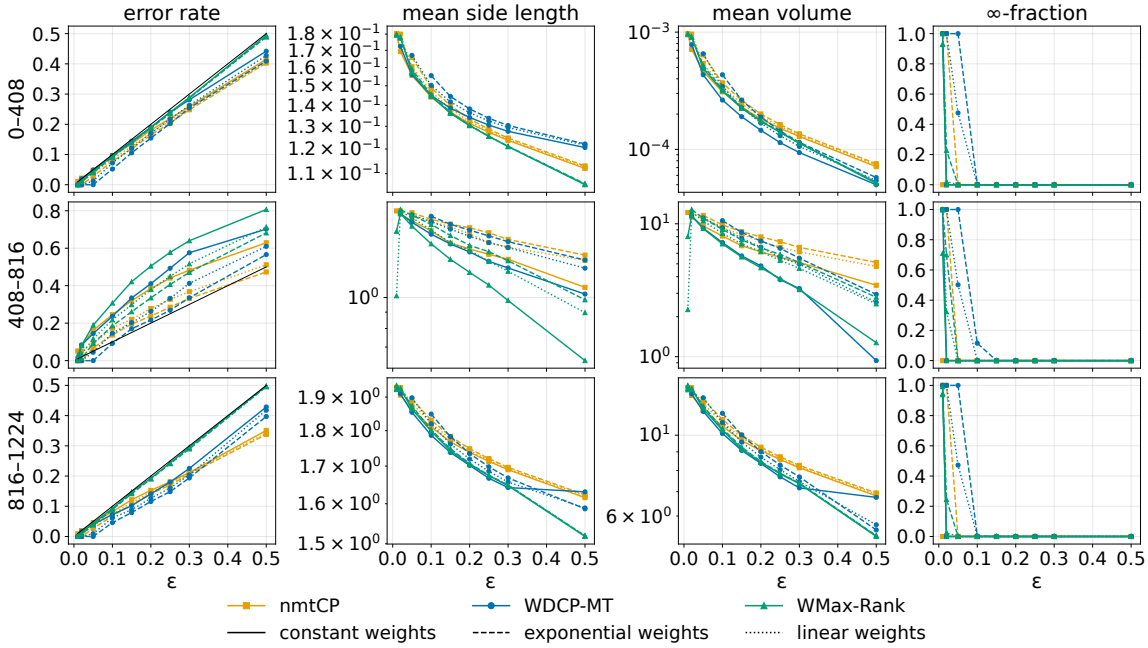


Figure S9: Mean error rate, mean hyperrectangle side length and mean hyperrectangle volume over the finite prediction regions, and fraction of prediction regions with at least one infinite bound (∞-fraction) as a function of the significance level ϵ for `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the synthetic changepoint data set with $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$, shown for three consecutive sections of the test set.

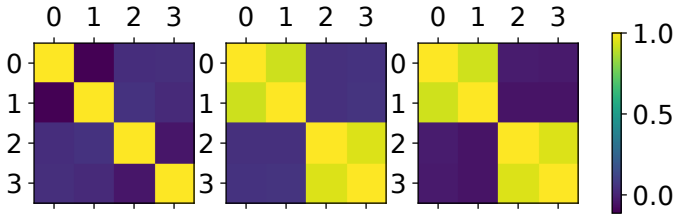


Figure S10: Correlation of the nonconformity scores across dimensions for the synthetic changepoint data set with $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$ and the three sections of the test set.

F.3. Synthetic Distribution Drift with Correlated Noise

Figure S11 displays the results from applying `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the Synthetic Distribution Drift with Correlated Noise data set. The correlation of the nonconformity scores across dimensions is shown in Figure S12.

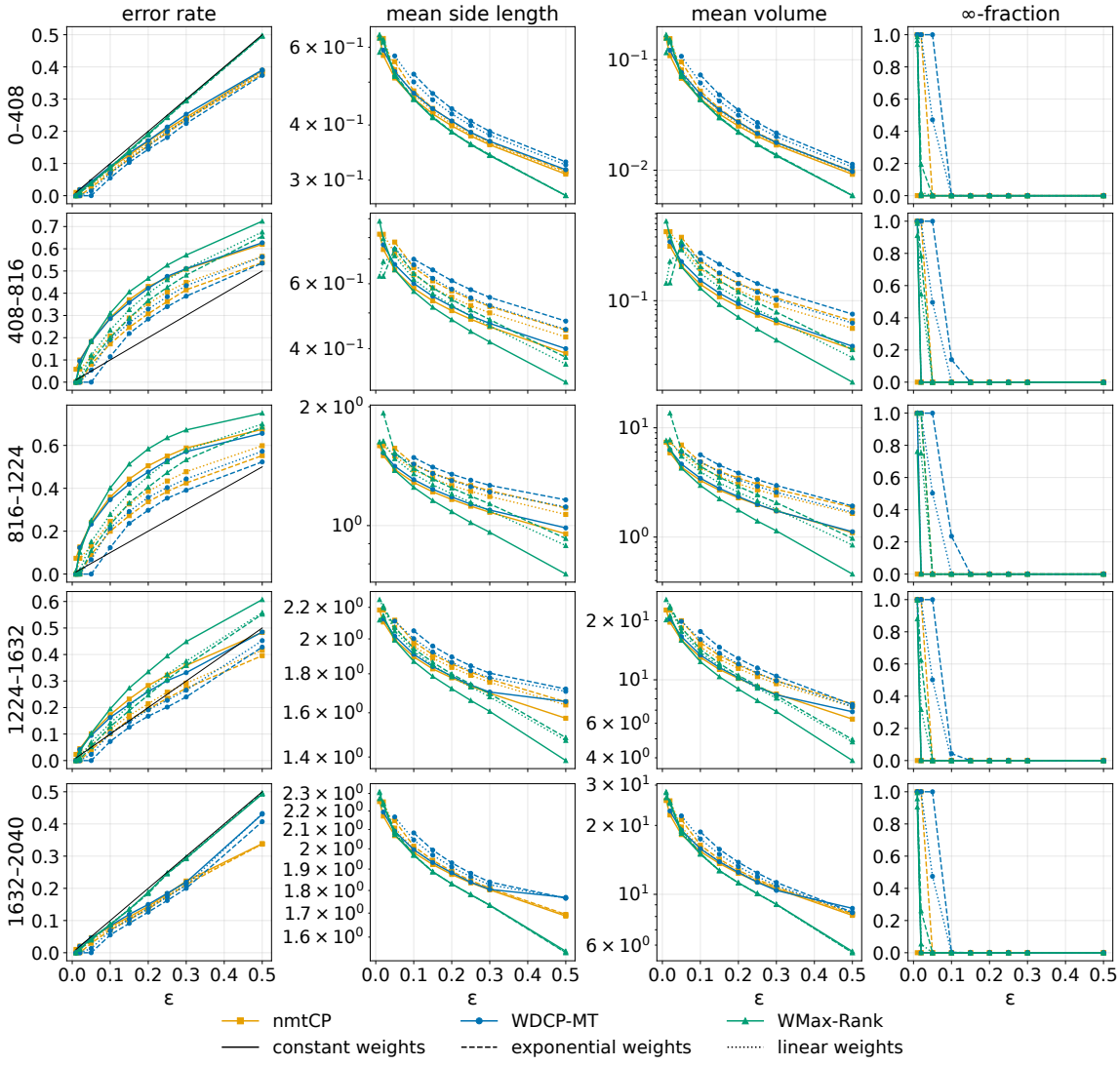


Figure S11: Mean error rate, mean hyperrectangle side length and mean hyperrectangle volume over the finite prediction regions, and fraction of prediction regions with at least one infinite bound (∞-fraction) as a function of the significance level ϵ for `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the synthetic distribution drift data set with $e \sim \mathcal{N}([0,0]^\top, \boldsymbol{\sigma})$, shown for five consecutive sections of the test set.

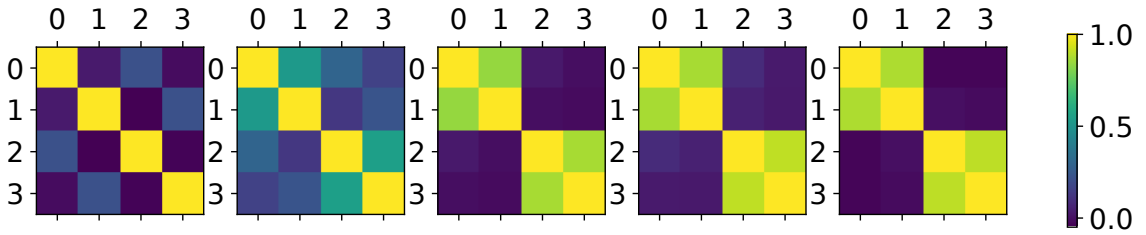


Figure S12: Correlation of the nonconformity scores across dimensions for the synthetic distribution drift data set with $e \sim \mathcal{N}([0,0]^\top, \boldsymbol{\sigma})$ and the five sections of the test set.

F.4. Synthetic Distribution Drift with Uncorrelated Noise

Figure S13 displays the results from applying `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the Synthetic Distribution Drift with Uncorrelated Noise data set. The correlation of the nonconformity scores across dimensions is shown in Figure S14.

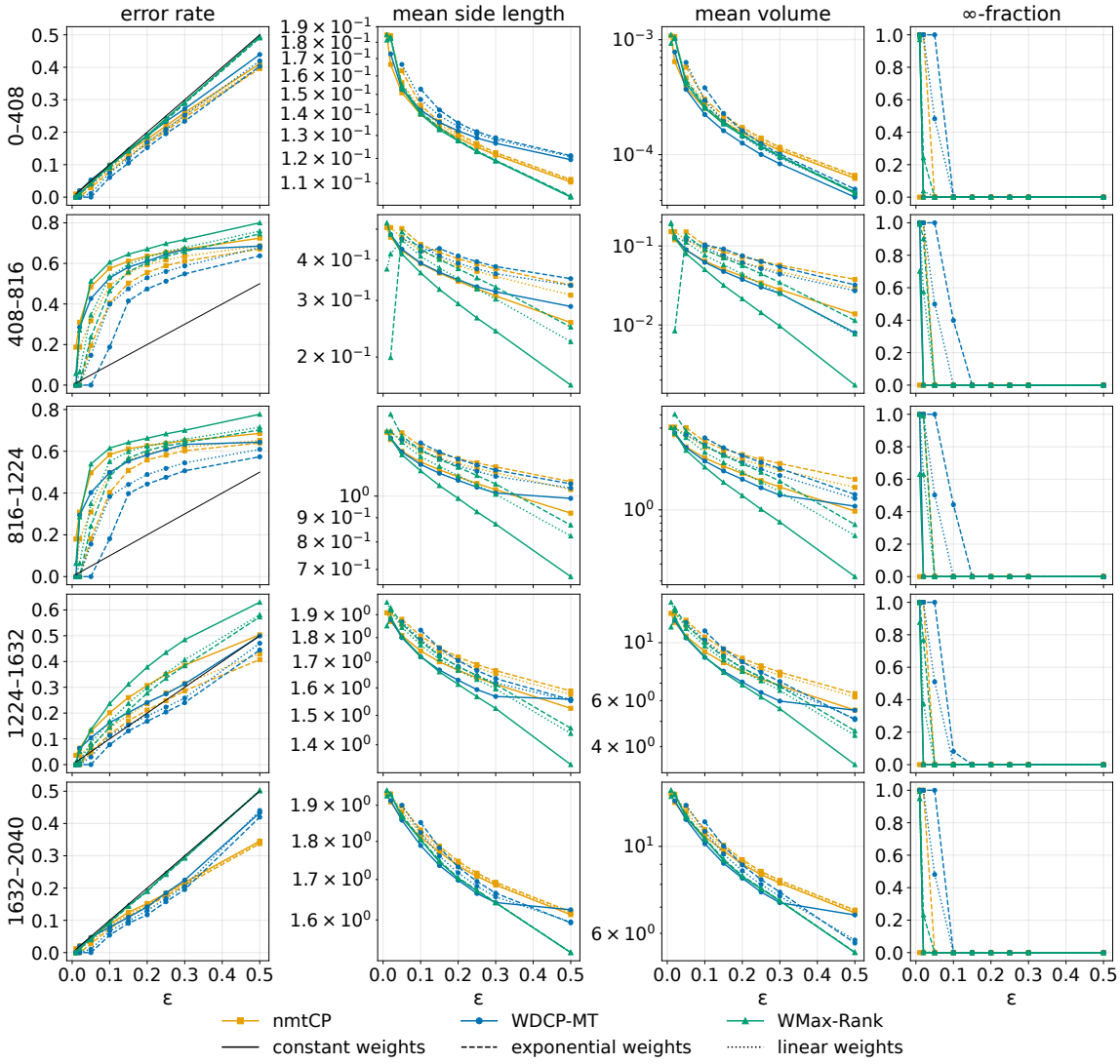


Figure S13: Mean error rate, mean hyperrectangle side length and mean hyperrectangle volume over the finite prediction regions, and fraction of prediction regions with at least one infinite bound (∞-fraction) as a function of the significance level ϵ for `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the synthetic distribution drift data set with $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$, shown for five consecutive sections of the test set.

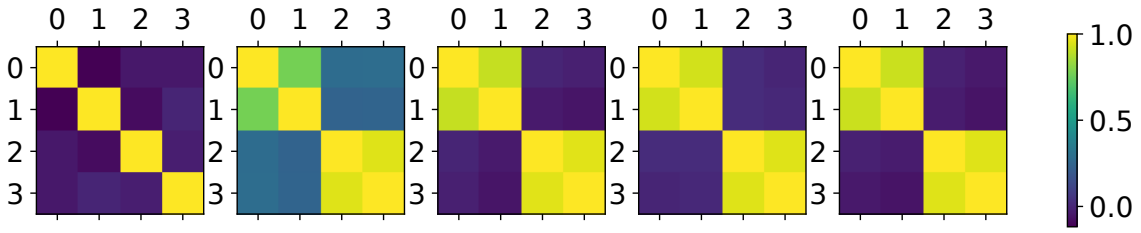


Figure S14: Correlation of the nonconformity scores across dimensions for the synthetic distribution drift data set with $e \sim [\mathcal{U}(-1, 1), \chi^2_2]^\top$ and the five sections of the test set.

F.5. ELEC2

Figure S15 displays the results from applying `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the ELEC2 data set (Harries, 1999). The correlation of the nonconformity scores across dimensions is shown in Figure S16.

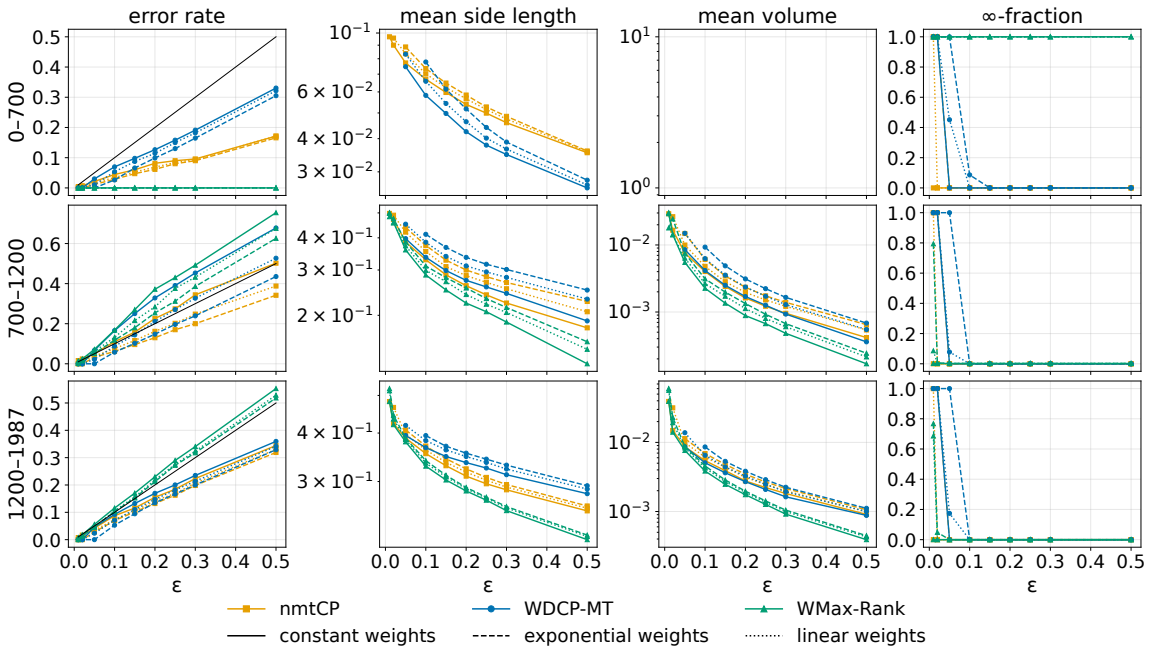


Figure S15: Mean error rate, mean hyperrectangle side length and mean hyperrectangle volume over the finite prediction regions, and fraction of prediction regions with at least one infinite bound (∞-fraction) as a function of the significance level ϵ for `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the ELEC2 (Harries, 1999) data set, shown for three consecutive sections of the test set.

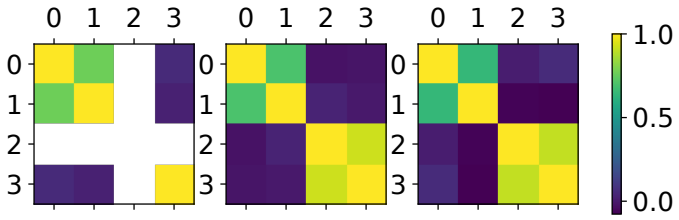


Figure S16: Nonconformity score correlation for the ELEC2 (Harries, 1999) data set and the three sections of the test set.

Degenerate dimensions in the first section. ELEC2 contains no data for Victoria in its first 17 424 time steps and stores a constant value for the Victorian demand instead. The proper training set, the calibration set, and the first section of the test set therefore contain only this value, the underlying model predicts it exactly, and the nonconformity scores of the Victorian demand at both horizons are zero in every calibration set of this section. These two dimensions are practically never missed, and the three methods treat them differently. `nmtCP` assigns ϵ/m to every dimension, including the degenerate ones, and stays far below the significance level. At $\epsilon = 0.5$ with constant weights, its error rate is 0.17. `WDCP-MT` assigns almost the entire budget to the two informative dimensions (mean dimension-wise significance levels 0.29 and 0.19, against 0.008 for each degenerate dimension) and reaches an error rate of 0.33, averaged over the ten replications. `WMax-Rank` returns $\mathcal{Y}$ at every significance level and for every weight function. All calibration scores of a degenerate dimension are tied, so every calibration instance has rank n in that dimension, all maximum ranks equal n, and the weighted quantile of the maximum ranks is n or $+\infty$, for which Algorithm S1 returns $\mathcal{Y}$. This is the conservative treatment of ties that the coverage-gap analysis of `WMax-Rank` assumes (Section B). Since the volume is zero for `nmtCP` and `WDCP-MT` and undefined for `WMax-Rank`, the volume panel of the first section is empty. A degenerate dimension thus disables the joint calibration of `WMax-Rank`, whereas the dimension-wise allocation of `WDCP-MT` turns it to its advantage. Randomised tie-breaking, which gives tied scores independent random ranks, would keep the regions of `WMax-Rank` finite in this case and is left for future work.

F.6. Tétouan City

Figure S17 displays the results from applying `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the Tétouan City (Salam and Hibaoui, 2018) data set. The correlation of the nonconformity scores across dimensions for the test set is shown in Figure S18.

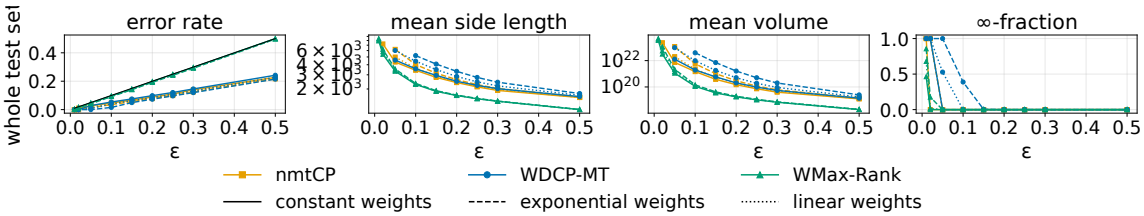


Figure S17: Mean error rate, mean hyperrectangle side length and mean hyperrectangle volume over the finite prediction regions, and fraction of prediction regions with at least one infinite bound (∞-fraction) as a function of the significance level ϵ for `nmtCP`, `WDCP-MT`, and `WMax-Rank` on the Tétouan City (Salam and Hibaoui, 2018) data set, shown for the whole test set.

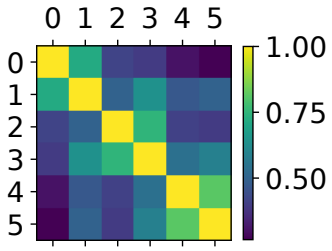


Figure S18: Correlation of the nonconformity scores across dimensions for the Tétouan City (Salam and Hibaoui, 2018) data set.

References

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc., 2021.

Michael Bonnell Harries. SPLICE-2 comparative evaluation: Electricity pricing, 1999.

Hans Kellerer, Ulrich Pferschy, and David Pisinger. The multiple-choice knapsack problem. In *Knapsack Problems*, pages 317–347. Springer, Berlin, Heidelberg, 2004. doi: 10.1007/978-3-540-24777-7_11.

Abdul Rahim Salam and Abdelaaziz El Hibaoui. Comparison of machine learning algorithms for the power consumption prediction: Case study of Tetouan city. In *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*, pages 1–5, 2018.

Filip Schlembach, Evgueni Smirnov, Irena Koprinska, and Mark H. M. Winands. Conformal multistep-ahead multivariate time-series forecasting. *Machine Learning*, 114(7):165, 2025a.

Filip Schlembach, Evgueni Smirnov, and Mark H. M. Winands. Dynamic conformal prediction for multi-target regression: Optimising informational efficiency under joint validity. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pages 193–213, 2025b.

Filip Schlembach, Evgueni Smirnov, and Mark H. M. Winands. Corrected Max-Rank for Finite-Sample Conformal Prediction. In *Proceedings of the Fifteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 329 of *Proceedings of Machine Learning Research*, pages 199–221. PMLR, September 2026. URL https://proceedings.mlr.press/v329/schlembach26a.html.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, Christian A. Naesseth, and Eric Nalisnick. Max-Rank: Efficient multiple testing for conformal prediction. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3898–3906, 2025.