Dynamic Conformal Prediction for Multi-Target Regression: Optimising Informational Efficiency under Joint Validity

Filip Schlembach FILIP.SCHLEMBACH@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

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

Inductive conformal prediction equips point regressors with finite-sample prediction sets that provably contain the unknown label with prescribed probability. For multi-target regression, joint coverage across all output dimensions can be guaranteed by combining one-dimensional conformal predictors, one for each output dimension, resulting in an axis-aligned hyperrectangular prediction region. The validity and informational efficiency of these hyperrectangular prediction regions depend on the choice of the targeted error rate for the individual one-dimensional conformal predictors. We cast this choice as an error-budget allocation problem and introduce Dynamic Conformal Prediction for Multi-Target Regression (DCP-MT), a method that finds the budget allocation which minimises the hyperrectangles' volumes while retaining joint coverage under exchangeability. Experiments on synthetic and real-world data sets demonstrate that DCP-MT reduces hyperrectangle volumes compared to state-of-the-art methods when nonconformity scores' correlations across target dimensions are weak or heterogeneous, while maintaining the nominal coverage. The proposed method thus offers a simple, drop-in solution for existing multi-target regression pipelines.

Keywords: Inductive conformal prediction, multi-target regression, hyperrectangular prediction regions, multiple hypothesis testing, uncertainty quantification.

1. Introduction

The reliability of machine-learning methods increasingly hinges on their ability not only to predict accurately but also to quantify the uncertainty associated with every prediction. *Conformal prediction* (CP) has emerged as a distribution-free, model agnostic framework that provides prediction regions for point forecasts with finite-sample validity guarantees (Vovk et al., 2005). While applying CP to single-target regression is straightforward, many real-world tasks require multi-target regression (Neeven and Smirnov, 2018). In this setting, practitioners need a joint prediction region that covers all label dimensions simultaneously with a prescribed error rate. Maintaining this joint validity while minimising the prediction region's volume is a challenging problem.

Various methods have been proposed to apply CP to multi-target regression problems (Vovk, 2013; Stankevičiūtė et al., 2021; Messoudi et al., 2021, 2022; Schlembach et al., 2022, 2025; Feldman et al., 2023; Ajroldi et al., 2023; Diquigiovanni et al., 2024), producing prediction regions of different shapes. In this article, we focus on CP variants that generate axis-aligned hyperrectangular prediction regions because (i) they retain the intuitive per-target interpretation of classic one-dimensional intervals making them explainable and easily

interpretable in human-in-the-loop workflows, (ii) they provide individual guarantees on each dimension, and (iii) they integrate seamlessly into tabular decision rules or rule-based post-processing pipelines.

Current methods that produce hyperrectangular prediction regions do this by combining multiple conformal predictors, one for each dimension of the label space, and assign all of them the same targeted error rate. Two approaches exist to determine the common targeted error rate. The first splits the global targeted error rate evenly, applying Bonferroni (Vovk, 2013) or Šidák corrections (Messoudi et al., 2021), which are often overly conservative. The second exploits the correlation between the output's dimensions (Messoudi et al., 2021; Timans et al., 2025), which is efficient when there is a strong positive correlation across targets. To improve the prediction region's informational efficiency[1] when there is no or weak correlation across the output's dimensions, we investigate how to set the targeted error rate for the individual single-target conformal predictors to minimise the hyperrectangular prediction region's volume.

Our contributions are fourfold:

- We formulate the choice of the local targeted error rates for the individual single-target conformal predictors as a global error-budget allocation problem and prove that joint validity is guaranteed whenever the sum of the local targeted error rates does not exceed the global targeted error rate.
- Building on this insight, we introduce *Dynamic Conformal Prediction for Multi-Target Regression* (DCP-MT), an approach that finds the budget allocation, which minimises the hyperrectangle's volume without distributional assumptions using integer linear programming.
- We provide a theoretical analysis comparing DCP-MT to related state-of-the-art methods producing hyperrectangular prediction regions.
- Extensive experiments on synthetic and public real-world data sets demonstrate that DCP-MT improves hyperrectangle volume compared to state-of-the-art methods when nonconformity scores' correlations across target dimensions are weak and have different marginal distributions, while maintaining the nominal coverage.

The remainder of this article is structured as follows. Section 2 discusses inductive CP for multi-target regression and highlights the difficulties of extending single-target CP to multi-target regression. Section 3 surveys existing work on applying CP in the multi-target regression setting, focussing on methods that produce axis-aligned hyperrectangular prediction regions. Our method DCP-MT is proposed in Section 4, where we prove its validity and discuss implementation details. Section 6 reports experimental results, and Section 7 discusses the findings and compares DCP-MT to related methods. Section 8 provides the conclusions and future research directions. Additional experimental details are given in the Appendix.

1. We use the term *informational efficiency* as a synonym for the *efficiency* of prediction sets, referring to their size, as introduced by Vovk et al. (2005, p.9). This is done to clearly distinguish informational efficiency from computational efficiency.

2. Inductive Conformal Prediction for Multi-Target Regression

In this section, we introduce the general concepts and associated notation used in this article in two parts. First, Subsection 2.1 recapitulates multi-target regression. Second, Subsection 2.2 describes inductive conformal prediction (ICP) in generic terms, how it is applied to single-target regression problems and then shows how and why extending ICP to the multi-target setting is non-trivial.

2.1. Multi-Target Regression

Let $\mathbf{X} \subseteq \mathbb{R}^L$ be the object space and $\mathbf{Y} \subseteq \mathbb{R}^M$ the M-dimensional label space. We are given a sample of N pairs $\{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N} \sim \mathcal{P}_{\mathbf{X},\mathbf{Y}}$, drawn from an arbitrary probability distribution $\mathcal{P}_{\mathbf{X},\mathbf{Y}}$, where each label $\mathbf{y}^{(n)} = (y_1^{(n)}, \ldots, y_M^{(n)})^\top$ contains M real-valued targets observed simultaneously with the object $\mathbf{x}^{(n)} \in \mathbf{X}$. A *multi-target regressor* is any learning algorithm that, after training on the sample, returns a fitted model $\hat{\mathbf{f}} : \mathbf{X} \longrightarrow \mathbf{Y},\ \mathbf{x} \mapsto \hat{\mathbf{f}}(\mathbf{x}) = \hat{\mathbf{y}}$. The fitted model $\hat{\mathbf{f}}$ is a point estimator which produces an estimate of the label $\hat{\mathbf{y}}$ for a given object $\mathbf{x}$. We drop the superscript (n) whenever the context is clear and denote generic observations simply by $(\mathbf{x}, \mathbf{y})$.

2.2. Inductive Conformal Prediction

In the regression setting, inductive conformal prediction (ICP) turns any point estimator into a *set-valued* predictor that satisfies a finite-sample coverage guarantee under two assumptions. First, the examples are exchangeable, and second, the learning algorithm treats them symmetrically, i.e. the model fitting algorithm produces the same model[2] independently of the order of the examples in the training set (Barber et al., 2022). We summarise the generic construction in a way that is agnostic to the output dimension M. The mechanism is the same whether $\mathbf{Y} \subseteq \mathbb{R}$ or $\mathbf{Y} \subseteq \mathbb{R}^M$.

Given a sample $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ of N examples, randomly split it into a *proper training set* $\mathcal{D}_{\text{tr}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{tr}}}$ of size N_{tr} and a *calibration set* $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=N_{\text{tr}}+1}^{N}$ of size $N_{\text{cal}} = N - N_{\text{tr}}$. Next, fit a model on $\mathcal{D}_{\text{tr}}$ and obtain a point predictor $\hat{\mathbf{f}} : \mathbf{X} \to \mathbf{Y},\ \mathbf{x} \mapsto \hat{\mathbf{f}}(\mathbf{x}) = \hat{\mathbf{y}}$.

A *nonconformity measure* (NCM) is a measurable function

$$A : \mathbf{X} \times \mathbf{Y} \longrightarrow \mathbb{R}, \qquad (\mathbf{x}, \mathbf{y}) \mapsto A\big((\mathbf{x}, \mathbf{y}), \hat{\mathbf{f}}\big), \tag{1}$$

that assigns a *single scalar* to the example $(\mathbf{x}, \mathbf{y})$. The NCM estimates how different any joint observation $(\mathbf{x}, \mathbf{y})$ is from the examples in $\mathcal{D}_{\text{tr}}$ (Vovk et al., 2005).

Use an NCM to compute the nonconformity scores $\alpha^{(n)} = A\big((\mathbf{x}^{(n)}, \mathbf{y}^{(n)}), \hat{\mathbf{f}}\big)$ for each joint observation in the calibration set $\mathcal{D}_{\text{cal}}$. Now let $\alpha_{(1-\epsilon)} = \mathsf{Q}_{1-\epsilon}(\{\alpha^{(n)}\}_{n=N_{\text{tr}}+1}^{N})$ be the empirical $(1-\epsilon)$ quantile of $\{\alpha^{(n)}\}_{n=N_{\text{tr}}+1}^{N}$. $\mathsf{Q}_{1-\epsilon}$ denotes the quantile function which returns the $\lceil (1-\epsilon)(N_{\text{cal}}+1) \rceil$th smallest value of the computed nonconformity scores.

Finally, for a new object $\mathbf{x}^{(N+1)} \in \mathbf{X}$, define the prediction region

$$\Gamma^{\epsilon}(\mathbf{x}^{(N+1)}) = \Big\{ \mathbf{y} \in \mathbf{Y} : A\big((\mathbf{x}^{(N+1)}, \mathbf{y}), \hat{\mathbf{f}}\big) \leq \alpha_{(1-\epsilon)} \Big\}. \tag{2}$$

2. The model fitting algorithm needs to produce models with the same distribution for randomized algorithms.

Any Γ^ϵ constructed in such a way is called an *inductive conformal predictor*. Inductive conformal predictors are conservatively valid meaning that

$$\mathbb{P}\big(\mathbf{y}^{(N+1)} \in \Gamma^\epsilon(\mathbf{x}^{(N+1)})\big) \geq 1 - \epsilon, \tag{3}$$

regardless of the distribution of $(\mathbf{X}, \mathbf{Y})$ and of the form of A (Vovk et al., 2005). We call ϵ the targeted error rate and say that an error occurs when $\mathbf{y}^{(N+1)} \notin \Gamma^\epsilon(\mathbf{x}^{(N+1)})$.

For a more in-depth overview of conformal prediction, please refer to (Vovk et al., 2005; Toccaceli, 2022; Fontana et al., 2023). Angelopoulos and Bates (2022) provide an excellent hands-on introduction to conformal prediction.

For single-target regression problems, a classic NCM is the absolute residual $\alpha^{(n)} = |\hat{y}^{(n)} - y^{(n)}|$. Using absolute residuals as the NCM produces the prediction region

$$\Gamma^\epsilon(\mathbf{x}) = \big[\hat{y} - \alpha_{(1-\epsilon)},\ \hat{y} + \alpha_{(1-\epsilon)}\big], \tag{4}$$

which takes the form of an interval centred around $\hat{y}$. When using absolute residuals as the NCM, the inductive conformal predictor Γ^ϵ assigns all test points intervals of the same size equal to $2\alpha_{(1-\epsilon)}$ (Papadopoulos et al., 2002). There exist more complex nonconformity measures for single-target regression, an example of which are normalized nonconformity measures (Papadopoulos et al., 2002; Papadopoulos and Haralambous, 2011).

We classify approaches applying conformal prediction to multi-target regression into two main groups:[3] *single-test* methods and *multi-tests* methods. This classification stems from observing conformal prediction through a hypothesis testing lens (Vovk et al., 2005; Timans et al., 2025). Given a candidate label $\bar{\mathbf{y}}$ for an object $\mathbf{x}$, a conformal predictor as defined in (2) tests whether it should be included in the prediction region.

Single-test methods employ an NCM $A : \mathbf{X} \times \mathbf{Y} \to \mathbb{R}$, with $\mathbf{Y} \subseteq \mathbb{R}^M$, which assigns a single nonconformity score to any pair $(\mathbf{x}, \mathbf{y})$. The single test for inclusion can then be performed by comparing the nonconformity score to the threshold $\alpha_{1-\epsilon}$ provided by the quantile function. A simple examples of such an NCM is the Euclidean norm $\alpha = \|\hat{\mathbf{y}} - \mathbf{y}\|$ which leads to a hyperspherical prediction region. Johnstone and Cox (2021); Messoudi et al. (2022); Dheur et al. (2025) present examples of NCMs producing prediction regions with more complex shapes. These methods can exploit rich dependence structures but the prediction regions' complex shapes make them difficult to interpret and therefore less attractive in human-in-the-loop scenarios. We mention them here for completeness but focus on multi-tests methods for the remainder of this article.

Multi-tests methods, on which we focus for the remainder of this article, combine M single-target conformal predictors $\{\Gamma^{\epsilon_m}_m\}_{m=1}^M$, one for each dimension m of the label space, into a single Γ^ϵ. These conformal predictors employ an NCM $A : \mathbf{X} \times \mathbb{R} \to \mathbb{R}$ which provides nonconformity scores for the associated dimension. The test for inclusion of a candidate label $\bar{\mathbf{y}}$ is therefore performed M times in parallel. Timans et al. (2025) provide a rigorous theoretical analysis relating the combination of multiple conformal predictors for multi-target regression to multiple hypothesis testing.

3. Ajroldi et al. (2023); Diquigiovanni et al. (2024) present a related third approach, building upon functional data analysis which is not directly applicable to point forecasts.

The prediction region of multi-tests methods is the Cartesian product of the M one-dimensional intervals constructed by the $\Gamma_m^{\epsilon_m}$ and therefore of hyperrectangular shape.[4] Prediction regions constructed in this way have the benefit that the sides of the hyperrectangle are parallel to the axes of the label space. They also provide individual coverage guarantees for each dimension of the label space. For any label $\mathbf{y}$ to be considered inside the prediction region, all of its components $\{y_m\}_{m=1}^M$ need to be contained within the prediction interval associated with their dimension m. Hence, we can write validity, as defined in (3), for multi-tests methods as

$$\begin{aligned}\mathbb{P}\big(\mathbf{y}^{(N+1)} \in \Gamma^{\epsilon_h}(\mathbf{x}^{(N+1)})\big) &= \mathbb{P}\left(\bigcap_{m=1}^{M}\left(y_m^{(n+1)} \in \Gamma_m^{\epsilon_m}(\mathbf{x}^{(N+1)})\right)\right) \\ &\geq 1-\epsilon_h\ , \end{aligned} \tag{5}$$

where $\Gamma^{\epsilon}(\mathbf{x}^{(N+1)})$ is the hyperrectangular prediction region with a global targeted error rate ϵ_h and $\Gamma_m^{\epsilon_m}(\mathbf{x}^{(N+1)})$ is the prediction interval for dimension m with a local targeted error rate ϵ_m (Timans et al., 2025). To avoid confusion $\epsilon_h = \epsilon$ denotes the global error rate moving forward.

The challenge in the design of multi-tests methods lies in the choice of the local targeted error rates ϵ_m. They need to be chosen in such a way, that the method is valid as described in (5). Next to validity, *informational efficiency* is a desired property for conformal predictors, meaning that the generated prediction region should be non-empty and as small as possible (Vovk et al., 2005). Exhaustively searching for a valid prediction region of minimal size is computationally inefficient because the quantile function each $\Gamma_m^{\epsilon_m}$ uses can return $N+1$ different values, leading to $(N+1)^M$ combinations that need to be tested (Vovk, 2013). The following section describes a number of methods that have been devised to reduce this computational cost.

3. Related Work

Vovk (2013) propose the first computationally efficient multi-tests method called *Bonferroni predictors*. A Bonferroni predictor assigns the same value $\epsilon_m = \epsilon_h/M$ to all local significance levels. Its name and the choice for ϵ_m stem from the Bonferroni family-wise error rate (FWER) correction method. The validity of Bonferroni predictors is directly derived from Boole's inequality. While Vovk (2013) introduces the method initially for transductive conformal predictors, which do not rely on a separate calibration set, it has subsequently been applied to ICP (Stankevičiūtė et al., 2021; Schlembach et al., 2022, 2025; Vovk et al., 2022). These applications show that Bonferroni predictors are generally too conservative, producing prediction regions that are larger than necessary for the targeted global error rate ϵ_h.

To improve the informational efficiency of the prediction regions, Messoudi et al. (2021) try to capture the dependencies across the nonconformity scores' dimensions by fitting copulas (Sklar, 1959) to them. Messoudi et al. (2021) test three copulas. The independent

4. It is possible to envision NCMs for single-target conformal predictors that do not lead to prediction regions that take the form of a single continuous interval. In our experience, they are uncommon in practice and their discussion is omitted in this article.

copula is equivalent to the Šidák FWER correction and sets $\epsilon_m = 1-(1-\epsilon_h)^{1/M}$, assuming independence between the nonconformity scores. It is less conservative than the Bonferroni predictors but makes independence assumptions about the nonconformity scores' distributions that are not always met. The use of the Gumbel copula results in setting $\epsilon_m = 1-(1-\epsilon_h)^{1/\sqrt[\theta]{M}}$, where θ is a parameter in the copula's generator function that Messoudi et al. (2021) estimate using the matrix of nonconformity scores and the Maximum Pseudo-Likelihood Estimator. For $\theta = 1$ the Gumbel copula is equivalent to the independent copula. For larger values of θ, it results in an even less conservative FWER correction that is justified if there is positive correlation between the nonconformity scores of the dimensions of the label space. The empirical copula estimates the cumulative distribution function (CDF) $[0,1]^M \to [0,1]$ which, given the M local targeted error rates ϵ_m, can be used to compute the global targeted error rate ϵ_h. Because there is no analytical expression for the inverse of this CDF, searching for an optimal solution leads to the same computational cost as described at the end of Subsection 2.2. For this reason Messoudi et al. (2021) only search the subspace where all ϵ_m have the same value, allowing for a simple dichotomic search.

Timans et al. (2025) propose `Max-Rank`, a method that is informationally efficient when the nonconformity scores in the dimensions of the label space are positively correlated. `Max-Rank` first computes the rank for all nonconformity scores $\{\text{rank}(\alpha_m^{(n)})\}_{n=N_{\text{tr}}+1}^{N}$ for each dimension m in the label space. It then computes the $\max(\{\text{rank}(\alpha_m^{(n)})_{m=1}^{M}\})$ over all dimensions for each element in the calibration set. Finally, applying the quantile function $\mathsf{Q}_{1-\epsilon_h}$ to the list of max ranks returns the rank $r_{1-\epsilon_h}$ which is used to compute the local targeted error rates $\epsilon_m = r_{1-\epsilon_h}/(N_{\text{cal}}+1)$. In practice, the ϵ_m do not need to be computed, as $r_{1-\epsilon_h}$ can directly be used to determine the empirical $(1-\epsilon_m)$ quantile for each dimension.[5]

The copula conformal predictors and Max-Rank are computationally and informationally efficient, performing as well as Bonferroni predictors in the worst case scenarios. All methods presented in this section assign the same local targeted error ϵ_m to all dimensions. In situations where the nonconformity scores are weakly correlated across the nonconformity scores' dimensions and when their marginal distributions differ, identical ϵ_ms will not produce optimal solutions. To solve this, we propose a new method in the following section.

4. Dynamic Conformal Prediction for Multi-Target Regression (DCP-MT)

We now present *Dynamic Conformal Prediction for Multi-Target Regression* (DCP-MT), a method that can assign different local targeted error rates to different dimensions in order to minimise the prediction region's volume. A sufficient condition for the validity of any such method is presented in Theorem 1.

Theorem 1 *Let ϵ_h be the global targeted error rate for a multi-tests conformal predictor and let $\{\epsilon_m\}_{m=1}^{M}$ be the local error rates associated with each dimension of the label space.*

5. For `Max-Rank` the distinction between single-test methods and multi-tests methods becomes ambiguous, as the computed max rank for each element in the calibration set could be interpreted as the nonconformity scores, collapsing the method to a single test.

A multi-tests conformal predictor is valid, if

$$\epsilon_h \leq \sum_{m=1}^{M} \epsilon_m \ . \tag{6}$$

Proof The result is derived by applying Boole's inequality to (5)

$$\begin{aligned}
\mathbb{P}\big(\mathbf{y}^{(N+1)} \in \Gamma^{\epsilon_h}(\mathbf{x}^{(N+1)})\big) &= 1 - \mathbb{P}\left(\bigcup_{m=1}^{M}\left(y_m^{(n+1)} \notin \Gamma_m^{\epsilon_m}(\mathbf{x}^{(N+1)})\right)\right) \\
&\geq 1 - \sum_{m=1}^{M} \mathbb{P}\left(y_m^{(n+1)} \notin \Gamma_m^{\epsilon_m}(\mathbf{x}^{(N+1)})\right) \\
&= 1 - \sum_{m=1}^{M} \epsilon_m \\
&\geq 1 - \epsilon_h \ , \qquad (7)
\end{aligned}$$

which can be rewritten as (6). ■

Equation (6) shows that for a multi-tests method to be valid, it is sufficient that the sum of the local targeted error rates ϵ_m does not exceed the global targeted error rate ϵ_h. We can therefore interpret the global targeted error rate as a global error budget. DCP-MT solves the problem of optimally allocating the global error budget ϵ_h between the M dimensions of the label space to minimise the volume of the resulting hyperrectangular prediction region. The volume of the prediction region is given by $\prod_{m=1}^{M} 2\mathsf{Q}_{1-\epsilon_m}(\{\alpha^{(m,n)}\}_{n=N_{\text{tr}}+1}^{N})$ where $\{\alpha^{(m,n)}\}_{n=N_{\text{tr}}+1}^{N}$ are all nonconformity scores associated with dimension m. Minimising this product with respect to the ϵ_m is equivalent to minimising the sum

$$\sum_{m=1}^{M} \ln\left(\mathsf{Q}_{1-\epsilon_m}(\{\alpha^{(m,n)}\}_{n=N_{\text{tr}}+1}^{N})\right) . \tag{8}$$

To find the minimum, DCP-MT first sorts the nonconformity scores for each dimension, such that $\alpha_m'^{(1)} \leq \alpha_m'^{(2)} \leq \cdots \leq \alpha_m'^{(N_{\text{cal}})}$, $\forall n \in \{1, \ldots, N\}$, where $\alpha_m'^{(n)}$ is a sorted nonconformity score with the associated local targeted error rate $\epsilon_m^{(n)} = 1 - n/(N_{\text{cal}}+1)$. Let $I_m^{(n)} \in \{0,\ 1\}$ be a binary variable indicating the choice of $\epsilon_m^{(n)}$ as the local targeted error rate for dimension m. Finding the optimal allocation of the error budget is then equivalent to minimising the cost function

$$\sum_{m=1}^{M}\sum_{n=1}^{N_{\text{cal}}} I_m^{(n)} \ln(\alpha_m^{(n)}) \tag{9}$$

under the constraints

$$\epsilon_h \leq \sum_{m=1}^{M} \epsilon_m \ , \tag{10}$$

and

$$\sum_{n=1}^{N_{\text{cal}}} I_m^{(n)} = 1 \ , \ \forall m \in \{1, \ldots, M\} \ . \tag{11}$$

Minimising the cost function (9) directly minimises the volume of the hyperrectangular prediction region as shown in (8). The transformation of the volume from a product into a sum is necessary to solve the problem via a standard mixed-integer linear programming (MILP) solver. Constraint (10) ensures validity while the constraints in (11) ensure that only one targeted error rate is chosen per dimension.

An MILP formulation fits this optimization problem because the choice of local significance levels e_m is equivalent to selecting one quantile from the discrete, finite set of nonconformity scores for each dimension m. These choices are represented by binary indicator variables $I_m^{(n)}$, which introduce integer constraints into the optimization. At the same time, the objective, which consists in minimizing the volume of the hyperrectangular prediction region, can be expressed as a sum of logarithms of the selected quantiles, and hence, as a linear function over the indicator variables. The MILP formulation enables us to efficiently explore this combinatorial space and find a globally optimal allocation of the error budget that respects the validity constraint in Equation (10).

In practice, DCP-MT is used by following the ICP procedure laid out in Section 2.2. First, a training algorithm fits the underlying model to the proper training set. Next, the individual inductive conformal predictors are calibrated on the calibration set, each on the nonconformity scores of their associated dimension. Finally, during inference, the hyperrectangular prediction region is built for each new object by solving the MILP optimisation problem presented in this section.

5. Comparison

This section discusses the theoretical results, comparing DCP-MT to Bonferroni predictors, copula conformal predictors, and Max-Rank.

In multi-target conformal prediction, guaranteeing joint coverage across all outputs is essential, but doing so efficiently is challenging. Among multi-tests methods, Bonferroni predictors (Vovk, 2013) provide FWER control by allocating the total significance level equally among targets, but they are highly conservative, resulting in unnecessarily wide prediction intervals. More sophisticated methods like copula conformal predictors (Messoudi et al., 2021; Sun and Yu, 2024) and Max-Rank (Timans et al., 2025) attempt to exploit dependencies between the dimensions' nonconformity scores. Copula-based methods explicitly model joint distributions, while Max-Rank leverages positive dependence using ranks. These methods are most efficient when outputs are strongly positively dependent and assign the same local significance levels ϵ_m to all targets. In the presence of positive dependencies between the dimensions' nonconformity scores copula conformal predictors and Max-Rank can violate the constraint (10) imposed by Boole's inequality while still maintaining theoretical validity guarantees.

DCP-MT offers a different approach. While it does not model output dependencies explicitly, it dynamically adjusts the local significance levels ϵ_m subject to constraint (10) imposed by Boole's inequality to decrease the volume of the predicted region. Therefore, DCP-MT achieves smaller prediction regions when the dimensions' nonconformity scores are weakly correlated and have different marginal distributions.

Once the local targeted error rates ϵ_m have been determined, all methods select the associated nonconformity score from the calibration set for each dimension using the quantile

function or the rank in the case of Max-Rank. Therefore, the difference in computational complexity between the methods stems from the way the local targeted error rates ϵ_m are determined. Bonferroni is the simplest method computationally, requiring only a simple division operation to compute the ϵ_m. Max-Rank is a little more complex. First Max-Rank requires a pass through the calibration set's nonconformity scores to compute the max rank. Then it applies the quantile function to the results and a single division operation to compute the ϵ_m. Finding the optimal solution for copula conformal predictors using the empirical copula would be computationally prohibitive due to the lack of an analytical expression for the CDF's inverse. Therefore, Messoudi et al. (2021) opted to assign all ϵ_m the same value, restricting the search space. Because the empirical copula is monotonically increasing when the ϵ_m are decreasing simultaneously, the largest value for the ϵ_m that satisfies the global targeted error rate can then be determined through a simple search operation. DCP-MT requires solving a single convex optimization problem to distribute the global error budget efficiently and set the ϵ_m. This is more efficient than finding a globally optimal hyperrectangle but remains computationally more complex than Bonferroni predictors, copula conformal predictors, and Max-Rank.

6. Experiments

In this section, we examine the empirical performance of the DCP-MT method presented in Section 4 on simulated and real-world data sets and compare it to other multi-tests methods. For the comparison we chose Bonferroni predictors as a baseline and copula conformal predictors using the empirical copula as well as Max-Rank as they are the best performing methods in (Messoudi et al., 2021) and (Timans et al., 2025). The code to reproduce the experiments is available at https://github.com/filipschlembach/dcp_mt.

Subsection 6.1 introduces the data sets. This is followed by the description of how the experiments are conducted and which underlying models we used in Subsection 6.2. We report the experimental results in Subsection 6.3.

6.1. Data Sets

This section presents the synthetic and real-world data sets used to evaluate DCP-MT and to compare it to other multi-tests methods.

Synthetic data sets. Inspired by Barber et al. (2022), the synthetic data sets used in this article have the structure

$$y = Ax + e \tag{12}$$

where A is a $M \times L$ shaped matrix of randomly generated coefficients and e is an added noise term. The data sets differ in the way the noise is generated.

For the first data set $L = 5$, $N = 2$, and the noise term $e \sim [\mathcal{U}(-1, 1), \chi^2_2]^\top$ is made up of two components that are sampled independently from a uniform and a Chi-squared distribution. This simulates a situation where the nonconformity scores are not correlated across the nonconformity scores' dimensions and when their marginal distributions differ. For the second, third, and fourth data set, $L = 5$, $N = 2$, and the noise term $e \sim \mathcal{N}([0, 0]^\top, \boldsymbol{\sigma})$ is sampled from a two-dimensional multivariate normal distribution, where $\boldsymbol{\sigma}$ is a 2×2 matrix with values equalling 1 on the main diagonal and σ otherwise. Setting σ to 0, 0.5, and 1 for

Table 1: Overview over the used real-world data sets. N is the number of examples in the data set, $L = \dim(\mathbf{X})$, and $M = \dim(\mathbf{Y})$.

Name	N **examples**	L	M
diabetes (Efron et al., 2004)	442	9	2
music origin (Zhou et al., 2014)	1059	68	2
rf1 (Tsoumakas et al., 2011)	9125	64	8
rf2 (Tsoumakas et al., 2011)	9125	576	8
scm1d (Tsoumakas et al., 2011)	9803	280	16
scm20d (Tsoumakas et al., 2011)	8966	61	16

the data sets respectively controls the correlation between the noise term's dimensions. We chose varying correlation to highlight the difference between DCP-MT, copula conformal predictors and Max-Rank, as copula conformal predictors and Max-Rank are expected to perform well for higher values of σ. For each data set, we generate a sample containing 1000 examples.

Real-world data sets. In addition to the experiments on synthetic data sets, we also apply DCP-MT and the methods presented in Section 3 to six real-world data sets shown in Table 1. For the diabetes data set we have chosen to predict features 3 and 4 as they are weakly correlated. Music origin (Zhou et al., 2014), rf1, rf2, scm1d, and scm20d (Tsoumakas et al., 2011) are taken from Messoudi et al. (2021) and Timans et al. (2025) tested Max-Rank on rf1, and scm1d.

6.2. Experimental Setup

The experiments use a 5-fold cross validation scheme. The proper training set $\mathcal{D}_{\text{tr}}$ and calibration set $\mathcal{D}_{\text{cal}}$ contain 67% and 33% of the examples in the training folds, respectively. As described in Subsection 2.2 we fit an underlying model to the proper training set and calibrate the conformal predictors on the calibration set. The test fold serves to evaluate the model and the conformal prediction methods. Matching the data generating process, we use scikit-learn's (Pedregosa et al., 2011) linear regressor as the underlying model for the synthetic data sets. For the diabetes (Efron et al., 2004) data set we employ XGBoost regressors as the underlying model which offer good performance at low computational cost. To maintain comparability, we preprocess the music origin (Zhou et al., 2014), rf1, rf2, scm1d, and scm20d (Tsoumakas et al., 2011) data sets following Messoudi et al. (2021). We also follow the lead of Messoudi et al. (2020) for the underlying model, using a neural network and replicating their architecture. All experiments are repeated 20 times.

6.3. Experimental Results

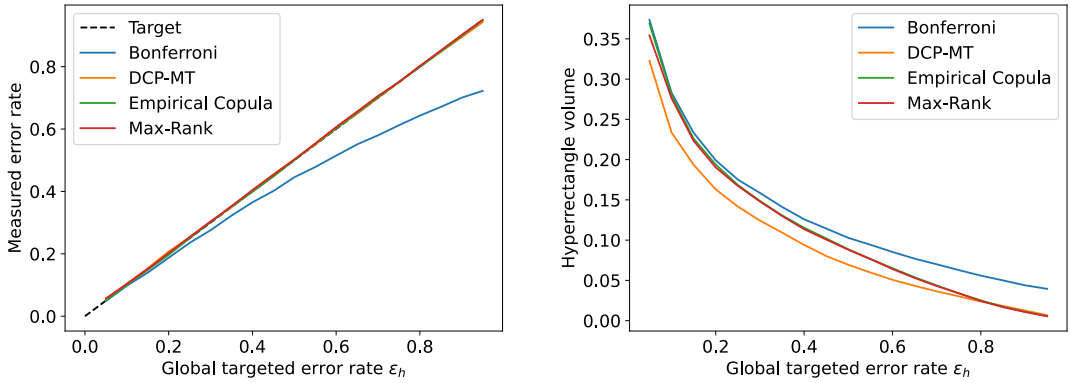
In this section, we first present the results for the synthetic data set with $e \sim [\mathcal{U}(-1, 1), \chi^2_2]^\top$ in detail, as it highlights the situation for which DCP-MT was designed. We then provide a summary of the results for all synthetic and real-world data sets. Appendix A contains the

Table 2: Multi-tests method's mean error rates and standard deviations for the synthetic data set with $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$. Underlined values indicate instances for which the measured error rate exceeds the targeted error rate while bold values indicate that the method produced the smallest prediction region.

ϵ_h	Bonferroni	DMT-CP	Empirical Copula	Max-Rank
0.05	0.049 ± 0.006	**0.052 ± 0.005**	0.051 ± 0.006	0.057 ± 0.007
0.10	0.098 ± 0.008	**0.104 ± 0.008**	0.101 ± 0.009	0.104 ± 0.007
0.15	0.140 ± 0.007	**0.154 ± 0.010**	0.148 ± 0.009	0.151 ± 0.010
0.20	0.188 ± 0.009	**0.207 ± 0.011**	0.197 ± 0.010	0.203 ± 0.011
0.25	0.235 ± 0.013	**0.253 ± 0.014**	0.250 ± 0.012	0.253 ± 0.012
0.30	0.276 ± 0.010	**0.304 ± 0.013**	0.301 ± 0.013	0.303 ± 0.013
0.35	0.322 ± 0.013	**0.349 ± 0.014**	0.351 ± 0.012	0.353 ± 0.012
0.40	0.365 ± 0.013	**0.404 ± 0.014**	0.399 ± 0.015	0.404 ± 0.016
0.45	0.401 ± 0.015	**0.455 ± 0.014**	0.448 ± 0.016	0.453 ± 0.016
0.50	0.445 ± 0.014	**0.502 ± 0.012**	0.500 ± 0.016	0.502 ± 0.015
0.55	0.478 ± 0.015	**0.550 ± 0.011**	0.552 ± 0.014	0.554 ± 0.014
0.60	0.515 ± 0.011	**0.602 ± 0.011**	0.602 ± 0.014	0.607 ± 0.015
0.65	0.551 ± 0.013	**0.649 ± 0.011**	0.652 ± 0.013	0.656 ± 0.013
0.70	0.580 ± 0.013	**0.700 ± 0.009**	0.702 ± 0.013	0.706 ± 0.012
0.75	0.612 ± 0.011	**0.750 ± 0.011**	0.751 ± 0.010	0.750 ± 0.010
0.80	0.643 ± 0.009	**0.803 ± 0.009**	0.799 ± 0.010	0.802 ± 0.010
0.85	0.671 ± 0.011	0.847 ± 0.009	0.849 ± 0.009	**0.852 ± 0.010**
0.90	0.701 ± 0.010	0.895 ± 0.009	0.898 ± 0.010	**0.902 ± 0.009**
0.95	0.722 ± 0.009	0.942 ± 0.009	0.949 ± 0.009	**0.950 ± 0.008**

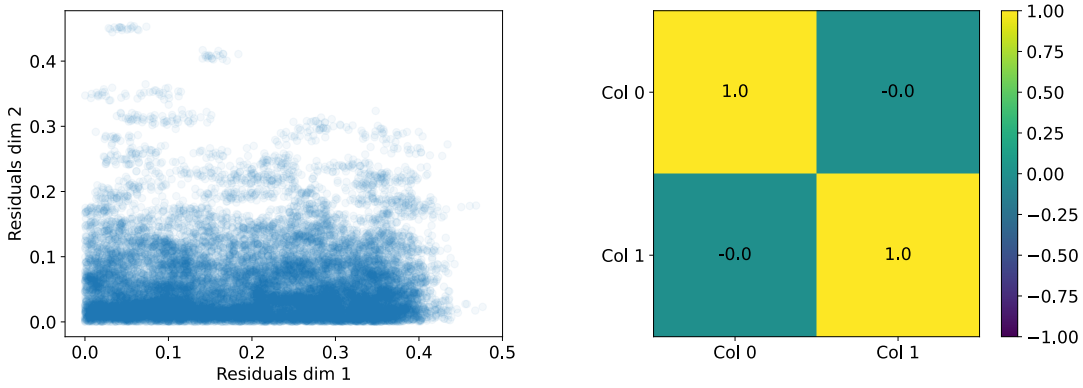
Table 3: Multi-tests method's mean hyperrectangle volumes for the synthetic data set with $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$. Bold values indicate that the method produced the smallest prediction region while underlined values highlight instances for which the measured error rate exceeds the targeted error rate.

ϵ_h	Bonferroni	DMT-CP	Empirical Copula	Max-Rank
0.05	0.374 ± 0.013	**0.323 ± 0.011**	0.369 ± 0.013	0.354 ± 0.011
0.10	0.283 ± 0.010	**0.234 ± 0.007**	0.280 ± 0.011	0.276 ± 0.011
0.15	0.233 ± 0.007	**0.194 ± 0.005**	0.226 ± 0.007	0.223 ± 0.007
0.20	0.199 ± 0.006	**0.163 ± 0.005**	0.194 ± 0.006	0.190 ± 0.006
0.25	0.175 ± 0.005	**0.142 ± 0.004**	0.169 ± 0.004	0.167 ± 0.004
0.30	0.159 ± 0.004	**0.124 ± 0.004**	0.149 ± 0.004	0.148 ± 0.004
0.35	0.141 ± 0.004	**0.109 ± 0.004**	0.131 ± 0.004	0.130 ± 0.004
0.40	0.126 ± 0.004	**0.094 ± 0.004**	0.115 ± 0.004	0.114 ± 0.004
0.45	0.114 ± 0.004	**0.080 ± 0.003**	0.102 ± 0.004	0.101 ± 0.004
0.50	0.103 ± 0.003	**0.069 ± 0.003**	0.089 ± 0.003	0.088 ± 0.003
0.55	0.094 ± 0.003	**0.060 ± 0.002**	0.077 ± 0.003	0.076 ± 0.003
0.60	0.085 ± 0.003	**0.051 ± 0.002**	0.065 ± 0.003	0.064 ± 0.003
0.65	0.077 ± 0.003	**0.043 ± 0.002**	0.054 ± 0.003	0.053 ± 0.003
0.70	0.070 ± 0.003	**0.036 ± 0.001**	0.044 ± 0.002	0.043 ± 0.002
0.75	0.063 ± 0.003	**0.030 ± 0.001**	0.034 ± 0.002	0.034 ± 0.002
0.80	0.056 ± 0.002	**0.024 ± 0.001**	0.025 ± 0.002	0.025 ± 0.002
0.85	0.050 ± 0.002	0.018 ± 0.001	0.017 ± 0.001	**0.017 ± 0.001**
0.90	0.044 ± 0.002	0.013 ± 0.001	0.011 ± 0.001	**0.011 ± 0.001**
0.95	0.040 ± 0.002	0.007 ± 0.001	0.006 ± 0.001	**0.005 ± 0.001**



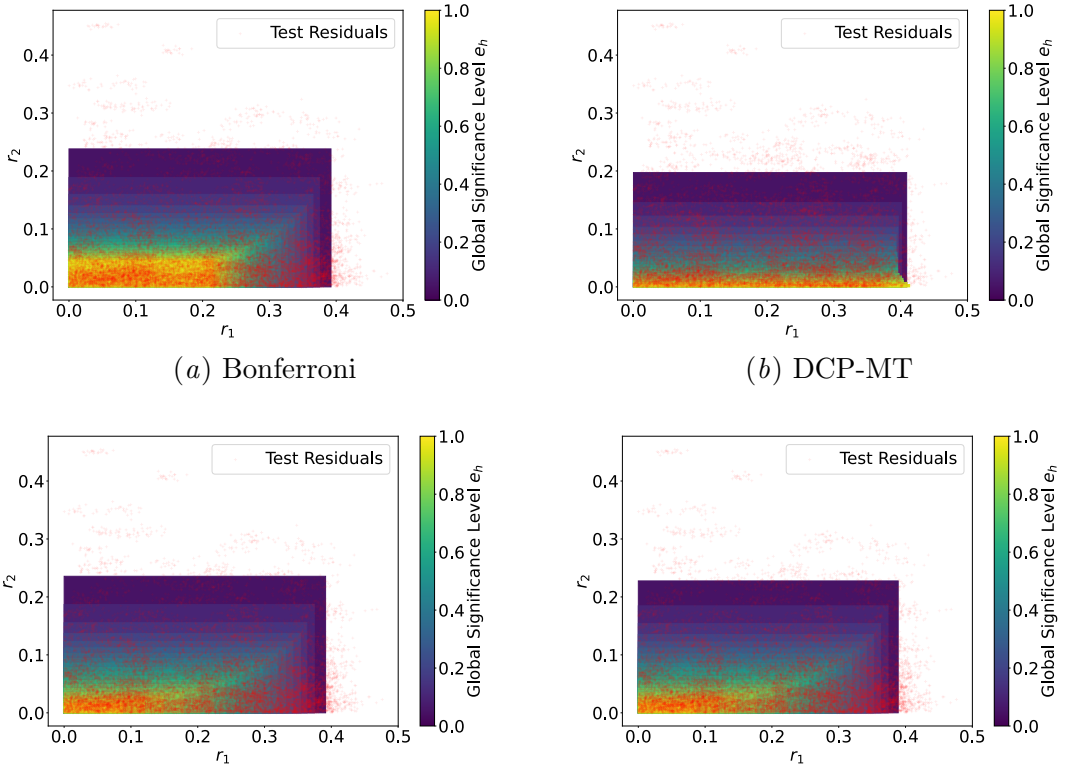
(*a*) Error rates (*b*) Volumes

Figure 1: Multi-tests method's mean error rates and volumes for the synthetic data set with $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$ over 20 repetitions.



(*a*) Residuals (*b*) Residual correlation matrix

Figure 2: Underlying model's absolute residuals and their correlation for the synthetic data set with $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$.



(a) Bonferroni (b) DCP-MT

(c) Empirical Copula (d) Max-Rank

Figure 3: Multi-tests method's partition of the residual space for the synthetic data set with $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$ averaged over 20 repetitions.

Table 4: Error rates acoss all experiments for $\epsilon_h = 0.05$. Bold values indicate that the method produced the smallest prediction region while underlined values highlight instances for which the measured error rate ≥ 0.05.

Data Set	Bonferroni	DMT-CP	Empirical Copula	Max-Rank
synt, $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$	0.049 ± 0.006	**0.052 ± 0.005**	0.051 ± 0.006	0.057 ± 0.007
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 0$	0.045 ± 0.006	**0.054 ± 0.008**	0.045 ± 0.006	0.051 ± 0.008
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 0.5$	0.043 ± 0.004	**0.052 ± 0.005**	0.048 ± 0.004	0.051 ± 0.005
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 1$	0.023 ± 0.005	0.035 ± 0.005	**0.049 ± 0.006**	**0.049 ± 0.006**
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = -0.5$	0.043 ± 0.004	**0.054 ± 0.006**	0.045 ± 0.005	0.051 ± 0.006
diabetes	0.035 ± 0.011	0.049 ± 0.013	0.038 ± 0.012	**0.050 ± 0.013**
music origin	0.042 ± 0.007	**0.047 ± 0.008**	0.048 ± 0.006	0.051 ± 0.007
rf1	0.037 ± 0.002	0.037 ± 0.002	0.049 ± 0.003	**0.051 ± 0.002**
rf2	0.037 ± 0.002	**0.036 ± 0.002**	0.048 ± 0.002	0.050 ± 0.002
scm1d	0.021 ± 0.001	0.026 ± 0.001	0.050 ± 0.002	**0.052 ± 0.002**
scm20d	0.021 ± 0.002	0.027 ± 0.002	0.049 ± 0.002	**0.051 ± 0.002**

full suite of experimental results. We evaluate the tested methods by measuring the error rate and the volume of the prediction region for different global targeted error rates ϵ_h.

Table 2 and Figure 1(a) show the measured error rate while Table 3 and Figure 1(b) present the prediction region's volume for the synthetic data set with $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$. For this data set, DMT-CP produces the smallest prediction regions for $\epsilon_h \leq 0.8$ while Max-Rank narrowly takes the lead for $\epsilon_h \geq 0.85$. Figure 2(a) shows that the underlying model's residuals follow the distribution of the noise term e and their lack of correlation is displayed in Figure 2(b). Finally, Figure 3 shows how the various methods partition the residual space. Because $y \in \mathbb{R}^2$, the rectangular prediction region is the Cartesian product of two prediction intervals. These intervals are produced by two conformal predictors $\Gamma^{\epsilon_m}_m$, one for each dimension of y. As shown in Equation (4), these $\Gamma^{\epsilon_m}_m$ utilise $\alpha_{m,(1-\epsilon_m)}$, the $1-\epsilon_m$ quantile of the residuals of dimension m. For each tested global significance level ϵ_h, Figure 3 contains a rectangle with vertices $(0,0), (\alpha_{1,(1-\epsilon_1)}, 0), (0, \alpha_{2,(1-\epsilon_2)})$ and $(\alpha_{1,(1-\epsilon_1)}, \alpha_{2,(1-\epsilon_2)})$. Unfolding such a rectangle along one axis and unfolding the result along the second axis gives the shape of the prediction region for a chosen global targeted error rate ϵ_h. This representation visualizes the different behaviours of the tested methods. Bonferroni predictors produce prediction regions that progressively grow in both dimensions when the global targeted error rate ϵ_h decreases, assigning equal and decreasing local targeted error rates ϵ_m to each dimension. Copula Conformal Predictors and Max-Rank display a similar behaviour but assign larger local targeted error rates than Bonferroni predictors, leading to smaller prediction regions. DCP-MT follows a different strategy, assigning a small portion of its total error budget to the first dimension, leading to a large interval which only changes minimally, while increasing the interval in the second dimension when the global targeted error rate ϵ_h decreases.

Tables 4 and 5 compare the tested method's measured error rates and prediction region volumes for all synthetic and real-world data sets at $\epsilon_h = 0.05$, providing a broader overview over tested methods' performance. DCP-MT produces the smallest prediction regions when the residuals' dimensions are weakly correlated. This is especially true for small values of ϵ_h on the as can be seen in Tables 7, 13, 15, and 17, where Max-Rank is generally more informationally efficient for larger ϵ_h. Max-Rank also produces smaller hyperrectangles when the residuals' dimensions are strongly correlated such as for the synthetic data set with $e \sim \mathcal{N}([0,0]^\top], \boldsymbol{\sigma})$ and $\sigma = 1$.

Table 5: Hyperrectangle volumes acoss all experiments for $\epsilon_h = 0.05$. Bold values indicate that the method produced the smallest prediction region while underlined values highlight instances for which the measured error rate ≥ 0.05.

Data Set	Bonferroni	DMT-CP	Empirical Copula	Max-Rank
synt, $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$	0.374 ± 0.013	**0.323 ± 0.011**	0.369 ± 0.013	0.354 ± 0.011
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 0$	6.644 ± 0.246	**6.312 ± 0.205**	6.628 ± 0.232	6.347 ± 0.201
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 0.5$	4.390 ± 0.126	**4.124 ± 0.114**	4.263 ± 0.122	4.159 ± 0.105
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 1$	5.928 ± 0.305	5.682 ± 0.281	**4.552 ± 0.221**	**4.552 ± 0.221**
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = -0.5$	6.068 ± 0.230	**5.621 ± 0.150**	5.962 ± 0.239	5.720 ± 0.230
diabetes	7.919 ± 0.943	6.604 ± 0.752	7.627 ± 0.923	**6.498 ± 0.784**
music origin	24.417 ± 2.147	**22.860 ± 1.602**	23.433 ± 1.969	23.237 ± 2.002
rf1	172.734 ± 81.099	32.901 ± 14.016	38.125 ± 18.702	**32.341 ± 15.305**
rf2	87.792 ± 46.649	**14.946 ± 8.429**	19.484 ± 12.388	16.539 ± 10.866
scm1d	11.387E9 ± 5.455E9	3.753E9 ± 1.420E9	41.820E6 ± 11.540E6	**32.774E6 ± 8.828E6**
scm20d	108.772E9 ± 36.307E9	35.955E9 ± 11.015E9	1.172E9 ± 0.345E9	**0.942E9 ± 0.301E9**

Appendix A contains the detailed experimental results for all data sets. This includes the measured error rates and prediction region volumes for different ϵ_h and the residual correlation matrix. For data sets with a two-dimensional label space it also provides a plot of the residuals and plots of the methods' partition of the residual space for different ϵ_h.

7. Discussion

This section discusses the empirical results, comparing DCP-MT to Bonferroni predictors, copula conformal predictors, and Max-Rank.

The theoretical differences discussed in Section 5 are confirmed by the experimental results shown in Section 6.3. DCP-MT produces smaller prediction regions than copula conformal predictors and Max-Rank for data sets whose nonconformity scores are weakly correlated across their dimensions and/or have differing marginal distributions. In the case of strong positive correlation between the dimensions' nonconformity scores Max-Rank excels as expected given the method's design. These results agree with the experimental results of Timans et al. (2025), reporting that Max-Rank is most efficient in the presence of strong positive correlation and will behave similarly to Bonferroni predictors in the absence of correlation. The behaviour of copula conformal predictors using the empirical copula generally resembles the one of Max-Rank while producing hyperrectangles that are slightly larger. Bonferroni conformal predictors are not competitive, producing the largest prediction regions. In some instances DCP-MT, Copula conformal predictors, and Max-Rank produce errors with a frequency slightly above the targeted error rate which is consistent with the results in Messoudi et al. (2021). This can not be observed for Bonferroni conformal predictors, which also produce the largest prediction regions.

Because of the different approaches the discussed methods take, we recommend to choose a method based on the characteristics of the nonconformity scores a user encounters. Depending on the data set's size and the number of dimension of the label space, the computational complexity might be another factor the user should consider. Appendix B contains a preliminary theoretical analysis of the discussed methods' computational complexity as well as the wall time measured for the experiments presented in Section 6.3.

8. Conclusion and Future Research

In this article, we present DCP-MT, a novel approach to determine the local targeted error rates of multiple single-target inductive conformal predictors when combining them to form a hyperrectangular prediction regions. DCP-MT minimises the hyperrectangle's volume by dynamically allocating a global error budget between the local targeted error rates using integer linear programming. The theoretical analysis of DCP-MT proves its validity under exchangeability of the examples in the data set and a symmetric training algorithm for the underlying model.

The experimental results on four synthetic and six public real-world data sets serve to validate DCP-MT's theoretical properties and provide insights into the method's behaviour. We compare DCP-MT to Bonferroni predictors, copula conformal predictors and Max-Rank. The results show that DCP-MT excels when the nonconformity scores are weakly correlated across the dimensions of the label space and when the distributions of the nonconformity scores differ between dimensions. In these situations, assigning the same local targeted error rate to all their single-target inductive conformal predictors limits the informational efficiency of the related methods. Plots that show the partition of the residual space for different global targeted error rates visualise the differences in the tested method's behaviours.

For future research, we would like to investigate the interactions of DCP-MT with input dependent nonconformity metrics such as utilised by Messoudi et al. (2021) and compare it to the method proposed by Cleaveland et al. (2024). It would also be interesting to extend the approach of optimising the prediction region's volume to other methods like copula conformal predictors. Finally, we believe that for multi-target regression applications with high-dimensional label spaces and very large calibration sets, DCP-MT might benefit from additional computational optimisation, such as using a minibatch approach.

References

Niccolò Ajroldi, Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for two-dimensional functional time series. *Computational Statistics & Data Analysis*, 187:107821, July 2023. ISSN 01679473. doi: 10.1016/j.csda.2023.107821. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947323001329.

Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, December 2022. URL http://arxiv.org/abs/2107.07511. arXiv:2107.07511 [cs, math, stat].

Rina Foygel Barber, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *arXiv:2202.13415 [stat]*, March 2022. URL http://arxiv.org/abs/2202.13415. arXiv: 2202.13415.

Matthew Cleaveland, Insup Lee, George J. Pappas, and Lars Lindemann. Conformal Prediction Regions for Time Series Using Linear Complementarity Programming. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):20984–20992, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i19.30089. URL https://ojs.aaai.org/index.php/AAAI/article/view/30089.

Victor Dheur, Matteo Fontana, Yorick Estievenart, Naomi Desobry, and Souhaib Ben Taieb. A Unified Comparative Study with Generalized Conformity Scores for Multi-Output Conformal Regression, February 2025. URL https://arxiv.org/abs/2501.10533. _eprint: 2501.10533.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Distribution-Free Prediction Bands for Multivariate Functional Time Series: an Application to the Italian Gas Market, January 2024. URL http://arxiv.org/abs/2107.00527. arXiv:2107.00527 [stat].

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, April 2004. doi: 10.1214/009053604000000067.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving Risk Control in Online Learning Settings, January 2023. URL http://arxiv.org/abs/2205.09095. arXiv:2205.09095 [cs, stat].

Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: A unified review of theory and new challenges. *Bernoulli*, 29 (1), February 2023. ISSN 1350-7265. doi: 10.3150/21-BEJ1447. URL https://projecteuclid.org/journals/bernoulli/volume-29/issue-1/Conformal-prediction--A-unified-review-of-theory-and-new/10.3150/21-BEJ1447.full.

Chancellor Johnstone and Bruce Cox. Conformal uncertainty sets for robust optimization. In Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen, editors, *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pages 72–90. PMLR, September 2021. URL https://proceedings.mlr.press/v152/johnstone21a.html.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Conformal multi-target regression using neural networks. In Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Giovanni Cherubin, editors, *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 65–83. PMLR, September 2020. URL https://proceedings.mlr.press/v128/messoudi20a.html.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101, June 2021. ISSN 00313203. doi: 10.1016/j.patcog.2021.108101. URL https://linkinghub.elsevier.com/retrieve/pii/S0031320321002880.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Ellipsoidal conformal inference for Multi-Target Regression. In Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo, and Lars Carlsson, editors, *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, volume 179 of *Proceedings of Machine Learning Research*, pages 294–306. PMLR, August 2022. URL https://proceedings.mlr.press/v179/messoudi22a.html.

Jelmer Neeven and Evgueni Smirnov. Conformal stacked weather forecasting. In Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Ralf Peeters, editors, *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*, volume 91 of *Proceedings of Machine Learning Research*, pages 220–233. PMLR, June 2018. URL https://proceedings.mlr.press/v91/neeven18a.html.

Harris Papadopoulos and Haris Haralambous. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851, October 2011. ISSN 08936080. doi: 10.1016/j.neunet.2011.05.008. URL https://linkinghub.elsevier.com/retrieve/pii/S089360801100150X.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive Confidence Machines for Regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002*, pages 345–356, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-36755-0.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, November 2011. ISSN 1532-4435. Publisher: JMLR.org.

Filip Schlembach, Evgueni Smirnov, and Irena Koprinska. Conformal Multistep-Ahead Multivariate Time-Series Forecasting. In Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo, and Lars Carlsson, editors, *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, volume 179 of *Proceedings of Machine Learning Research*, pages 316–318. PMLR, August 2022. URL https://proceedings.mlr.press/v179/schlembach22a.html.

Filip Schlembach, Evgueni Smirnov, Irena Koprinska, and Mark H. M. Winands. Conformal multistep-ahead multivariate time-series forecasting. *Machine Learning*, 114(7):165, June 2025. ISSN 1573-0565. doi: 10.1007/s10994-024-06722-9. URL https://doi.org/10.1007/s10994-024-06722-9.

M. Sklar. Fonctions de répartition à N dimensions et leurs marges. *Annales de l'ISUP*, VIII (3):229–231, 1959. URL https://hal.science/hal-04094463. Publisher: Publications de l'Institut de Statistique de l'Université de Paris.

Kamilė Stankevičiūtė, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal Time-Series Forecasting. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/312f1ba2a72318edaaa995a67835fad5-Abstract.html.

Sophia Sun and Rose Yu. Copula Conformal Prediction for Multi-step Time Series Forecasting, March 2024. URL https://arxiv.org/abs/2212.03281. _eprint: 2212.03281.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, Christian A. Naesseth, and Eric Nalisnick. Max-Rank: Efficient Multiple Testing for Conformal Prediction, March 2025. URL https://arxiv.org/abs/2311.10900. _eprint: 2311.10900.

Paolo Toccaceli. Introduction to conformal predictors. *Pattern Recognition*, 124:108507, April 2022. ISSN 00313203. doi: 10.1016/j.patcog.2021.108507. URL https://linkinghub.elsevier.com/retrieve/pii/S003132032100683X.

Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. MULAN: A Java Library for Multi-Label Learning. *Journal of Machine Learning Research*, 12(71):2411–2414, 2011. URL http://jmlr.org/papers/v12/tsoumakas11a.html.

Vladimir Vovk. Transductive conformal predictors. In Harris Papadopoulos, Andreas S. Andreou, Lazaros Iliadis, and Ilias Maglogiannis, editors, *Artificial Intelligence Applications and Innovations*, pages 348–360, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-41142-7. doi: 10.1007/978-3-642-41142-7_36.

Vladimir Vovk, A. Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, New York, 2005. ISBN 978-0-387-00152-4 978-0-387-25061-8.

Vladimir Vovk, Bin Wang, and Ruodu Wang. Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, 50(1), February 2022. ISSN 0090-5364. doi: 10.1214/21-AOS2109. URL https://projecteuclid.org/journals/annals-of-statistics/volume-50/issue-1/Admissible-ways-of-merging-p-values-under-arbitrary-dependence/10.1214/21-AOS2109.full.

Fang Zhou, Q. Claire, and Ross D. King. Predicting the Geographical Origin of Music. In *2014 IEEE International Conference on Data Mining*, pages 1115–1120, 2014. doi: 10.1109/ICDM.2014.73.

Appendix A. Additional Experimental Results

This Section contains the additional experimental results. Each subsection corresponds to one of the data sets from Section 6.1. In the tables, underlined values indicate instances for which the measured error rate exceeds the targeted error rate while bold values indicate that the method produced the smallest prediction region.

A.1. Synthetic, Multivariate Normal Error, $\sigma = 0$

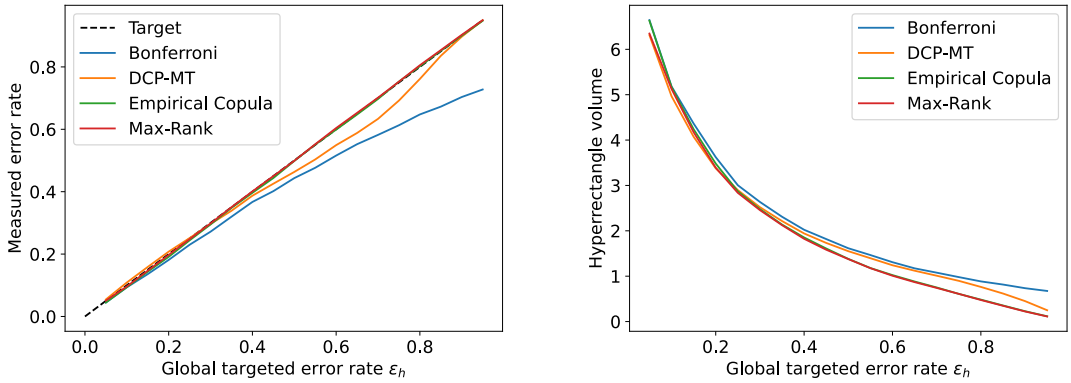
Additional results for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 0$

Table 6: Multi-tests method's mean error rates and standard deviations for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 0$.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	0.045 ± 0.006	**0.054 ± 0.008**	0.045 ± 0.006	0.051 ± 0.008
0.10	0.093 ± 0.006	**0.109 ± 0.007**	0.094 ± 0.006	0.096 ± 0.008
0.15	0.136 ± 0.007	**0.159 ± 0.010**	0.144 ± 0.009	0.147 ± 0.009
0.20	0.181 ± 0.008	**0.208 ± 0.011**	0.191 ± 0.010	0.197 ± 0.010
0.25	0.230 ± 0.009	0.251 ± 0.012	0.244 ± 0.012	**0.248 ± 0.012**
0.30	0.272 ± 0.013	0.297 ± 0.014	0.295 ± 0.013	**0.298 ± 0.013**
0.35	0.320 ± 0.012	0.339 ± 0.009	0.348 ± 0.012	**0.350 ± 0.011**
0.40	0.367 ± 0.012	0.387 ± 0.011	0.396 ± 0.013	**0.401 ± 0.013**
0.45	0.402 ± 0.011	0.426 ± 0.014	0.444 ± 0.014	**0.450 ± 0.013**
0.50	0.444 ± 0.012	0.463 ± 0.015	0.498 ± 0.017	**0.499 ± 0.017**
0.55	0.477 ± 0.014	0.503 ± 0.013	0.551 ± 0.015	**0.552 ± 0.014**
0.60	0.516 ± 0.014	0.549 ± 0.016	0.599 ± 0.015	**0.604 ± 0.015**
0.65	0.552 ± 0.012	0.588 ± 0.018	0.647 ± 0.014	**0.653 ± 0.014**
0.70	0.582 ± 0.014	0.633 ± 0.017	0.697 ± 0.009	**0.701 ± 0.010**
0.75	0.613 ± 0.013	0.692 ± 0.018	**0.753 ± 0.012**	0.753 ± 0.011
0.80	0.648 ± 0.013	0.761 ± 0.026	0.802 ± 0.010	**0.805 ± 0.009**
0.85	0.673 ± 0.011	0.835 ± 0.016	0.851 ± 0.009	**0.854 ± 0.010**
0.90	0.703 ± 0.011	0.897 ± 0.013	0.900 ± 0.008	**0.903 ± 0.007**
0.95	0.728 ± 0.010	0.950 ± 0.007	0.948 ± 0.008	**0.950 ± 0.008**

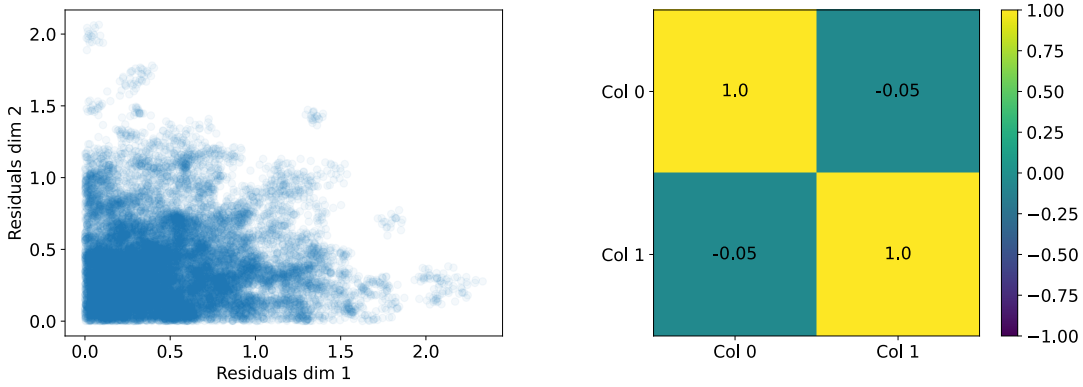
Table 7: Multi-tests method's mean hyperrectangle volumes for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 0$.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	6.644 ± 0.246	**6.312 ± 0.205**	6.628 ± 0.232	6.347 ± 0.201
0.10	5.182 ± 0.115	**4.966 ± 0.115**	5.171 ± 0.114	5.127 ± 0.138
0.15	4.364 ± 0.107	**4.073 ± 0.102**	4.238 ± 0.130	4.174 ± 0.117
0.20	3.623 ± 0.085	**3.378 ± 0.078**	3.482 ± 0.121	3.391 ± 0.113
0.25	3.005 ± 0.079	2.897 ± 0.075	2.872 ± 0.083	**2.837 ± 0.084**
0.30	2.636 ± 0.077	2.520 ± 0.057	2.478 ± 0.072	**2.457 ± 0.073**
0.35	2.310 ± 0.063	2.216 ± 0.056	2.138 ± 0.063	**2.125 ± 0.061**
0.40	2.023 ± 0.058	1.942 ± 0.055	1.853 ± 0.066	**1.824 ± 0.064**
0.45	1.818 ± 0.053	1.733 ± 0.051	1.610 ± 0.067	**1.583 ± 0.063**
0.50	1.616 ± 0.052	1.551 ± 0.048	1.381 ± 0.060	**1.377 ± 0.059**
0.55	1.466 ± 0.049	1.396 ± 0.042	1.180 ± 0.051	**1.176 ± 0.051**
0.60	1.308 ± 0.045	1.241 ± 0.043	1.024 ± 0.046	**1.010 ± 0.045**
0.65	1.173 ± 0.041	1.121 ± 0.037	0.883 ± 0.035	**0.869 ± 0.036**
0.70	1.077 ± 0.036	1.010 ± 0.034	0.751 ± 0.026	**0.741 ± 0.026**
0.75	0.980 ± 0.031	0.897 ± 0.030	**0.610 ± 0.026**	0.610 ± 0.024
0.80	0.885 ± 0.029	0.763 ± 0.027	0.484 ± 0.020	**0.475 ± 0.019**
0.85	0.816 ± 0.028	0.618 ± 0.027	0.354 ± 0.016	**0.348 ± 0.017**
0.90	0.736 ± 0.026	0.451 ± 0.024	0.228 ± 0.011	**0.222 ± 0.011**
0.95	0.674 ± 0.023	0.249 ± 0.023	0.116 ± 0.013	**0.112 ± 0.013**



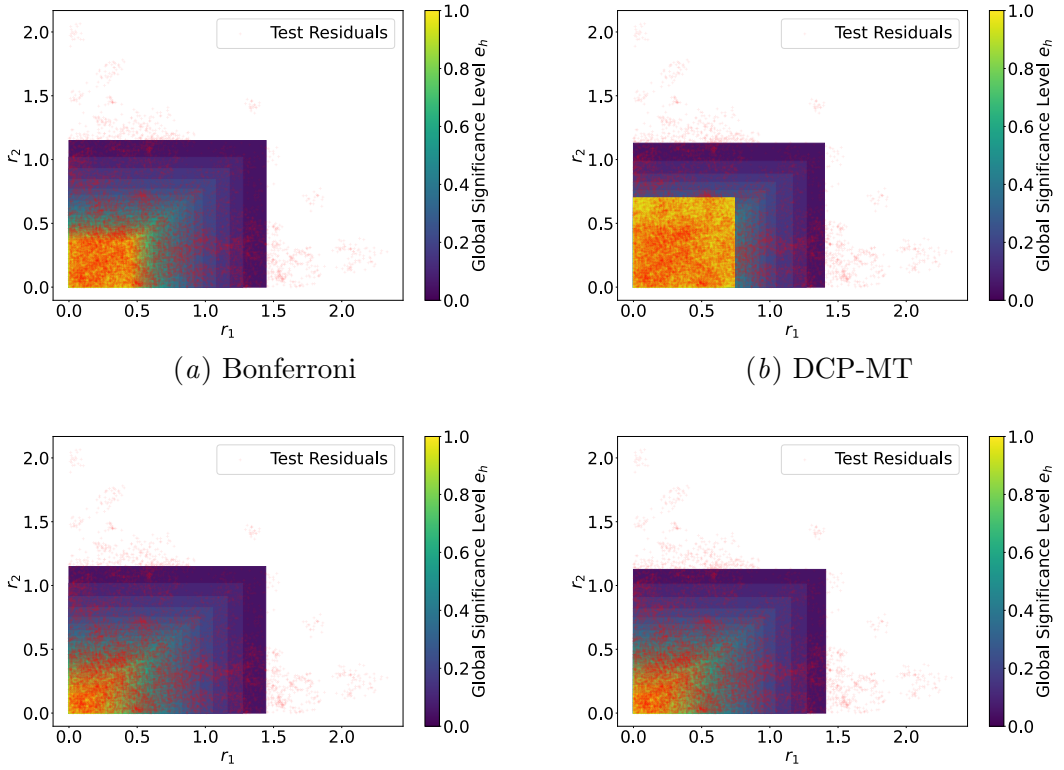
(*a*) Error Rate (*b*) Volume

Figure 4: Multi-tests method's mean error rates and volumes for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 0$.



(*a*) Residuals (*b*) Residual correlation matrix

Figure 5: Underlying model's absolute residuals and their correlation for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 0$.



(*a*) Bonferroni (*b*) DCP-MT

(*c*) Empirical Copula (*d*) Max-Rank

Figure 6: Multi-tests method's partition of the residual space for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 0$ averaged over 20 repetitions.

A.2. Synthetic, Multivariate Normal Error, $\sigma = 0.5$

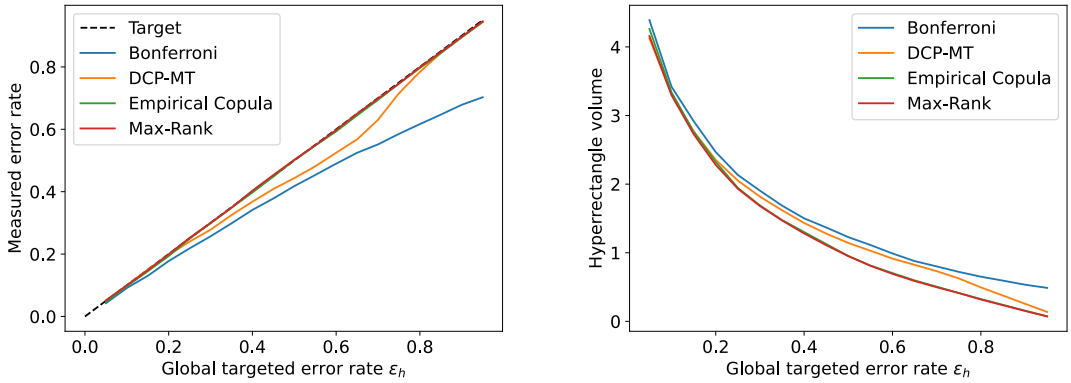
Additional results for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 0.5$

Table 8: Multi-tests method's mean error rates and standard deviations for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 0.5$.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	0.043 ± 0.004	**0.052 ± 0.005**	0.048 ± 0.004	0.051 ± 0.005
0.10	0.091 ± 0.007	0.099 ± 0.007	0.097 ± 0.007	**0.100 ± 0.008**
0.15	0.130 ± 0.007	0.150 ± 0.009	0.145 ± 0.006	**0.148 ± 0.007**
0.20	0.178 ± 0.007	0.197 ± 0.010	0.194 ± 0.009	**0.199 ± 0.008**
0.25	0.218 ± 0.007	0.240 ± 0.008	0.250 ± 0.008	**0.252 ± 0.008**
0.30	0.257 ± 0.008	0.278 ± 0.009	0.298 ± 0.009	**0.300 ± 0.009**
0.35	0.298 ± 0.008	0.325 ± 0.015	0.348 ± 0.011	**0.349 ± 0.011**
0.40	0.342 ± 0.010	0.368 ± 0.012	0.397 ± 0.011	**0.403 ± 0.011**
0.45	0.378 ± 0.011	0.409 ± 0.012	0.449 ± 0.010	**0.453 ± 0.011**
0.50	0.418 ± 0.010	0.443 ± 0.011	0.501 ± 0.012	**0.502 ± 0.011**
0.55	0.454 ± 0.008	0.481 ± 0.015	0.548 ± 0.011	**0.549 ± 0.012**
0.60	0.490 ± 0.009	0.525 ± 0.016	0.593 ± 0.014	**0.598 ± 0.014**
0.65	0.525 ± 0.010	0.568 ± 0.015	0.645 ± 0.012	**0.650 ± 0.012**
0.70	0.552 ± 0.013	0.630 ± 0.017	0.694 ± 0.012	**0.699 ± 0.012**
0.75	0.585 ± 0.013	0.715 ± 0.022	0.747 ± 0.013	**0.746 ± 0.013**
0.80	0.617 ± 0.010	0.784 ± 0.017	0.796 ± 0.010	**0.799 ± 0.010**
0.85	0.647 ± 0.010	0.845 ± 0.016	0.844 ± 0.009	**0.848 ± 0.007**
0.90	0.678 ± 0.009	0.896 ± 0.012	0.895 ± 0.008	**0.897 ± 0.008**
0.95	0.703 ± 0.010	0.944 ± 0.008	0.944 ± 0.006	**0.946 ± 0.005**

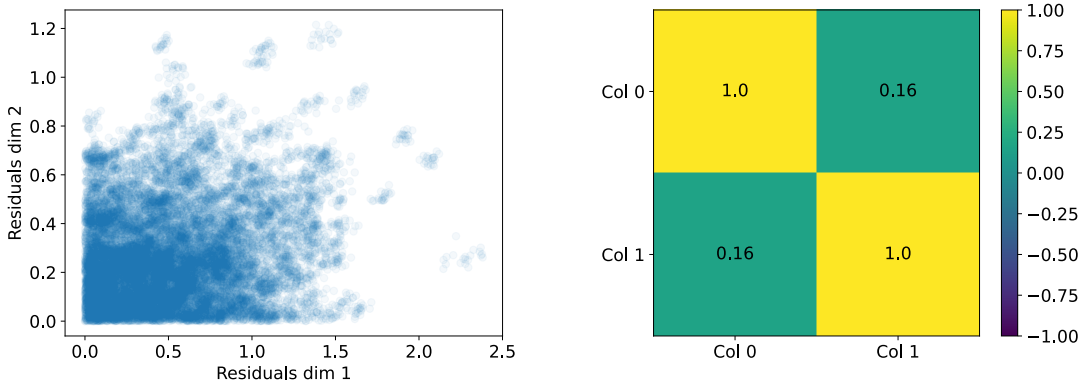
Table 9: Multi-tests method's mean hyperrectangle volumes for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 0.5$.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	4.390 ± 0.126	**4.124 ± 0.114**	4.263 ± 0.122	4.159 ± 0.105
0.10	3.419 ± 0.088	3.313 ± 0.080	3.338 ± 0.082	**3.296 ± 0.079**
0.15	2.919 ± 0.073	2.770 ± 0.071	2.763 ± 0.058	**2.729 ± 0.066**
0.20	2.468 ± 0.062	2.353 ± 0.048	2.323 ± 0.057	**2.280 ± 0.053**
0.25	2.134 ± 0.045	2.053 ± 0.036	1.946 ± 0.044	**1.933 ± 0.040**
0.30	1.905 ± 0.039	1.818 ± 0.033	1.690 ± 0.037	**1.681 ± 0.038**
0.35	1.687 ± 0.031	1.619 ± 0.034	1.477 ± 0.040	**1.471 ± 0.036**
0.40	1.501 ± 0.032	1.434 ± 0.032	1.302 ± 0.034	**1.281 ± 0.033**
0.45	1.368 ± 0.031	1.279 ± 0.031	1.127 ± 0.035	**1.110 ± 0.037**
0.50	1.228 ± 0.028	1.143 ± 0.029	0.954 ± 0.036	**0.950 ± 0.034**
0.55	1.113 ± 0.028	1.030 ± 0.027	0.813 ± 0.026	**0.810 ± 0.027**
0.60	0.990 ± 0.026	0.915 ± 0.023	0.703 ± 0.022	**0.693 ± 0.021**
0.65	0.879 ± 0.025	0.824 ± 0.021	0.597 ± 0.022	**0.588 ± 0.022**
0.70	0.801 ± 0.024	0.731 ± 0.016	0.505 ± 0.022	**0.497 ± 0.023**
0.75	0.721 ± 0.020	0.626 ± 0.015	0.413 ± 0.021	**0.413 ± 0.021**
0.80	0.651 ± 0.017	0.495 ± 0.017	0.325 ± 0.017	**0.319 ± 0.017**
0.85	0.594 ± 0.020	0.377 ± 0.014	0.241 ± 0.016	**0.235 ± 0.015**
0.90	0.534 ± 0.018	0.256 ± 0.015	0.156 ± 0.010	**0.152 ± 0.010**
0.95	0.488 ± 0.017	0.137 ± 0.013	0.074 ± 0.006	**0.071 ± 0.006**



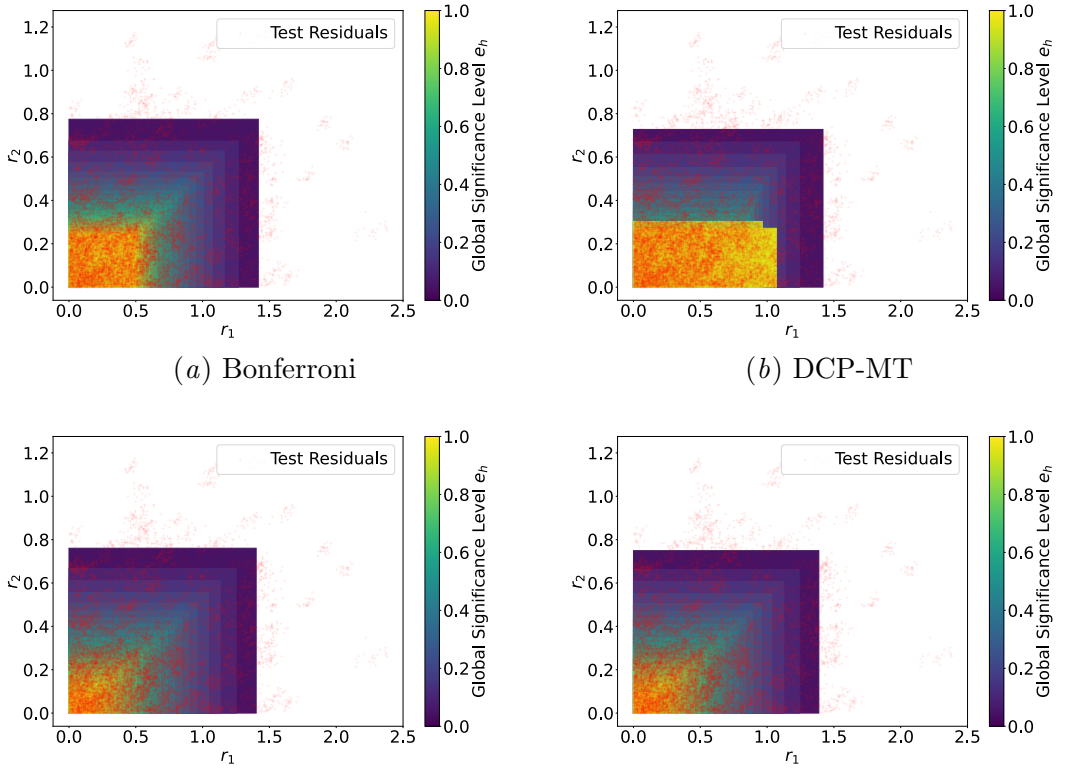
(a) Error Rate (b) Volume

Figure 7: Multi-tests method's mean error rates and volumes for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 0.5$.



(a) Residuals (b) Residual correlation matrix

Figure 8: Underlying model's absolute residuals and their correlation for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 0.5$.



(*a*) Bonferroni (*b*) DCP-MT

(*c*) Empirical Copula (*d*) Max-Rank

Figure 9: Multi-tests method's partition of the residual space for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 0.5$ averaged over 20 repetitions.

A.3. Synthetic, Multivariate Normal Error, $\sigma = 1$

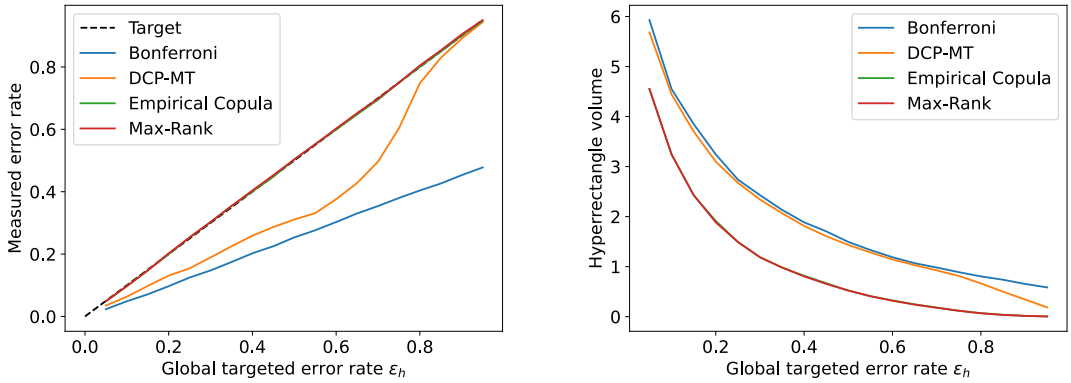
Additional results for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 1$.

Table 10: Multi-tests method's mean error rates and standard deviations for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 1$.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	0.023 ± 0.005	0.035 ± 0.005	**0.049 ± 0.006**	0.049 ± 0.006
0.10	0.049 ± 0.006	0.063 ± 0.007	**0.097 ± 0.008**	0.097 ± 0.008
0.15	0.071 ± 0.006	0.098 ± 0.010	**0.148 ± 0.009**	0.148 ± 0.009
0.20	0.097 ± 0.008	0.131 ± 0.010	0.200 ± 0.008	**0.203 ± 0.009**
0.25	0.125 ± 0.008	0.154 ± 0.011	**0.254 ± 0.009**	0.254 ± 0.009
0.30	0.148 ± 0.009	0.189 ± 0.015	**0.303 ± 0.010**	0.303 ± 0.010
0.35	0.174 ± 0.009	0.225 ± 0.021	**0.354 ± 0.011**	0.354 ± 0.011
0.40	0.203 ± 0.009	0.259 ± 0.020	0.401 ± 0.013	**0.404 ± 0.013**
0.45	0.225 ± 0.009	0.287 ± 0.013	0.448 ± 0.014	**0.453 ± 0.014**
0.50	0.254 ± 0.009	0.311 ± 0.018	**0.504 ± 0.009**	0.504 ± 0.009
0.55	0.276 ± 0.010	0.331 ± 0.019	**0.552 ± 0.011**	0.552 ± 0.011
0.60	0.303 ± 0.010	0.376 ± 0.035	0.599 ± 0.016	**0.603 ± 0.016**
0.65	0.330 ± 0.011	0.428 ± 0.046	0.648 ± 0.015	**0.652 ± 0.015**
0.70	0.354 ± 0.011	0.496 ± 0.044	0.695 ± 0.014	**0.698 ± 0.014**
0.75	0.380 ± 0.011	0.603 ± 0.048	**0.750 ± 0.013**	0.750 ± 0.013
0.80	0.404 ± 0.013	0.748 ± 0.042	0.801 ± 0.011	**0.805 ± 0.011**
0.85	0.426 ± 0.014	0.829 ± 0.021	0.849 ± 0.010	**0.853 ± 0.009**
0.90	0.453 ± 0.014	0.891 ± 0.009	0.900 ± 0.009	**0.905 ± 0.009**
0.95	0.478 ± 0.013	0.944 ± 0.009	0.947 ± 0.008	**0.950 ± 0.008**

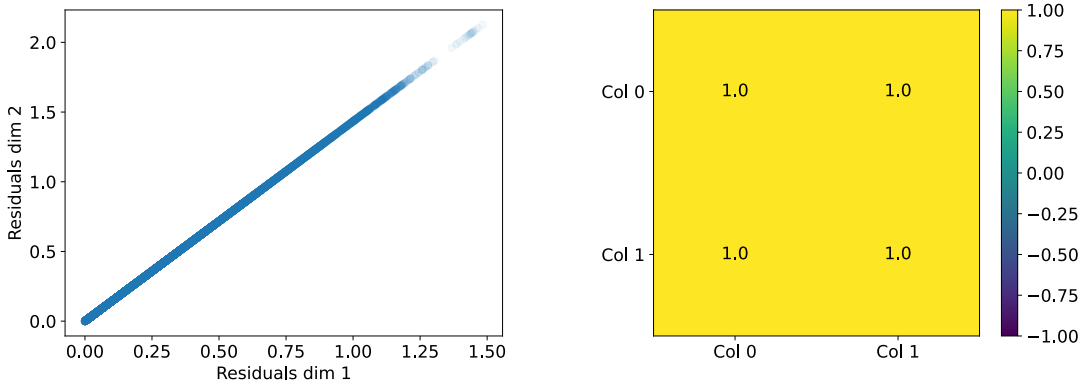
Table 11: Multi-tests method's mean hyperrectangle volumes for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 1$.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	5.928 ± 0.305	5.682 ± 0.281	**4.552 ± 0.221**	4.552 ± 0.221
0.10	4.552 ± 0.221	4.445 ± 0.189	**3.245 ± 0.149**	3.245 ± 0.149
0.15	3.852 ± 0.146	3.703 ± 0.152	**2.430 ± 0.100**	2.430 ± 0.100
0.20	3.245 ± 0.149	3.100 ± 0.131	1.913 ± 0.065	**1.886 ± 0.064**
0.25	2.741 ± 0.118	2.678 ± 0.108	**1.494 ± 0.066**	1.494 ± 0.066
0.30	2.430 ± 0.100	2.342 ± 0.093	**1.188 ± 0.045**	1.188 ± 0.045
0.35	2.136 ± 0.084	2.068 ± 0.077	**0.983 ± 0.039**	0.983 ± 0.039
0.40	1.886 ± 0.064	1.814 ± 0.068	0.817 ± 0.037	**0.805 ± 0.034**
0.45	1.706 ± 0.067	1.609 ± 0.057	0.669 ± 0.032	**0.655 ± 0.032**
0.50	1.494 ± 0.066	1.433 ± 0.052	**0.522 ± 0.023**	0.522 ± 0.023
0.55	1.332 ± 0.059	1.284 ± 0.048	**0.410 ± 0.019**	0.410 ± 0.019
0.60	1.188 ± 0.045	1.143 ± 0.043	0.324 ± 0.022	**0.318 ± 0.022**
0.65	1.069 ± 0.039	1.031 ± 0.038	0.247 ± 0.018	**0.241 ± 0.018**
0.70	0.983 ± 0.039	0.926 ± 0.037	0.182 ± 0.015	**0.178 ± 0.015**
0.75	0.890 ± 0.036	0.814 ± 0.036	**0.118 ± 0.013**	0.118 ± 0.013
0.80	0.805 ± 0.034	0.666 ± 0.036	0.071 ± 0.009	**0.068 ± 0.009**
0.85	0.739 ± 0.033	0.500 ± 0.031	0.037 ± 0.005	**0.035 ± 0.004**
0.90	0.655 ± 0.032	0.343 ± 0.024	0.017 ± 0.002	**0.015 ± 0.002**
0.95	0.585 ± 0.028	0.186 ± 0.020	0.005 ± 0.001	**0.004 ± 0.001**



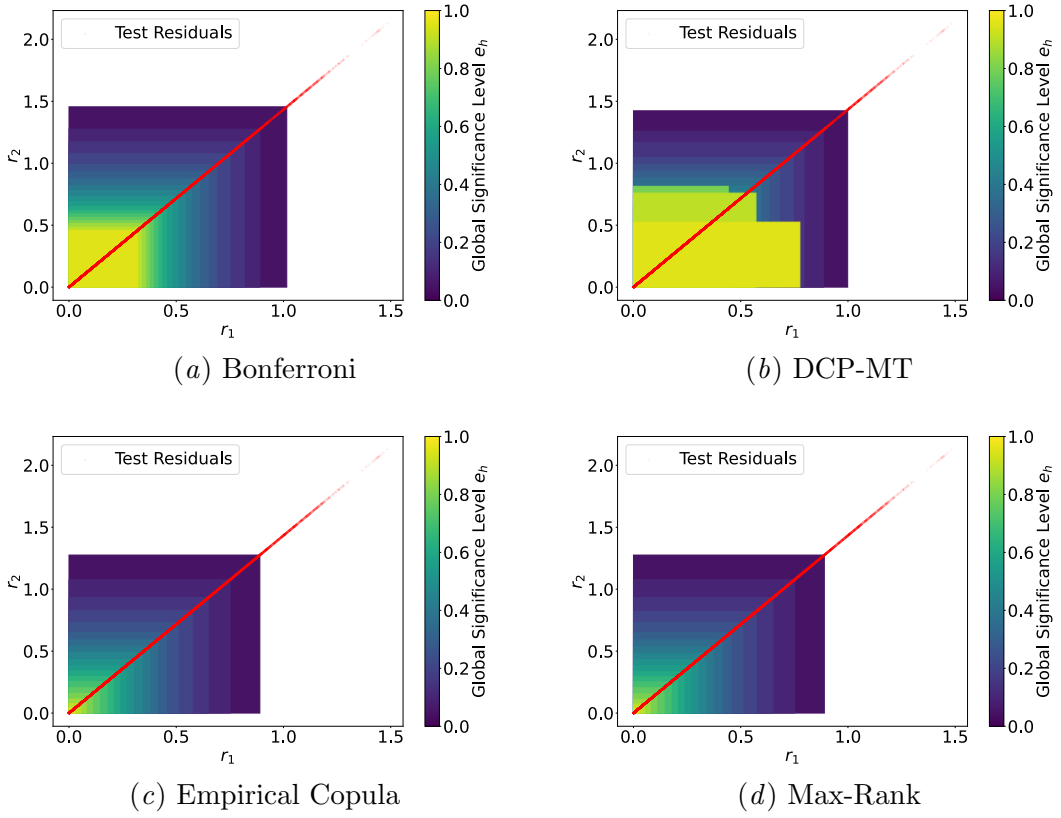
(*a*) Error Rate (*b*) Volume

Figure 10: Multi-tests method's mean error rates and volumes for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 1$.



(*a*) Residuals (*b*) Residual correlation matrix

Figure 11: Underlying model's absolute residuals and their correlation for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 1$.



(*a*) Bonferroni (*b*) DCP-MT

(*c*) Empirical Copula (*d*) Max-Rank

Figure 12: Multi-tests method's partition of the residual space for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = 1$ averaged over 20 repetitions.

A.4. Synthetic, Multivariate Normal Error, $\sigma = -0.5$

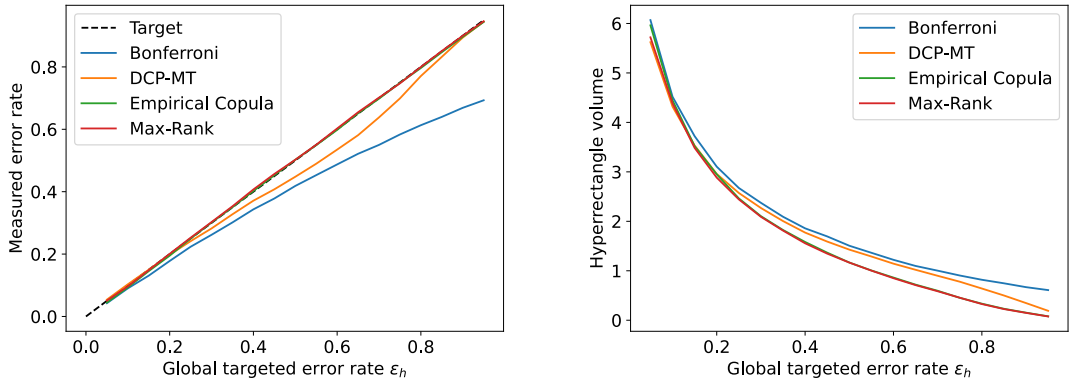
Additional results for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = -0.5$

Table 12: Multi-tests method's mean error rates and standard deviations for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = -0.5$.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	0.043 ± 0.004	**0.054 ± 0.006**	0.045 ± 0.005	0.051 ± 0.006
0.10	0.090 ± 0.007	**0.103 ± 0.011**	0.093 ± 0.008	0.097 ± 0.008
0.15	0.131 ± 0.007	0.150 ± 0.011	0.145 ± 0.010	**0.148 ± 0.009**
0.20	0.178 ± 0.010	0.198 ± 0.010	0.195 ± 0.011	**0.201 ± 0.011**
0.25	0.224 ± 0.009	0.241 ± 0.010	0.250 ± 0.010	**0.253 ± 0.011**
0.30	0.262 ± 0.011	0.282 ± 0.015	0.301 ± 0.012	**0.303 ± 0.012**
0.35	0.301 ± 0.013	0.328 ± 0.018	0.350 ± 0.013	**0.353 ± 0.014**
0.40	0.343 ± 0.014	0.371 ± 0.018	0.402 ± 0.014	**0.407 ± 0.014**
0.45	0.378 ± 0.015	0.408 ± 0.017	0.453 ± 0.013	**0.457 ± 0.013**
0.50	0.418 ± 0.015	0.448 ± 0.014	0.502 ± 0.015	**0.502 ± 0.015**
0.55	0.453 ± 0.013	0.490 ± 0.020	0.549 ± 0.016	**0.550 ± 0.016**
0.60	0.487 ± 0.016	0.535 ± 0.024	0.598 ± 0.017	**0.603 ± 0.016**
0.65	0.521 ± 0.013	0.581 ± 0.026	0.651 ± 0.013	**0.655 ± 0.013**
0.70	0.550 ± 0.014	0.638 ± 0.021	0.698 ± 0.014	**0.702 ± 0.014**
0.75	0.583 ± 0.013	0.699 ± 0.016	**0.749 ± 0.012**	0.747 ± 0.011
0.80	0.613 ± 0.013	0.771 ± 0.017	0.796 ± 0.009	**0.800 ± 0.009**
0.85	0.640 ± 0.012	0.832 ± 0.015	0.847 ± 0.011	**0.851 ± 0.011**
0.90	0.669 ± 0.010	0.894 ± 0.013	0.896 ± 0.007	**0.900 ± 0.007**
0.95	0.693 ± 0.010	0.945 ± 0.007	0.944 ± 0.006	**0.946 ± 0.005**

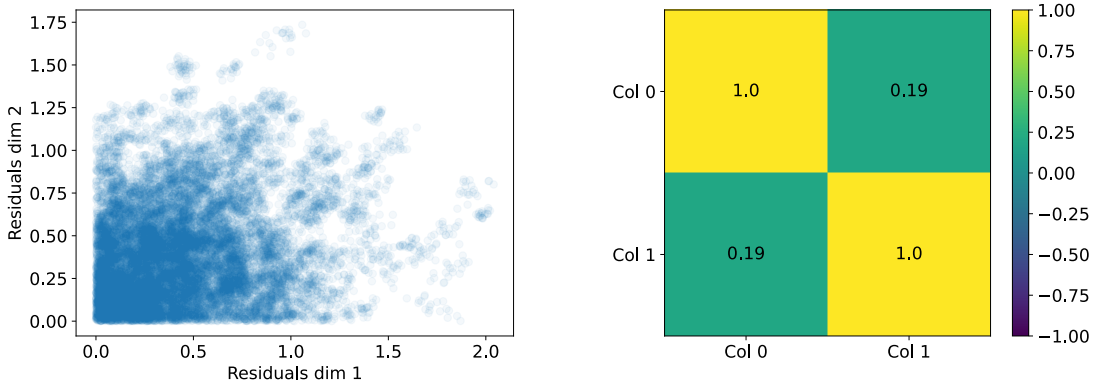
Table 13: Multi-tests method's mean hyperrectangle volumes for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = -0.5$.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	6.068 ± 0.230	**5.621 ± 0.150**	5.962 ± 0.239	5.720 ± 0.230
0.10	4.519 ± 0.109	**4.319 ± 0.123**	4.446 ± 0.130	4.384 ± 0.129
0.15	3.721 ± 0.105	3.532 ± 0.095	3.520 ± 0.102	**3.484 ± 0.087**
0.20	3.105 ± 0.078	2.958 ± 0.069	2.950 ± 0.071	**2.884 ± 0.069**
0.25	2.674 ± 0.063	2.573 ± 0.063	2.463 ± 0.068	**2.443 ± 0.076**
0.30	2.374 ± 0.070	2.269 ± 0.070	2.104 ± 0.064	**2.091 ± 0.059**
0.35	2.092 ± 0.057	2.005 ± 0.065	1.825 ± 0.056	**1.812 ± 0.059**
0.40	1.858 ± 0.056	1.772 ± 0.057	1.583 ± 0.050	**1.556 ± 0.048**
0.45	1.695 ± 0.059	1.590 ± 0.051	1.366 ± 0.047	**1.347 ± 0.045**
0.50	1.506 ± 0.049	1.428 ± 0.042	1.167 ± 0.048	**1.165 ± 0.048**
0.55	1.364 ± 0.049	1.289 ± 0.039	1.006 ± 0.039	**1.003 ± 0.038**
0.60	1.221 ± 0.051	1.144 ± 0.038	0.862 ± 0.042	**0.850 ± 0.041**
0.65	1.097 ± 0.039	1.020 ± 0.034	0.719 ± 0.036	**0.710 ± 0.035**
0.70	1.002 ± 0.035	0.898 ± 0.032	0.595 ± 0.028	**0.586 ± 0.028**
0.75	0.904 ± 0.035	0.779 ± 0.026	**0.453 ± 0.029**	0.455 ± 0.028
0.80	0.818 ± 0.033	0.641 ± 0.022	0.335 ± 0.026	**0.327 ± 0.025**
0.85	0.747 ± 0.030	0.500 ± 0.021	0.231 ± 0.018	**0.226 ± 0.018**
0.90	0.669 ± 0.022	0.347 ± 0.020	0.153 ± 0.011	**0.149 ± 0.011**
0.95	0.609 ± 0.022	0.190 ± 0.015	0.079 ± 0.009	**0.076 ± 0.008**



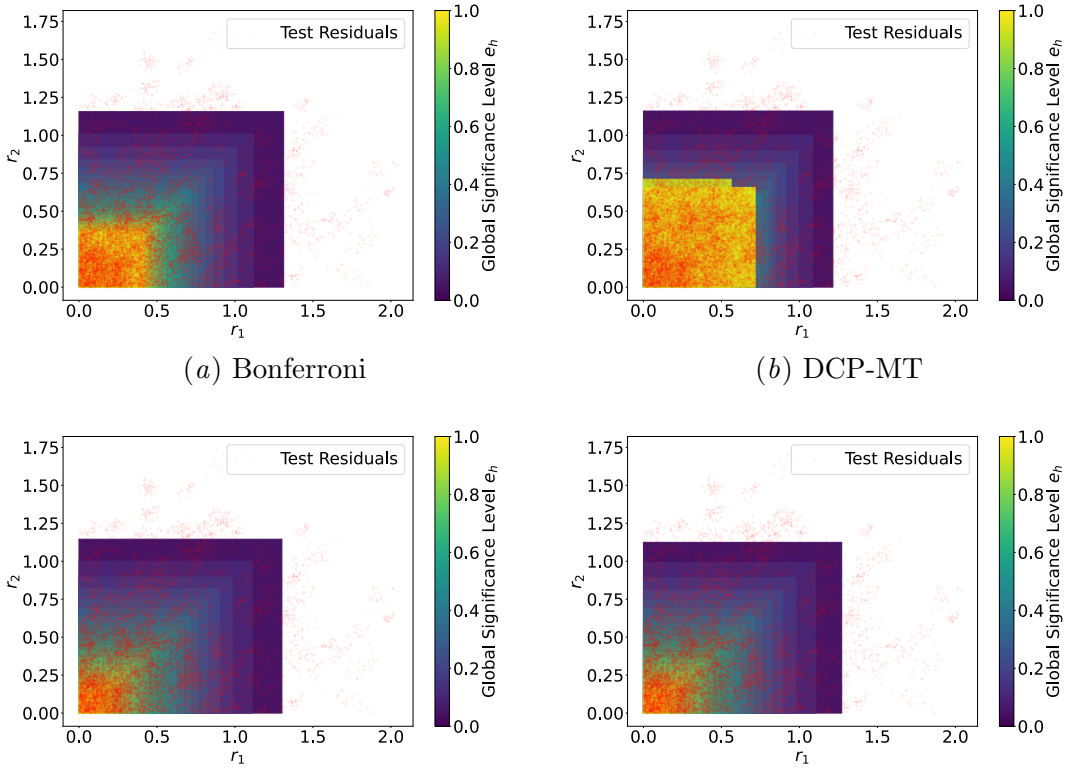
(a) Error Rate (b) Volume

Figure 13: Multi-tests method's mean error rates and volumes for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = -0.5$.



(a) Residuals (b) Residual correlation matrix

Figure 14: Underlying model's absolute residuals and their correlation for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = -0.5$.



(*a*) Bonferroni (*b*) DCP-MT

(*c*) Empirical Copula (*d*) Max-Rank

Figure 15: Multi-tests method's partition of the residual space for the synthetic data set with $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, and $\sigma = -0.5$ averaged over 20 repetitions.

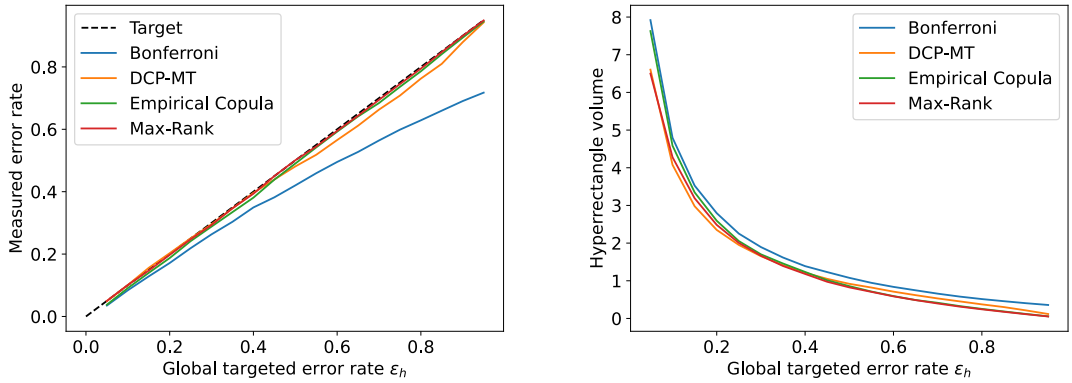
A.5. Diabetes

Table 14: Multi-tests method's mean error rates and standard deviations for the diabetes (Efron et al., 2004) data set.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	0.035 ± 0.011	0.049 ± 0.013	0.038 ± 0.012	**0.050 ± 0.013**
0.10	0.084 ± 0.017	**0.101 ± 0.020**	0.091 ± 0.018	0.101 ± 0.017
0.15	0.129 ± 0.021	**0.156 ± 0.025**	0.140 ± 0.024	0.147 ± 0.023
0.20	0.171 ± 0.024	**0.203 ± 0.024**	0.187 ± 0.024	0.197 ± 0.024
0.25	0.219 ± 0.025	**0.251 ± 0.023**	0.241 ± 0.026	0.247 ± 0.024
0.30	0.263 ± 0.025	**0.296 ± 0.025**	0.288 ± 0.025	0.293 ± 0.026
0.35	0.303 ± 0.024	0.347 ± 0.028	0.335 ± 0.024	**0.347 ± 0.024**
0.40	0.349 ± 0.020	0.396 ± 0.023	0.381 ± 0.024	**0.394 ± 0.025**
0.45	0.382 ± 0.021	0.439 ± 0.025	0.439 ± 0.028	**0.450 ± 0.028**
0.50	0.420 ± 0.025	0.481 ± 0.025	0.490 ± 0.027	**0.501 ± 0.027**
0.55	0.459 ± 0.027	0.518 ± 0.027	0.542 ± 0.026	**0.545 ± 0.027**
0.60	0.495 ± 0.029	0.566 ± 0.030	0.592 ± 0.029	**0.595 ± 0.026**
0.65	0.527 ± 0.025	0.612 ± 0.024	0.641 ± 0.022	**0.644 ± 0.023**
0.70	0.564 ± 0.023	0.663 ± 0.028	0.684 ± 0.023	**0.694 ± 0.022**
0.75	0.599 ± 0.021	0.708 ± 0.027	0.736 ± 0.022	**0.745 ± 0.022**
0.80	0.629 ± 0.022	0.762 ± 0.027	0.786 ± 0.020	**0.794 ± 0.020**
0.85	0.660 ± 0.023	0.810 ± 0.028	0.841 ± 0.017	**0.848 ± 0.016**
0.90	0.691 ± 0.022	0.879 ± 0.019	0.894 ± 0.018	**0.899 ± 0.017**
0.95	0.718 ± 0.023	0.943 ± 0.016	0.946 ± 0.013	**0.949 ± 0.013**

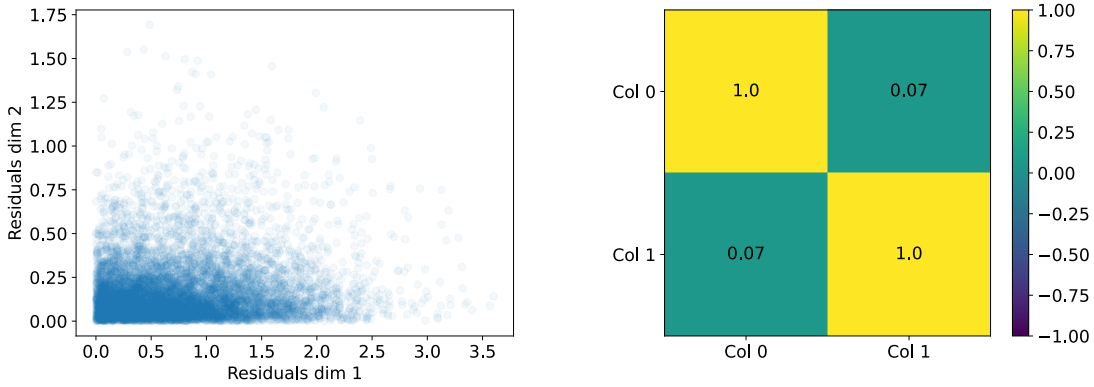
Table 15: Multi-tests method's mean hyperrectangle volumes for the diabetes (Efron et al., 2004) data set.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	7.919 ± 0.943	6.604 ± 0.752	7.627 ± 0.923	**6.498 ± 0.784**
0.10	4.795 ± 0.446	**4.076 ± 0.369**	4.584 ± 0.405	4.279 ± 0.362
0.15	3.526 ± 0.295	**2.979 ± 0.226**	3.346 ± 0.283	3.191 ± 0.274
0.20	2.795 ± 0.255	**2.349 ± 0.170**	2.593 ± 0.217	2.483 ± 0.200
0.25	2.252 ± 0.159	**1.951 ± 0.129**	2.053 ± 0.155	2.005 ± 0.138
0.30	1.895 ± 0.134	**1.653 ± 0.107**	1.702 ± 0.130	1.665 ± 0.135
0.35	1.617 ± 0.119	1.414 ± 0.089	1.458 ± 0.103	**1.392 ± 0.093**
0.40	1.389 ± 0.081	1.212 ± 0.073	1.234 ± 0.079	**1.181 ± 0.077**
0.45	1.230 ± 0.074	1.051 ± 0.062	1.013 ± 0.073	**0.973 ± 0.072**
0.50	1.078 ± 0.069	0.919 ± 0.056	0.858 ± 0.055	**0.824 ± 0.052**
0.55	0.945 ± 0.064	0.820 ± 0.053	0.714 ± 0.046	**0.706 ± 0.048**
0.60	0.838 ± 0.054	0.711 ± 0.048	0.594 ± 0.039	**0.586 ± 0.034**
0.65	0.747 ± 0.046	0.618 ± 0.039	0.489 ± 0.025	**0.487 ± 0.029**
0.70	0.658 ± 0.041	0.531 ± 0.037	0.415 ± 0.027	**0.397 ± 0.025**
0.75	0.580 ± 0.035	0.453 ± 0.034	0.328 ± 0.027	**0.314 ± 0.027**
0.80	0.514 ± 0.028	0.374 ± 0.034	0.254 ± 0.021	**0.242 ± 0.020**
0.85	0.458 ± 0.030	0.302 ± 0.032	0.185 ± 0.020	**0.176 ± 0.018**
0.90	0.405 ± 0.028	0.213 ± 0.029	0.119 ± 0.016	**0.112 ± 0.015**
0.95	0.357 ± 0.026	0.118 ± 0.018	0.056 ± 0.011	**0.052 ± 0.010**



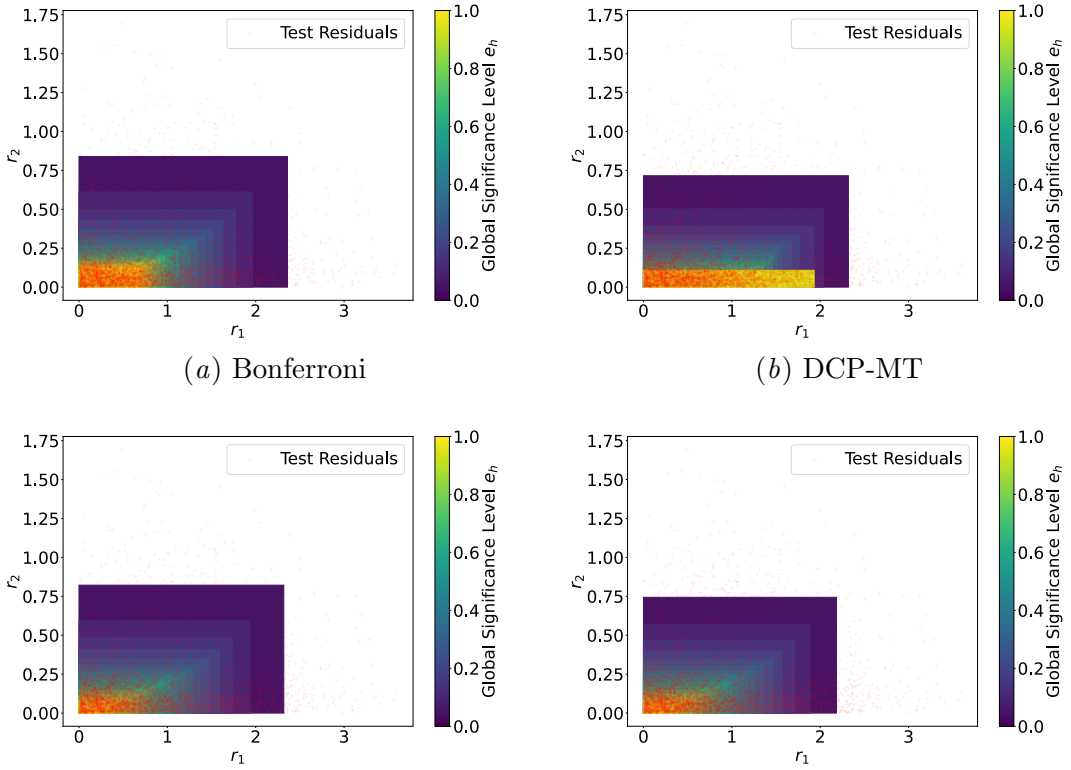
(*a*) Error Rate (*b*) Volume

Figure 16: Multi-tests method's mean error rates and volumes for the diabetes (Efron et al., 2004) data set.



(*a*) Residuals (*b*) Residual correlation matrix

Figure 17: Underlying model's absolute residuals and their correlation for the diabetes (Efron et al., 2004) data set.



(*a*) Bonferroni (*b*) DCP-MT

(*c*) Empirical Copula (*d*) Max-Rank

Figure 18: Multi-tests method's partition of the residual space for the diabetes (Efron et al., 2004) data set.

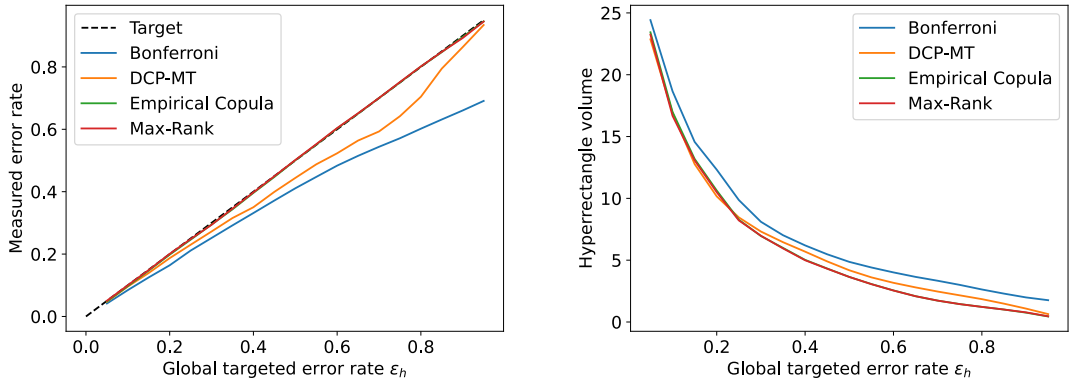
A.6. Music Origin

Table 16: Multi-tests method's mean error rates and standard deviations for the music origin (Zhou et al., 2014) data set.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	0.042 ± 0.007	**0.047 ± 0.008**	0.048 ± 0.006	0.051 ± 0.007
0.10	0.085 ± 0.009	0.096 ± 0.011	0.097 ± 0.010	**0.100 ± 0.010**
0.15	0.125 ± 0.010	**0.140 ± 0.011**	0.147 ± 0.011	0.149 ± 0.011
0.20	0.164 ± 0.012	**0.187 ± 0.011**	0.198 ± 0.011	0.201 ± 0.010
0.25	0.211 ± 0.013	0.230 ± 0.017	0.246 ± 0.009	**0.248 ± 0.008**
0.30	0.252 ± 0.012	0.272 ± 0.015	**0.295 ± 0.013**	0.294 ± 0.013
0.35	0.292 ± 0.012	0.316 ± 0.017	0.343 ± 0.015	**0.346 ± 0.016**
0.40	0.332 ± 0.015	0.350 ± 0.018	0.396 ± 0.016	**0.398 ± 0.015**
0.45	0.372 ± 0.014	0.399 ± 0.019	0.447 ± 0.016	**0.450 ± 0.016**
0.50	0.411 ± 0.016	0.444 ± 0.020	**0.500 ± 0.016**	0.501 ± 0.016
0.55	0.447 ± 0.019	0.488 ± 0.016	0.551 ± 0.016	**0.552 ± 0.016**
0.60	0.483 ± 0.019	0.523 ± 0.017	0.603 ± 0.016	**0.604 ± 0.017**
0.65	0.515 ± 0.020	0.564 ± 0.017	0.650 ± 0.017	**0.651 ± 0.018**
0.70	0.544 ± 0.024	0.593 ± 0.023	**0.700 ± 0.016**	0.700 ± 0.016
0.75	0.571 ± 0.027	0.642 ± 0.033	0.750 ± 0.015	**0.751 ± 0.015**
0.80	0.602 ± 0.028	0.704 ± 0.042	0.801 ± 0.013	**0.801 ± 0.013**
0.85	0.631 ± 0.026	0.795 ± 0.026	0.847 ± 0.013	**0.848 ± 0.013**
0.90	0.660 ± 0.027	0.864 ± 0.026	**0.896 ± 0.012**	0.893 ± 0.013
0.95	0.691 ± 0.026	0.934 ± 0.014	**0.946 ± 0.008**	0.945 ± 0.008

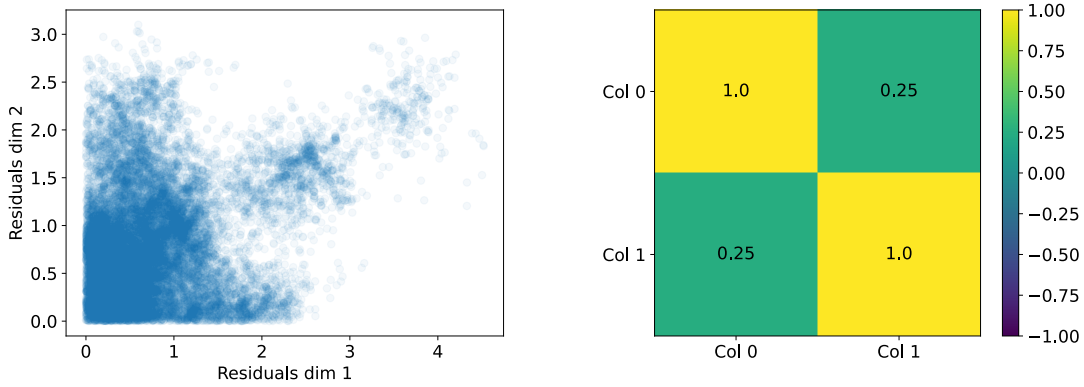
Table 17: Multi-tests method's mean hyperrectangle volumes for the music origin (Zhou et al., 2014) data set.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	24.417 ± 2.147	**22.860 ± 1.602**	23.433 ± 1.969	23.237 ± 2.002
0.10	18.667 ± 1.229	16.928 ± 1.251	16.994 ± 1.215	**16.677 ± 1.155**
0.15	14.569 ± 1.012	**12.774 ± 0.900**	13.190 ± 0.951	13.074 ± 0.909
0.20	12.321 ± 0.962	**10.161 ± 0.586**	10.635 ± 0.868	10.499 ± 0.849
0.25	9.866 ± 0.731	8.460 ± 0.575	8.271 ± 0.642	**8.217 ± 0.646**
0.30	8.105 ± 0.517	7.317 ± 0.471	**6.954 ± 0.482**	6.977 ± 0.482
0.35	7.019 ± 0.448	6.456 ± 0.441	5.992 ± 0.481	**5.951 ± 0.487**
0.40	6.195 ± 0.418	5.692 ± 0.348	5.025 ± 0.383	**4.995 ± 0.386**
0.45	5.476 ± 0.401	4.900 ± 0.338	4.323 ± 0.366	**4.309 ± 0.363**
0.50	4.863 ± 0.394	4.183 ± 0.282	**3.642 ± 0.355**	3.644 ± 0.356
0.55	4.409 ± 0.386	3.610 ± 0.227	3.067 ± 0.345	**3.061 ± 0.346**
0.60	4.011 ± 0.368	3.168 ± 0.166	2.553 ± 0.330	**2.544 ± 0.333**
0.65	3.648 ± 0.346	2.799 ± 0.135	2.086 ± 0.286	**2.085 ± 0.284**
0.70	3.336 ± 0.280	2.463 ± 0.131	**1.728 ± 0.250**	1.730 ± 0.248
0.75	3.002 ± 0.231	2.157 ± 0.149	1.449 ± 0.214	**1.446 ± 0.211**
0.80	2.631 ± 0.198	1.849 ± 0.150	1.223 ± 0.209	**1.221 ± 0.206**
0.85	2.296 ± 0.192	1.483 ± 0.134	1.008 ± 0.186	**1.008 ± 0.188**
0.90	1.987 ± 0.167	1.083 ± 0.145	**0.757 ± 0.169**	0.779 ± 0.171
0.95	1.763 ± 0.160	0.639 ± 0.104	**0.447 ± 0.131**	0.462 ± 0.133



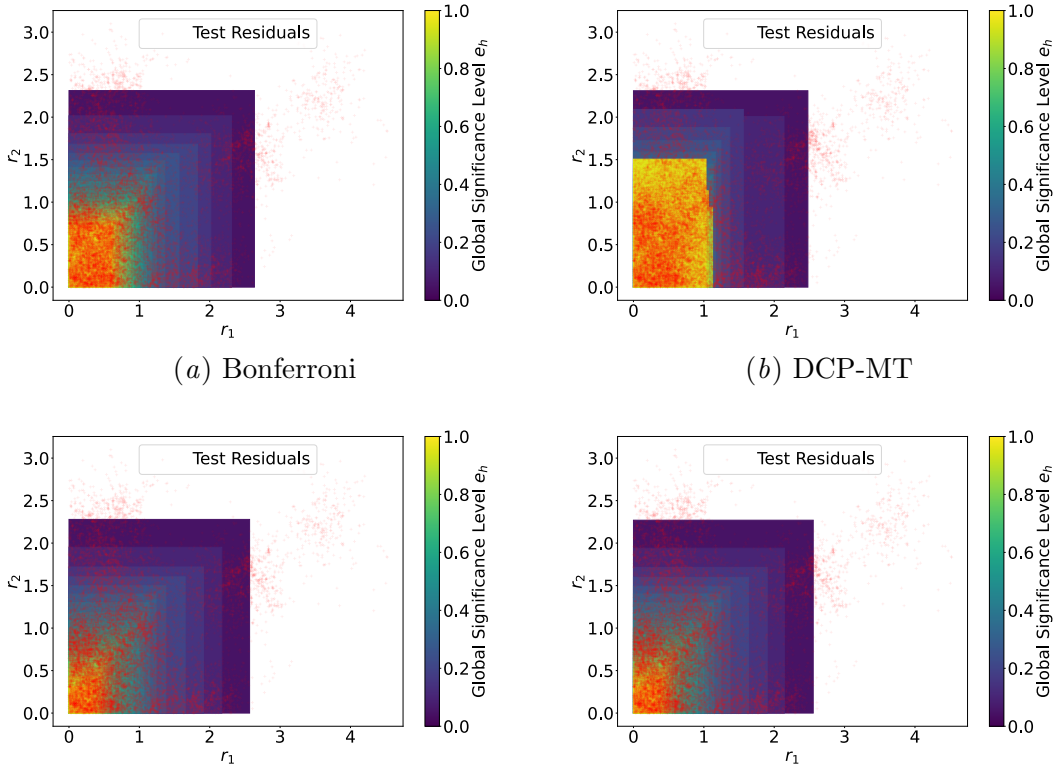
(a) Error Rate (b) Volume

Figure 19: Multi-tests method's mean error rates and volumes for the music origin (Zhou et al., 2014) data set.



(a) Residuals (b) Residual correlation matrix

Figure 20: Underlying model's absolute residuals and their correlation for the music origin (Zhou et al., 2014) data set.



(*a*) Bonferroni (*b*) DCP-MT

(*c*) Empirical Copula (*d*) Max-Rank

Figure 21: Multi-tests method's partition of the residual space for the music origin (Zhou et al., 2014) data set.

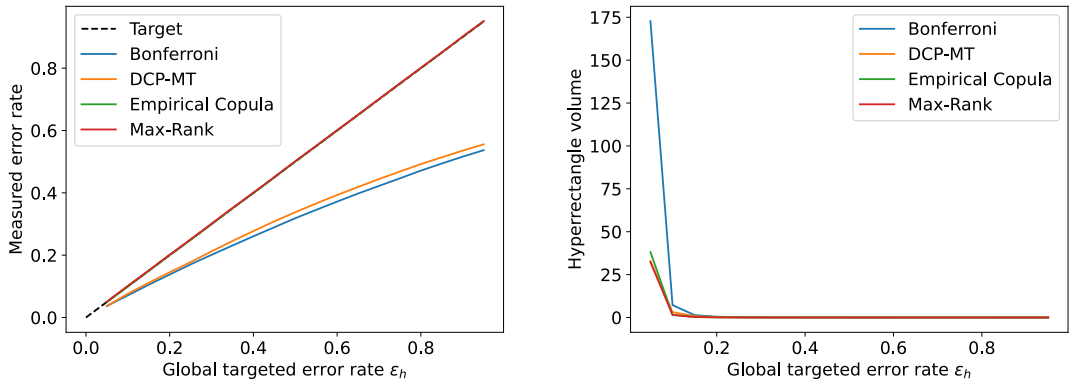
A.7. rf1

Table 18: Multi-tests method's mean error rates and standard deviations for the rf1 (Tsoumakas et al., 2011) data set.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	0.037 ± 0.002	0.037 ± 0.002	0.049 ± 0.003	**0.051 ± 0.002**
0.10	0.071 ± 0.004	0.076 ± 0.005	0.099 ± 0.003	**0.102 ± 0.004**
0.15	0.105 ± 0.004	0.112 ± 0.005	0.149 ± 0.004	**0.151 ± 0.004**
0.20	0.138 ± 0.005	0.145 ± 0.004	0.200 ± 0.005	**0.202 ± 0.005**
0.25	0.171 ± 0.005	0.178 ± 0.005	0.249 ± 0.005	**0.251 ± 0.005**
0.30	0.201 ± 0.006	0.212 ± 0.007	0.299 ± 0.006	**0.301 ± 0.006**
0.35	0.231 ± 0.006	0.245 ± 0.006	0.350 ± 0.006	**0.352 ± 0.006**
0.40	0.261 ± 0.006	0.278 ± 0.007	0.399 ± 0.006	**0.401 ± 0.006**
0.45	0.290 ± 0.006	0.308 ± 0.007	**0.449 ± 0.005**	0.451 ± 0.005
0.50	0.319 ± 0.007	0.338 ± 0.007	**0.500 ± 0.005**	0.502 ± 0.005
0.55	0.345 ± 0.007	0.366 ± 0.008	**0.550 ± 0.004**	0.551 ± 0.004
0.60	0.372 ± 0.008	0.393 ± 0.008	**0.600 ± 0.003**	0.601 ± 0.003
0.65	0.398 ± 0.009	0.419 ± 0.009	**0.650 ± 0.005**	0.651 ± 0.005
0.70	0.422 ± 0.009	0.444 ± 0.009	**0.700 ± 0.005**	0.701 ± 0.005
0.75	0.447 ± 0.009	0.469 ± 0.009	**0.751 ± 0.005**	0.751 ± 0.005
0.80	0.471 ± 0.009	0.492 ± 0.009	**0.800 ± 0.004**	0.801 ± 0.004
0.85	0.494 ± 0.009	0.514 ± 0.010	**0.850 ± 0.004**	0.851 ± 0.004
0.90	0.516 ± 0.009	0.535 ± 0.010	**0.901 ± 0.003**	0.901 ± 0.003
0.95	0.537 ± 0.008	0.556 ± 0.009	**0.951 ± 0.003**	0.951 ± 0.003

Table 19: Multi-tests method's mean hyperrectangle volumes for the rf1 (Tsoumakas et al., 2011) data set.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	172.734 ± 81.099	32.901 ± 14.016	38.125 ± 18.702	**32.341 ± 15.305**
0.10	7.227 ± 3.014	3.135 ± 1.077	1.595 ± 0.617	**1.446 ± 0.548**
0.15	1.378 ± 0.529	0.750 ± 0.265	0.296 ± 0.105	**0.282 ± 0.099**
0.20	0.441 ± 0.158	0.264 ± 0.098	0.090 ± 0.027	**0.086 ± 0.026**
0.25	0.187 ± 0.067	0.120 ± 0.044	0.036 ± 0.011	**0.035 ± 0.011**
0.30	0.094 ± 0.032	0.062 ± 0.023	0.016 ± 0.005	**0.015 ± 0.004**
0.35	0.053 ± 0.018	0.035 ± 0.013	0.008 ± 0.002	**0.007 ± 0.002**
0.40	0.032 ± 0.011	0.021 ± 0.007	0.004 ± 0.001	**0.004 ± 0.001**
0.45	0.020 ± 0.007	0.013 ± 0.005	**0.002 ± 0.001**	0.002 ± 0.001
0.50	0.013 ± 0.004	0.009 ± 0.003	**0.001 ± 0.000**	0.001 ± 0.000
0.55	0.009 ± 0.003	0.006 ± 0.002	**0.001 ± 0.000**	0.001 ± 0.000
0.60	0.006 ± 0.002	0.004 ± 0.001	**0.000 ± 0.000**	0.000 ± 0.000
0.65	0.004 ± 0.002	0.003 ± 0.001	**0.000 ± 0.000**	0.000 ± 0.000
0.70	0.003 ± 0.001	0.002 ± 0.001	**0.000 ± 0.000**	0.000 ± 0.000
0.75	0.002 ± 0.001	0.002 ± 0.001	**0.000 ± 0.000**	0.000 ± 0.000
0.80	0.002 ± 0.001	0.001 ± 0.000	**0.000 ± 0.000**	0.000 ± 0.000
0.85	0.001 ± 0.000	0.001 ± 0.000	**0.000 ± 0.000**	0.000 ± 0.000
0.90	0.001 ± 0.000	0.001 ± 0.000	**0.000 ± 0.000**	0.000 ± 0.000
0.95	0.001 ± 0.000	0.001 ± 0.000	**0.000 ± 0.000**	0.000 ± 0.000



(*a*) Error Rate (*b*) Volume

Figure 22: Multi-tests method's mean error rates and volumes for the rf1 (Tsoumakas et al., 2011) data set.

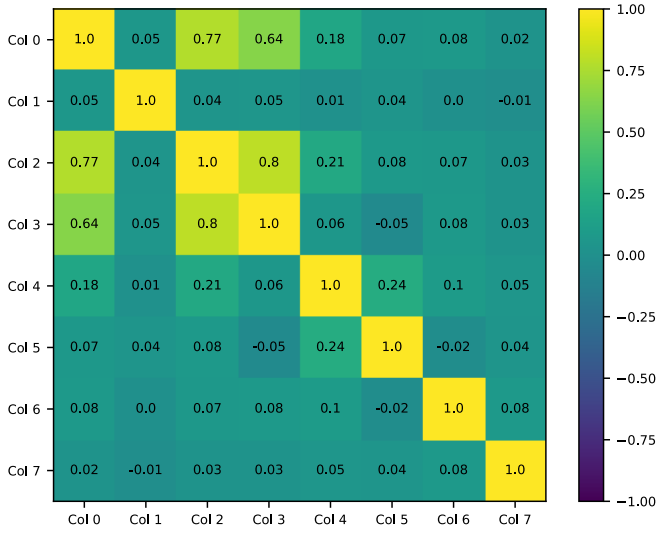


Figure 23: Underlying model's residual correlation matrix for the rf1 (Tsoumakas et al., 2011) data set.

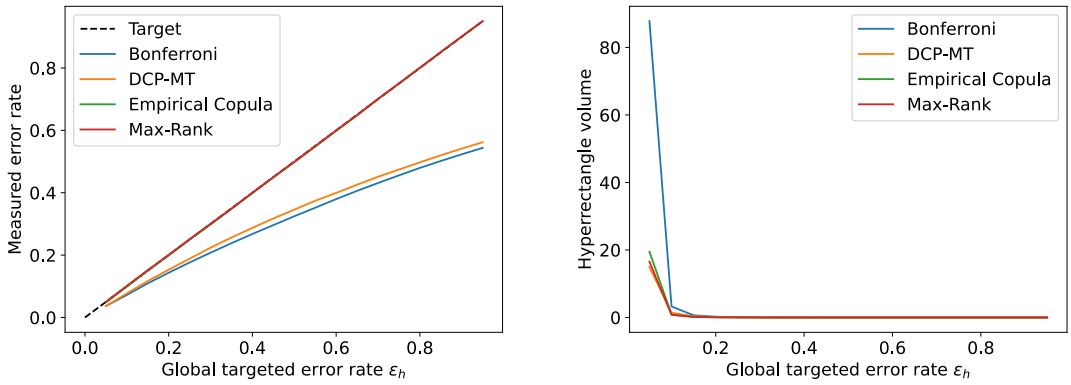
A.8. rf2

Table 20: Multi-tests method's mean error rates and standard deviations for the rf2 (Tsoumakas et al., 2011) data set.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	0.037 ± 0.002	**0.036 ± 0.002**	0.048 ± 0.002	0.050 ± 0.002
0.10	0.072 ± 0.003	0.077 ± 0.004	0.099 ± 0.002	**0.101 ± 0.002**
0.15	0.110 ± 0.004	0.117 ± 0.005	0.150 ± 0.003	**0.152 ± 0.003**
0.20	0.143 ± 0.004	0.153 ± 0.006	0.199 ± 0.004	**0.201 ± 0.004**
0.25	0.176 ± 0.005	0.189 ± 0.006	0.250 ± 0.004	**0.251 ± 0.004**
0.30	0.208 ± 0.005	0.224 ± 0.006	0.298 ± 0.004	**0.300 ± 0.004**
0.35	0.238 ± 0.006	0.256 ± 0.005	0.348 ± 0.003	**0.349 ± 0.003**
0.40	0.268 ± 0.005	0.287 ± 0.006	**0.399 ± 0.004**	0.400 ± 0.004
0.45	0.296 ± 0.005	0.318 ± 0.006	**0.449 ± 0.004**	0.450 ± 0.004
0.50	0.325 ± 0.006	0.346 ± 0.006	**0.498 ± 0.005**	0.499 ± 0.005
0.55	0.352 ± 0.006	0.374 ± 0.006	**0.548 ± 0.005**	0.549 ± 0.005
0.60	0.380 ± 0.006	0.400 ± 0.006	**0.598 ± 0.005**	0.600 ± 0.005
0.65	0.406 ± 0.007	0.426 ± 0.007	**0.648 ± 0.005**	0.649 ± 0.005
0.70	0.431 ± 0.007	0.451 ± 0.008	**0.700 ± 0.005**	0.701 ± 0.005
0.75	0.456 ± 0.007	0.474 ± 0.008	**0.750 ± 0.005**	0.750 ± 0.005
0.80	0.480 ± 0.008	0.497 ± 0.009	**0.800 ± 0.004**	0.801 ± 0.004
0.85	0.502 ± 0.009	0.520 ± 0.009	**0.851 ± 0.004**	0.851 ± 0.004
0.90	0.523 ± 0.008	0.541 ± 0.009	**0.900 ± 0.003**	0.901 ± 0.003
0.95	0.544 ± 0.008	0.562 ± 0.009	**0.950 ± 0.002**	0.950 ± 0.002

Table 21: Multi-tests method's mean hyperrectangle volumes for the rf2 (Tsoumakas et al., 2011) data set.

ϵ	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	87.792 ± 46.649	**14.946 ± 8.429**	19.484 ± 12.388	16.539 ± 10.866
0.10	3.260 ± 2.041	1.404 ± 0.725	0.859 ± 0.398	**0.793 ± 0.369**
0.15	0.651 ± 0.343	0.361 ± 0.172	0.178 ± 0.073	**0.171 ± 0.069**
0.20	0.229 ± 0.108	0.137 ± 0.065	0.059 ± 0.024	**0.057 ± 0.023**
0.25	0.103 ± 0.049	0.065 ± 0.030	0.023 ± 0.009	**0.023 ± 0.008**
0.30	0.054 ± 0.027	0.035 ± 0.016	0.011 ± 0.004	**0.010 ± 0.004**
0.35	0.031 ± 0.015	0.020 ± 0.009	0.005 ± 0.002	**0.005 ± 0.002**
0.40	0.019 ± 0.009	0.013 ± 0.005	**0.003 ± 0.001**	0.003 ± 0.001
0.45	0.012 ± 0.005	0.008 ± 0.004	**0.002 ± 0.000**	0.002 ± 0.000
0.50	0.008 ± 0.004	0.005 ± 0.002	**0.001 ± 0.000**	0.001 ± 0.000
0.55	0.005 ± 0.002	0.004 ± 0.002	**0.001 ± 0.000**	0.001 ± 0.000
0.60	0.004 ± 0.002	0.003 ± 0.001	**0.000 ± 0.000**	0.000 ± 0.000
0.65	0.003 ± 0.001	0.002 ± 0.001	**0.000 ± 0.000**	0.000 ± 0.000
0.70	0.002 ± 0.001	0.001 ± 0.001	**0.000 ± 0.000**	0.000 ± 0.000
0.75	0.002 ± 0.001	0.001 ± 0.000	**0.000 ± 0.000**	0.000 ± 0.000
0.80	0.001 ± 0.000	0.001 ± 0.000	**0.000 ± 0.000**	0.000 ± 0.000
0.85	0.001 ± 0.000	0.001 ± 0.000	**0.000 ± 0.000**	0.000 ± 0.000
0.90	0.001 ± 0.000	0.001 ± 0.000	**0.000 ± 0.000**	0.000 ± 0.000
0.95	0.001 ± 0.000	0.000 ± 0.000	**0.000 ± 0.000**	0.000 ± 0.000



(*a*) Error Rate (*b*) Volume

Figure 24: Multi-tests method's mean error rates and volumes for the rf2 (Tsoumakas et al., 2011) data set.

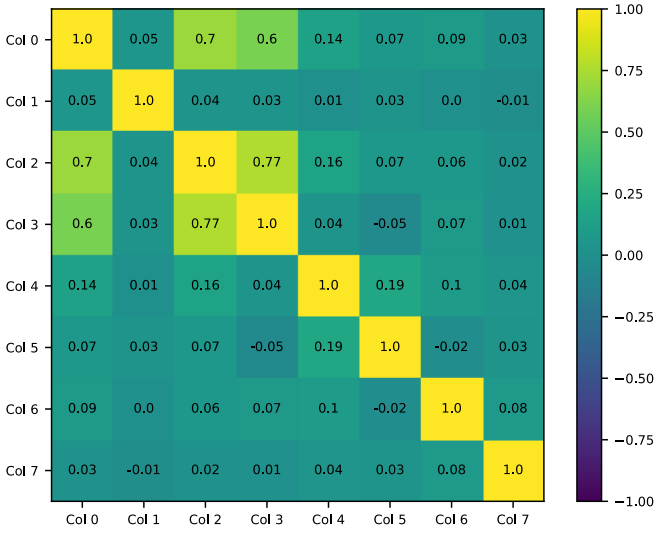


Figure 25: Underlying model's residual correlation matrix for the rf2 (Tsoumakas et al., 2011) data set.

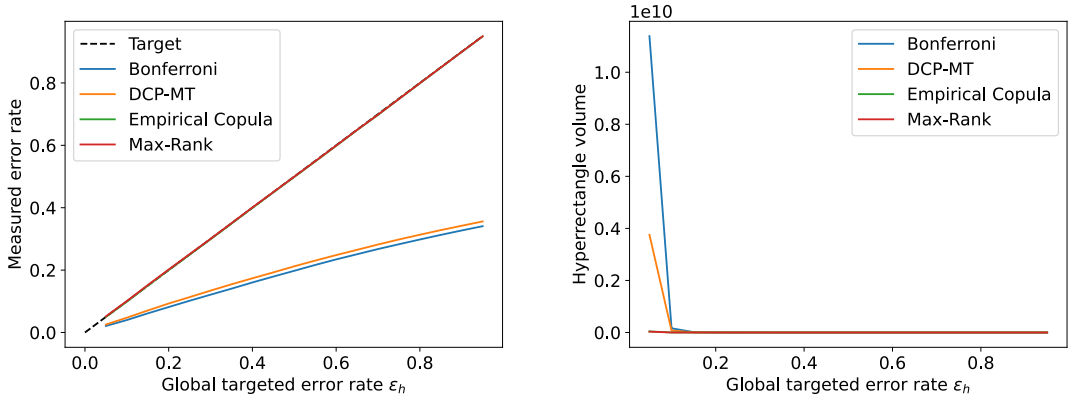
A.9. scm1d

Table 22: Multi-tests method's mean error rates and standard deviations for the scm1d (Tsoumakas et al., 2011) data set.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	0.021 ± 0.001	0.026 ± 0.001	0.050 ± 0.002	**0.052 ± 0.002**
0.10	0.040 ± 0.002	0.047 ± 0.003	0.099 ± 0.003	**0.101 ± 0.002**
0.15	0.061 ± 0.003	0.070 ± 0.004	0.150 ± 0.003	**0.152 ± 0.003**
0.20	0.082 ± 0.003	0.093 ± 0.003	0.200 ± 0.004	**0.202 ± 0.004**
0.25	0.102 ± 0.004	0.113 ± 0.004	0.249 ± 0.003	**0.251 ± 0.003**
0.30	0.121 ± 0.003	0.134 ± 0.004	0.299 ± 0.003	**0.300 ± 0.003**
0.35	0.141 ± 0.004	0.154 ± 0.004	0.349 ± 0.004	**0.350 ± 0.004**
0.40	0.160 ± 0.004	0.174 ± 0.005	0.400 ± 0.004	**0.401 ± 0.004**
0.45	0.179 ± 0.004	0.193 ± 0.004	0.449 ± 0.004	**0.450 ± 0.004**
0.50	0.198 ± 0.004	0.212 ± 0.004	0.498 ± 0.004	**0.499 ± 0.004**
0.55	0.216 ± 0.004	0.231 ± 0.004	0.548 ± 0.004	**0.549 ± 0.004**
0.60	0.234 ± 0.004	0.248 ± 0.003	0.598 ± 0.003	**0.599 ± 0.003**
0.65	0.251 ± 0.004	0.265 ± 0.004	0.649 ± 0.003	**0.650 ± 0.003**
0.70	0.268 ± 0.003	0.282 ± 0.004	0.698 ± 0.004	**0.698 ± 0.004**
0.75	0.283 ± 0.004	0.298 ± 0.004	0.748 ± 0.004	**0.749 ± 0.004**
0.80	0.298 ± 0.003	0.313 ± 0.004	**0.799 ± 0.004**	0.799 ± 0.004
0.85	0.313 ± 0.003	0.328 ± 0.004	**0.849 ± 0.003**	0.850 ± 0.003
0.90	0.327 ± 0.003	0.342 ± 0.004	**0.899 ± 0.003**	0.900 ± 0.003
0.95	0.341 ± 0.003	0.356 ± 0.004	**0.949 ± 0.003**	0.950 ± 0.003

Table 23: Multi-tests method's mean hyperrectangle volumes for the scm1d (Tsoumakas et al., 2011) data set.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	11387195021 ± 5455494297	3753203987 ± 1420280347	41820923 ± 11540509	**32774816 ± 8828963**
0.10	161280296 ± 54689221	74214589 ± 23661934	643216 ± 117978	**557265 ± 109330**
0.15	13872041 ± 4321745	7001601 ± 1918128	38361 ± 5326	**35158 ± 4657**
0.20	2268132 ± 548707	1253055 ± 312528	5553 ± 838	**5195 ± 786**
0.25	544677 ± 126477	321505 ± 73152	1084 ± 144	**1029 ± 138**
0.30	167680 ± 36681	103735 ± 22282	251 ± 31	**243 ± 31**
0.35	61579 ± 12268	39680 ± 8105	64.915 ± 9.348	**62.322 ± 8.743**
0.40	25642 ± 4441	17172 ± 3199	17.357 ± 2.438	**16.811 ± 2.355**
0.45	12098 ± 2042	8198 ± 1427	4.803 ± 0.642	**4.664 ± 0.619**
0.50	6127 ± 1008	4216 ± 686	1.360 ± 0.194	**1.327 ± 0.187**
0.55	3238 ± 530	2311 ± 360	0.380 ± 0.052	**0.371 ± 0.050**
0.60	1777 ± 283	1327 ± 204	0.108 ± 0.015	**0.105 ± 0.015**
0.65	1052 ± 162	794 ± 119	0.029 ± 0.004	**0.029 ± 0.004**
0.70	650.046 ± 98.984	489.282 ± 71.034	0.008 ± 0.001	**0.008 ± 0.001**
0.75	410.726 ± 59.717	311.254 ± 44.093	0.002 ± 0.000	**0.002 ± 0.000**
0.80	265.403 ± 37.348	202.153 ± 28.171	**0.000 ± 0.000**	0.000 ± 0.000
0.85	174.552 ± 23.357	134.516 ± 18.488	**0.000 ± 0.000**	0.000 ± 0.000
0.90	118.474 ± 15.136	90.933 ± 12.394	**0.000 ± 0.000**	0.000 ± 0.000
0.95	81.466 ± 10.949	62.658 ± 8.493	**0.000 ± 0.000**	0.000 ± 0.000



(a) Error Rate (b) Volume

Figure 26: Multi-tests method's mean error rates and volumes for the scm1d (Tsoumakas et al., 2011) data set.

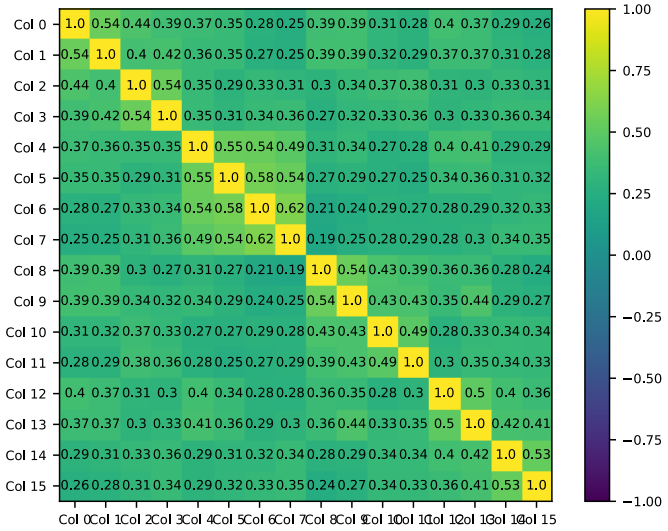


Figure 27: Underlying model's residual correlation matrix for the scm1d (Tsoumakas et al., 2011) data set.

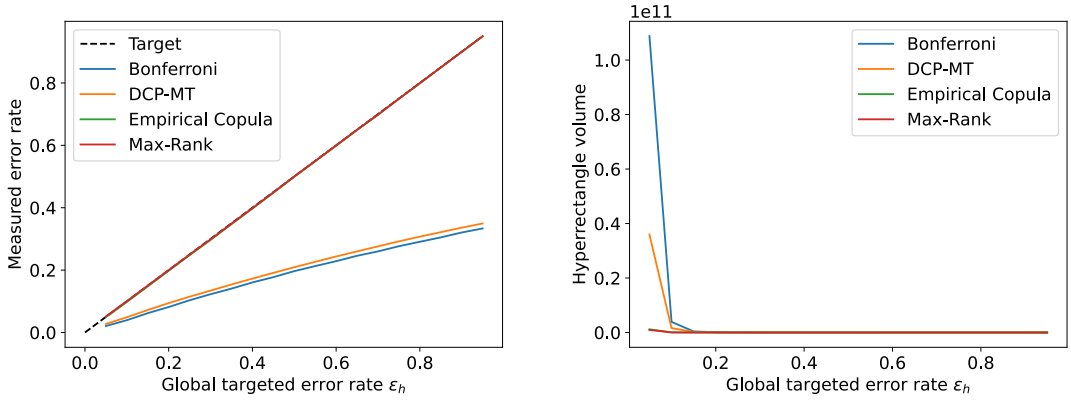
A.10. scm20d

Table 24: Multi-tests method's mean error rates and standard deviations for the scm20d (Tsoumakas et al., 2011) data set.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	0.021 ± 0.002	0.027 ± 0.002	0.049 ± 0.002	**0.051 ± 0.002**
0.10	0.039 ± 0.003	0.049 ± 0.003	0.098 ± 0.004	**0.100 ± 0.004**
0.15	0.062 ± 0.004	0.072 ± 0.004	0.148 ± 0.004	**0.150 ± 0.005**
0.20	0.082 ± 0.004	0.094 ± 0.005	0.198 ± 0.004	**0.200 ± 0.004**
0.25	0.103 ± 0.005	0.115 ± 0.006	0.248 ± 0.004	**0.250 ± 0.004**
0.30	0.123 ± 0.006	0.134 ± 0.005	0.297 ± 0.004	**0.298 ± 0.004**
0.35	0.141 ± 0.005	0.154 ± 0.006	0.347 ± 0.004	**0.349 ± 0.004**
0.40	0.161 ± 0.005	0.173 ± 0.006	0.397 ± 0.005	**0.399 ± 0.005**
0.45	0.177 ± 0.005	0.191 ± 0.006	0.448 ± 0.004	**0.449 ± 0.005**
0.50	0.197 ± 0.005	0.209 ± 0.005	0.499 ± 0.005	**0.500 ± 0.005**
0.55	0.213 ± 0.005	0.227 ± 0.005	0.549 ± 0.005	**0.550 ± 0.005**
0.60	0.229 ± 0.006	0.244 ± 0.005	0.599 ± 0.005	**0.600 ± 0.005**
0.65	0.246 ± 0.006	0.260 ± 0.005	0.649 ± 0.006	**0.650 ± 0.006**
0.70	0.260 ± 0.005	0.276 ± 0.006	0.698 ± 0.005	**0.699 ± 0.005**
0.75	0.277 ± 0.005	0.292 ± 0.006	0.749 ± 0.004	**0.750 ± 0.004**
0.80	0.291 ± 0.006	0.307 ± 0.005	0.799 ± 0.003	**0.799 ± 0.003**
0.85	0.305 ± 0.006	0.322 ± 0.005	0.849 ± 0.003	**0.850 ± 0.003**
0.90	0.321 ± 0.006	0.336 ± 0.005	0.899 ± 0.004	**0.900 ± 0.004**
0.95	0.334 ± 0.006	0.349 ± 0.006	**0.950 ± 0.003**	0.950 ± 0.003

Table 25: Multi-tests method's mean hyperrectangle volumes for the music scm20d (Tsoumakas et al., 2011) data set.

ϵ_h	Bonferroni	DCP-MT	Empirical Copula	Max-Rank
0.05	108772471099 ± 36307263659	35955379066 ± 11015273007	1172582220 ± 345982686	**942799215 ± 301176043**
0.10	3891399957 ± 1418997924	1592281229 ± 546138137	40943128 ± 14226060	**36005596 ± 12547808**
0.15	399793160 ± 145477607	238981538 ± 83581288	4674176 ± 1819219	**4284514 ± 1678530**
0.20	103537820 ± 36730140	63341800 ± 23666577	829646 ± 337978	**784011 ± 313692**
0.25	31089891 ± 11904414	21814753 ± 8697915	195841 ± 80515	**187253 ± 76667**
0.30	13073237 ± 5395580	9026934 ± 3807150	55315 ± 24220	**53257 ± 23192**
0.35	6206213 ± 2708349	4190937 ± 1817270	17306 ± 7919	**16618 ± 7623**
0.40	2943052 ± 1311661	2131334 ± 930811	5734 ± 2745	**5587 ± 2678**
0.45	1625242 ± 719661	1158625 ± 504736	1884 ± 911	**1838 ± 893**
0.50	889071 ± 385238	668919 ± 293430	630 ± 326	**616 ± 318**
0.55	549855 ± 240086	405368 ± 179573	213 ± 113	**208 ± 110**
0.60	349811 ± 157026	254117 ± 113466	70.085 ± 37.083	**68.639 ± 36.429**
0.65	215962 ± 94763	164406 ± 73801	22.013 ± 11.712	**21.671 ± 11.500**
0.70	144915 ± 63992	109267 ± 49440	6.669 ± 3.705	**6.531 ± 3.626**
0.75	94756 ± 42453	74518 ± 34104	1.778 ± 1.068	**1.756 ± 1.047**
0.80	66412 ± 29466	51747 ± 23860	0.419 ± 0.265	**0.412 ± 0.261**
0.85	47633 ± 21547	36683 ± 17061	0.077 ± 0.052	**0.075 ± 0.051**
0.90	33008 ± 15138	26339 ± 12375	0.009 ± 0.006	**0.009 ± 0.006**
0.95	24367 ± 11492	19200 ± 9078	**0.000 ± 0.000**	0.000 ± 0.000



(*a*) Error Rate (*b*) Volume

Figure 28: Multi-tests method's mean error rates and volumes for the scm20d (Tsoumakas et al., 2011) data set.

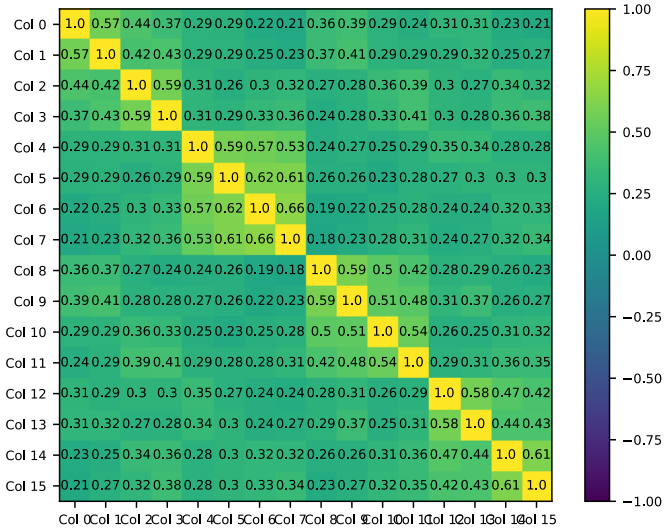


Figure 29: Underlying model's residual correlation matrix for the scm20d (Tsoumakas et al., 2011) data set.

Appendix B. Computational Complexity

This Section presents a preliminary analysis of the computational complexity of the multi-tests methods DCP-MT, Bonferroni predictors (Vovk, 2013), Copula CP (Messoudi et al., 2021) using the empirical copula, and Max-Rank (Timans et al., 2025). These are the methods used for the experiments in this article.

B.1. Theoretical Analysis

We deduce the computational complexity of the compared methods when computing the prediction region for a new object $\mathbf{x}$ and a single global significance level ϵ_h. Further work is necessary to describe computational efficiency gains that can be achieved for each method when prediction regions for multiple global significance levels ϵ_h are computed simultaneously.

Table 26: Multi-tests methods' computational complexity.

	Bonferroni	**DMT-CP**	**Empirical Copula**	**Max-Rank**
Complexity	$\mathcal{O}(MN_{\text{cal}}\log(N_{\text{cal}}))$	$\mathcal{O}(2^{MN_{\text{cal}}})$	$\mathcal{O}(MN_{\text{cal}}\log(N_{\text{cal}}))$	$\mathcal{O}(MN_{\text{cal}}\log(N_{\text{cal}}))$

Bonferroni. Computing the local significance levels ϵ_m can be done in a single step and has a complexity $\mathcal{O}(1)$. Next, the method determines the associated quantile for each dimension. This can be achieved through sorting and using the index $\lceil(1-\epsilon_m)(N_{\text{cal}}+1)\rceil$ to get the right element, costing $\mathcal{O}(N_{\text{cal}}\log(N_{\text{cal}}))$ and $\mathcal{O}(1)$, respectively. Therefore the total computational of Bonferroni CP complexity is $\mathcal{O}(1) + M(\mathcal{O}(N_{\text{cal}}\log(N_{\text{cal}})) + \mathcal{O}(1)) = \mathcal{O}(MN_{\text{cal}}\log(N_{\text{cal}}))$.

DCP-MT. The MILP problem has MN_{cal} binary variables and therefore a worst-case complexity of $\mathcal{O}(2^{MN_{\text{cal}}})$.

Copula CP. Using the empirical copula, Copula CP first computes the p-values for the residuals in each dimension in $\mathcal{O}(MN_{\text{cal}}\log(N_{\text{cal}}))$. Next, binary search is used to find the largest local significance ϵ_m such that the global targeted error rate of ϵ_h is not exceeded, requiring $\mathcal{O}(\log(N_{\text{cal}}))$ steps, which each takes $\mathcal{O}(MN_{\text{cal}})$ operations to evaluate. Finally, because the residuals are already sorted from the computation of the p-values, computing the $1-\epsilon_m$ quantile in each dimension can be done in $\mathcal{O}(M)$ time. Therefore, the total computational complexity of Copula CP is $\mathcal{O}(MN_{\text{cal}}\log(N_{\text{cal}})) + \mathcal{O}(N_{\text{cal}}) \cdot \mathcal{O}(MN_{\text{cal}}) + \mathcal{O}(M) = \mathcal{O}(MN_{\text{cal}}\log(N_{\text{cal}}))$.

Max-Rank. Getting the ranks of the residuals for each dimension costs $\mathcal{O}(MN_{\text{cal}}\log(N_{\text{cal}}))$. Computing the maximum rank for each example in the calibration set adds $\mathcal{O}(MN_{\text{cal}})$ operations. The $1-\epsilon_h$ quantile of the maximum ranks is computed in $\mathcal{O}(N_{\text{cal}}\log(N_{\text{cal}}))$.

Finally, because the residuals are already sorted, computing the $1-\epsilon_m$ quantile in each dimension can be done in $\mathcal{O}(M)$. Therefore, the total computational complexity of Max-Rank is $\mathcal{O}(MN_{\text{cal}}\log(N_{\text{cal}})) + \mathcal{O}(MN_{\text{cal}}) + \mathcal{O}(N_{\text{cal}}\log(N_{\text{cal}})) + \mathcal{O}(M) = \mathcal{O}(MN_{\text{cal}}\log(N_{\text{cal}}))$.

B.2. Experimental Wall Time

Table 27: Total wall time across all experiments for multi-tests methods. The column $M = \dim(\mathbf{Y})$ indicates the number of dimensions in the labels space. Bold values indicate the fastest method.

Data Set	Bonferroni	DMT-CP	Empirical Copula	Max-Rank	M
synt, $e \sim [\mathcal{U}(-1,1), \chi^2_2]^\top$	**0.009**	31.618	1.292	0.013	2
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 0$	**0.010**	31.684	1.295	0.013	2
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 0.5$	**0.010**	32.026	1.295	0.013	2
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = 1$	**0.011**	35.666	1.410	0.016	2
synt, $e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$, $\sigma = -0.5$	**0.010**	30.828	1.272	0.013	2
diabetes	0.337	63.747	2.763	**0.035**	2
music origin	0.063	146.643	8.242	**0.048**	2
rf1	**0.050**	696.357	50.787	0.192	8
rf2	**0.048**	698.977	50.939	0.196	8
scm1d	**0.090**	1472.797	59.414	0.400	16
scm20d	**0.076**	1326.763	50.244	0.359	16

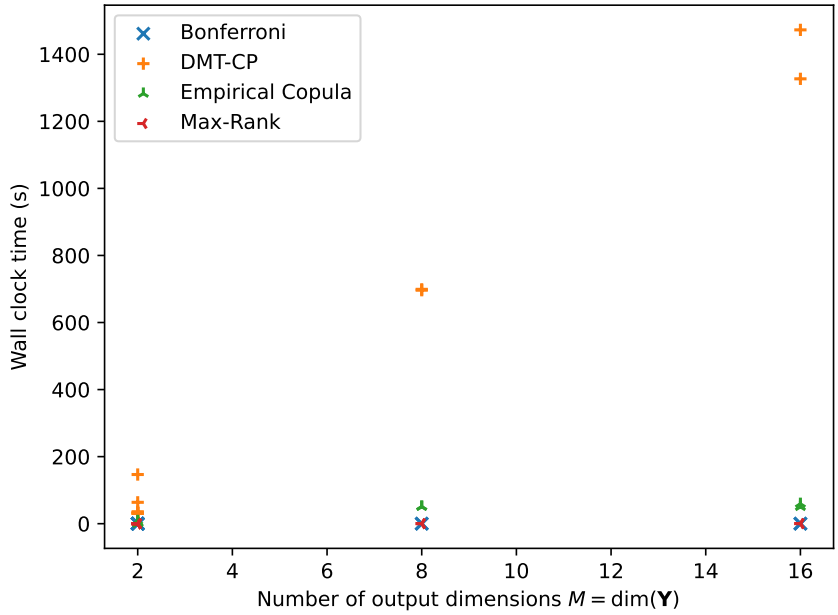


Figure 30: Total wall time in seconds across all experiments for multi-tests methods as a function of the number of dimensions in the labels space $M = \dim(\mathbf{Y})$.

Table 27 and Figure 30 show the wall time for each tested method. They do not include any compute time related to training or the computation of the nonconformity scores and

measure exclusively the time it took the methods to compute the forecasting regions. The conclusions that can be drawn from comparing methods based on these results are limited as these methods are not all optimised to the same degree. Furthermore, we conducted these experiments on a virtual server whose performance depends on the load on the host system which may vary over time.