Interpretable Multivariate Conformal Prediction with Balanced and Jointly Calibrated Rectangular Envelopes

Nabil Alami
Department of Statistics
and Data Science, MBZUAI

Rafael Izbicki
Department of Statistics,
UFSCar

Souhaib Ben Taieb
Department of Statistics
and Data Science, MBZUAI

Abstract

Multi-output conformal prediction methods often produce flexible but hard-to-interpret prediction sets, while alternative approaches yield hyperrectangular regions that can be overly conservative. We propose a framework that bridges these two paradigms by constructing prediction sets as Cartesian products of regions defined on low-dimensional output partitions. To ensure joint coverage, we introduce three calibration strategies and clarify their connection to the notion of asymptotic balance, which seeks to equalize miscoverage across partitions asymptotically. Motivated by interpretability in high-dimensional output settings, we further propose a novel nonconformity score that generates efficient hyperrectangular sets. Experiments on real-world datasets demonstrate that our method improves efficiency over existing baselines while preserving interpretability and achieving asymptotic balance.

1 Introduction

Conformal Prediction (CP) provides a powerful, model-agnostic framework for constructing prediction sets with guaranteed finite-sample marginal coverage (Vovk et al., 2005; Shafer and Vovk, 2008). Substantial research has focused on univariate prediction problems (Vovk et al., 2005; Lei et al., 2018; Angelopoulos et al., 2023). For multi-output CP problems, most existing methods implicitly or explicitly rely on estimating the conditional density of the output (Izbicki et al., 2020; Braun et al., 2025b; Dheur et al., 2025). These methods can be statistically efficient, yielding prediction sets of small volume while maintaining finite-sample coverage, but often incur substantial computational costs and produce non-analytic, difficult-to-interpret sets. In contrast, other approaches construct separate intervals for each output dimension, resulting in hyperrectangular prediction sets (Neeven and Smirnov, 2018; Messoudi et al., 2021; Zhou et al., 2024; Timans et al., 2025; Sampson and Chan, 2024). Such methods offer strong interpretability, as the interval for each output can be visualized, making them appealing in practical applications. However, by ignoring dependencies between outputs, they may suffer from statistical inefficiency. We provide a review of related work on multi-output CP in Appendix C.

In this work, we seek a middle ground. We develop efficient multi-output CP methods that remain interpretable by constructing prediction sets as Cartesian products of lower-dimensional regions. This design captures part of the dependence structure between outputs while preserving much of the simplicity of hyperrectangular methods. Our contributions can be summarized as follows:

(i) We propose a general framework for constructing CP prediction sets as Cartesian products of low-dimensional sets with finite-sample joint coverage guarantee, based on three different calibration strategies; (ii) We introduce a novel nonconformity score, *the rectangular envelope*, a density-based score that produces small rectangular prediction sets; (iii) We develop a new method for constructing rectangular regions that can achieve asymptotic balance and demonstrates improved efficiency compared to several existing baselines.

2 Split-Conformal Prediction

We consider a supervised regression problem with i.i.d. samples $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{M} \sim P$, where each input $X_i \in \mathcal{X} \subset \mathbb{R}^K$ is a feature vector and $Y_i \in \mathcal{Y} \subset \mathbb{R}^d$

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

is a multivariate output. We also consider a new test point $(X_{n+1}, Y_{n+1}) \sim P$, where Y_{n+1} is unobserved at prediction time. We adopt the Split Conformal Prediction (SCP) framework (Papadopoulos et al., 2002; Lei et al., 2018). The dataset $\mathcal{D}$ is divided into a training set and a calibration set $\mathcal{S}_{\text{cal}} := \{(X_i, Y_i)\}_{i=1}^n$. A predictive model $h : \mathcal{X} \to \mathcal{Y}$ is trained on the training set. Given a miscoverage level $\alpha \in (0,1)$, the goal is to construct a prediction set $\mathcal{C}_\alpha(X_{n+1})$ for the new input X_{n+1} such that $\mathbb{P}(Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1})) \geq 1-\alpha$.

This procedure requires a nonconformity score (NCS) function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which measures the discrepancy between a candidate output y and the model prediction at x (for example, $s(x,y) = \|y - h(x)\|$). We compute the calibration scores $s_i := s(X_i, Y_i)$ for $i = 1, \ldots, n$, and define $\hat{q}_\alpha := \text{Quantile}(\{s_i\}_{i=1}^n, 1-\alpha)$. The SCP prediction set for X_{n+1} is

$$\mathcal{C}_\alpha(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}_\alpha\}. \tag{1}$$

Theorem 2.1. *Under exchangeability of* $\{(X_i, Y_i)\}_{i=1}^n$ *and* (X_{n+1}, Y_{n+1})*, the SCP set* (1) *satisfies the finite-sample coverage guarantee*

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1})) \;\geq\; 1-\alpha. \tag{2}$$

3 A New Framework for Interpretable Multi-output CP

We propose the Product of Interpretable Conformal Prediction (PICP), a general framework for constructing prediction sets as Cartesian products of low dimensional regions.

PICP Methodology: For an output $Y \in \mathbb{R}^d$, let $[d] := \{1, \ldots, d\}$, and consider $\mathcal{G} := \{G_1, \ldots, G_m\}$ a partition of $[d]$ into $m \geq 1$ groups of nonempty subsets, $\bigcup_{j=1}^m G_j = [d], G_i \cap G_j = \emptyset, \forall i, j \in [d], i \neq j$. We denote $Y^{G_j} := \{y_i, i \in G_j\}$. We also note y_{-i} the $d-1$ dimensional output vector without the i-th output.

Consider $m \leq d$ NCS functions, each defined on a subset of the output coordinates $s^j : \mathcal{X} \times \mathbb{R}^{|G_j|} \to \mathbb{R}_+, \quad j = 1, \ldots, m$. We refer to $s^1, \ldots, s^m$ as *sub-scores*. This leads to prediction regions of the form

$$\mathcal{C}_\alpha := \mathcal{C}_\alpha^1 \times \cdots \times \mathcal{C}_\alpha^m, \tag{3}$$

where $\mathcal{C}_\alpha^j, j = 1, \ldots, m$ are low-dimensional regions. We refer to them as *sub-prediction sets*.

Given a miscoverage level $\alpha \in (0,1)$, we aim to calibrate each sub-prediction set induced by the corresponding sub-score so that their Cartesian product in (3) satisfies (2), using three calibration strategies (CS).

Bonferroni Calibration. We choose $\alpha_1, \ldots, \alpha_m \in (0,1)$, $\sum_{j=1}^m \alpha_j \leq \alpha$, so that the Cartesian product $\mathcal{C}_{\text{Bonf}}(x) := \prod_{j=1}^m \mathcal{C}_{\alpha_j}^j(x)$, achieves joint coverage of at least $1-\alpha$, where the sub-prediction sets are

$$\mathcal{C}_{\alpha_j}^j(x) := \left\{u \in \mathbb{R}^{|G_j|} : s^j(x,u) \leq \widehat{Q}_{\alpha_j}^j\right\}, \; j = 1, \ldots, m,$$

and $\widehat{Q}_{\alpha_j}^j := \text{Quantile}(\{s_i^j\}_{i=1}^n, 1-\alpha_j)$.

Max-aggregation. We aggregate the sub-scores by taking their maximum: $s_i^\infty := \max_j s^j(X_i, Y_i^{G_j})$, $i = 1, \ldots, n$, and compute the corresponding empirical quantile $\widehat{Q}_\alpha := \text{Quantile}(\{s_i^\infty\}_{i=1}^n, 1-\alpha)$. Then,

$$\mathcal{C}_\alpha(x) := \{y \in \mathbb{R}^d : \max_j s^j(x, y^{G_j}) \leq \widehat{Q}_\alpha\}$$

satisfies (2) thanks to Theorem 2.1 and is a Cartesian product of the sub-prediction sets

$$\mathcal{C}_\alpha^j(x) := \left\{u \in \mathbb{R}^{|G_j|} : s^j(x,u) \leq \widehat{Q}_\alpha\right\}, \; j = 1, \ldots, m.$$

Max-Rank (Timans et al. 2025). Let $s_{(1)}^j \leq \cdots \leq s_{(n)}^j$ denote the order statistics of the calibration sub-scores $\{s_i^j\}_{i=1}^n$. The Max-Rank procedure selects the maximum integer rank $r_{\max} \in \{1, \ldots, n\}$ such that the product set $\mathcal{C}_{\text{MR}}(x) = \prod_{j=1}^m \mathcal{C}_{\text{MR}}^j(x)$ attains joint coverage at least $1-\alpha$, where

$$\mathcal{C}_{\text{MR}}^j(x) := \left\{u \in \mathbb{R}^{|G_j|} : s^j(x,u) \leq s_{(r_{\max})}^j\right\}.$$

Empirically, the Max-Rank correction is very efficient (Timans et al., 2025; Schlembach et al., 2025). More details about these strategies are given in Appendix A. The proofs of all results are provided in Appendix B.

When PICP is applied with arbitrary sub-scores, we write **Bonf-PICP**, **Max-PICP**, and **MR-PICP** to denote calibration of the sub-prediction sets using Bonferroni correction, Max-aggregation, and Max-Rank correction, respectively.

Particular cases of PICP. Our PICP framework recovers many multi-output CP methods as special cases. For e.g., with a single group ($m = 1$) and Mahalanobis score it recovers ellipsoidal sets (Johnstone and Cox, 2021; Messoudi et al., 2022), and with density-based scores it matches density/sampling approaches (Dheur et al., 2025). With per-coordinate groups ($m = d$), it also recovers hyperrectangular methods such as Timans et al. (2025) and the M-CP method (Zhou et al., 2024; Dheur et al., 2025). More broadly, PICP allows choosing any non-conformity score with different CS within one framework.

3.1 Asymptotic Balance

While our CSs guarantee joint coverage over the Cartesian product in (3), they do not control how coverage is distributed across individual sub-prediction sets. In a clinical example predicting blood pressure (BP) and cholesterol (Chol), an 80% joint target could yield 95% marginal coverage for BP but only 85% for Chol. Such an imbalance is undesirable, as it reflects an uneven allocation of uncertainty across outputs. We illustrate this concept in Fig. 1.

To quantify this, Sampson and Chan (2024) introduce the notion of **asymptotic balance**, which we restate below, adapted to our notation and framework.

Definition 3.1. *Define the marginal miscoverage of the j-th sub-prediction set as $\alpha_j := \mathbb{P}(Y^{G_j} \notin \mathcal{C}^j)$ for $j = 1, \dots, m$. We say that the Cartesian product $\prod_{j=1}^m \mathcal{C}^j$ achieves* asymptotic balance *if $\max_{1 \le j \le m} |\alpha_j - \bar{\alpha}| \xrightarrow{p} 0$ as $n \to \infty$, where $\bar{\alpha} := \frac{1}{m}\sum_{j=1}^m \alpha_j$.*

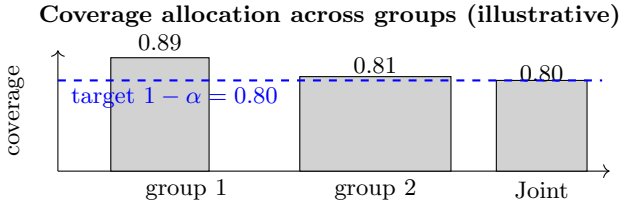

Figure 1: Illustration of groupwise imbalance.

Sampson and Chan (2024) establish asymptotic balance for their hyperrectangular set calibrated via the max-aggregation CS, but under a strong assumption[1]. A limitation of max-aggregation CS is that all sub-prediction sets use a common quantile threshold, which can lead to imbalance across groups. This motivates alternative CS, which naturally promote balance when sub-scores have no ties.

Proposition 3.1. *Let $\mathcal{G} = \{G_1, \dots, G_m\}$ be a partition of $[d]$. Assume that, for each group G_j, the corresponding calibration sub-scores are continuously distributed. Then Bonf-PICP with $\alpha_j = \alpha/m, j = 1, \dots, m$ is asymptotically balanced.*

4 A PICP Instance with Pairwise Rectangular Sets

We introduce a new instance of the PICP framework that combines (*i*) pairwise rectangular prediction sets and (*ii*) improved efficiency through a novel NCS. The pairwise construction is motivated by the fact that prediction regions in dimension $d > 2$ are often difficult to interpret. Specifically, we build the overall prediction set as a Cartesian product of *pairwise* regions by partitioning the outputs into consecutive pairs, i.e., $\mathcal{G}_{\text{pair}} := \{(1,2),(3,4),\dots\}$[2] if d is even. Otherwise, we can consider a union of pairs and a singleton if the dimension is odd.

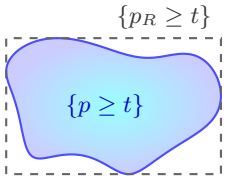

Figure 2: Illustration of a 2D level set for a density p at level t and its rectangle envelope level set.

A new NCS for rectangular envelopes. Flexible multi-output CP methods that account for output dependencies can yield tight prediction sets. In particular, the empirical results of Dheur et al. (2025) show that the DR-CP method, relying on the density-based NCS $s_{\text{DR-CP}}(x, y) = -\log p(y|x)$ (Sadinle et al., 2019), achieves strong efficiency. Motivated by the goal of obtaining similarly efficient yet interpretable hyperrectangular prediction sets, we introduce a new nonconformity score derived from $p(y|x)$, and ask: *which densities admit hyperrectangular superlevel-sets?*

In what follows, hyperrectangle is used in a generalized sense to denote a Cartesian product of one-dimensional sets.

Proposition 4.1. *Consider a continuous pdf $p(y)$.*

Assume that its superlevel-sets $L_t := \{y : p(y) \ge t\}$ are hyperrectangles $\forall t$. Then there exist functions $p_1, \dots, p_d : \mathbb{R} \to \mathbb{R}_+$ such that $p(y) = \min\{p_1(y_1), \dots, p_d(y_d)\}\, \forall y \in \mathbb{R}^d$.

Following Proposition 4.1, we define what we call the *rectangular envelope* (RE) of $p(y)$ as

$$p^R(y) := \min\{p_1^R(y_1), \dots p_d^R(y_d)\},$$

where $p_i^R(y_i) := \sup_{y_{-i} \in \mathbb{R}^{d-1}} p(y)$. The corresponding envelope superlevel-set is $R_t^\star := \{y : p^R(y) \ge t\}$.

Proposition 4.2. *Assume p is continuous and L_τ are compact $\forall \tau > 0$. Then, for any $t > 0$, $R_t^\star$ is the hyperrectangle with smallest Lebesgue volume containing the superlevel-set L_t of the original density.*

An illustration of Proposition 4.2 is shown in Fig. 2.

[1] They assume there exist constants $c_j \ge 0$ and functions $\mu_j(X)$ such that the conditional distributions of $c_j\{Y_j - \mu_j(X)\} \mid X$ are identical for all $j = 1, \dots, m$ almost surely.

[2] Although other pairwise partitions are possible, we restrict to this consecutive pairing in this work for simplicity.

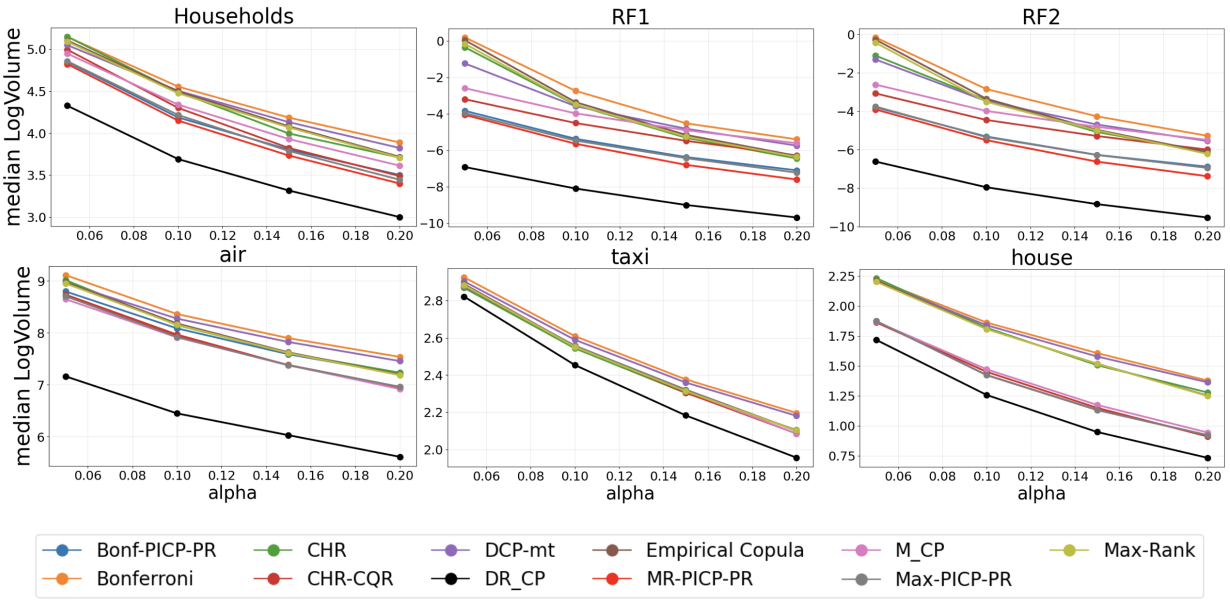

Figure 3: Median logvolume vs $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$ for many methods across 10 seeds.

Our PICP-PR methods. For each pair G_j, assume we have an estimate of the conditional density $\hat{p}(y^{G_j} \mid x)$. We define the corresponding RE nonconformity score as

$$s^j_{\mathrm{RE}}(x, y) := -\log \hat{p}^R(y^{G_j} \mid x).$$

Using these pairwise sub-scores, the resulting prediction set is a Cartesian product of two-dimensional rectangles. We then calibrate these sub-scores using one of the CS introduced in Section 3, which yields finite-sample joint coverage guarantee (2). In this work, we consider two such methods: **Bonf-PICP-PR** and **MR-PICP-PR**. Moreover, by Proposition 3.1, **Bonf-PICP-PR** is asymptotically balanced.

5 Experimental Study

We compare our methods with various baselines on six datasets commonly used in multivariate regression (Dheur et al., 2025); see Appendix D for details. To ensure a fair comparison, all methods share the same underlying model, a conditional multivariate Gaussian predictor $\hat{p}(y|x) = \mathcal{N}(y; f_\theta(x), \Sigma_\phi(x))$, following Braun et al. (2025b). We adopt this model as it gives an explicit form of the estimated density $\hat{p}(y|x)$, and crucially, yields a closed-form expression for our pairwise RE; See Appendix B.

Methods. We consider DR-CP as a strong efficiency-oriented baseline, but the weakest in terms of interpretability. We also evaluate variants of PICP methods, namely **Max-PICP-PR** which uses max-aggregation CS. For hyperrectangular CP baselines, we include **Bonferroni** and **Max-Rank** (Timans et al., 2025), both based on absolute residuals scores. Additional methods considered are **Empirical Copula** (Messoudi et al., 2021), **DCP-mt** (Schlembach et al., 2025), **CHR** and **CHR-CQR** (Sampson and Chan, 2024), and **M-CP** (Dheur et al., 2025).

Figure 3 reports the median prediction-set volume for all methods. As expected, DR-CP consistently produces the most efficient sets, but at the cost of interpretability. Among hyperrectangular methods, our approaches generally achieve the smallest median volume, with **MR-PICP-PR** being the most efficient overall. The performance gap between our methods and the baselines becomes more pronounced on higher-dimensional datasets (RF1 and RF2). Although **M-CP** and **CHR-CQR** also produce relatively small prediction sets, they exhibit weaker balance (see Fig. 4). The remaining baselines tend to be more conservative, resulting in larger prediction sets.

For illustration, we plot in the top panel of Figure 4 example sub-prediction sets for the `Household` dataset of our method alongside selected baselines.[3] Notably, **DCP-mt**, which optimizes miscoverage across sub-prediction set intervals to improve efficiency, produces a large rectangle in the left panel but a much tighter one, similar to **MR-PICP-PR**, in the right panel. This highlights a practical manifestation of imbalance. The bottom panel of Figure 4 confirms that **DCP-mt**, methods based on max-aggregation CS: **M-CP**,

[3]The outputs are $y[0]$: income, $y[1]$: food, $y[2]$: housing, and $y[3]$: utilities, predicted from household covariates.

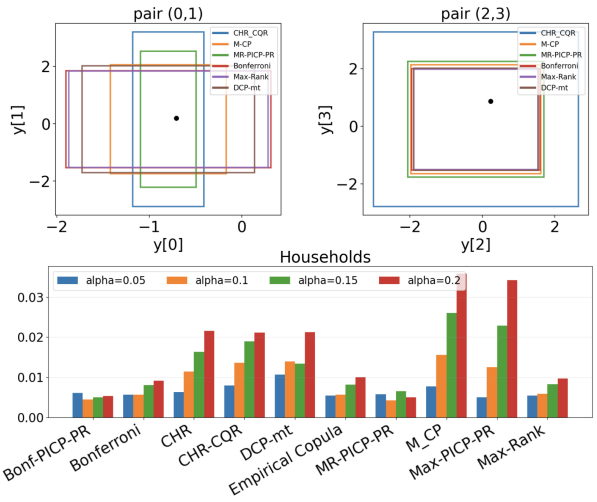

Figure 4: Top: sub-prediction sets for a single test point. Bottom: mean balance per method and α.

Max-PICP-PR, and Sampson and Chan (2024) approaches **CHR**, **CHR-CQR** exhibit the weakest balance.

In contrast, we observe in Fig. 4 that **Bonf-PICP-PR** and **MR-PICP-PR** are the most balanced across the pairs. However, we note that for our methods the balance is relative to the sub-prediction sets of each pair, which is different from hyperrectangles who evaluate balance across all intervals. Additional results are available in Appendix D.

6 Conclusion

We proposed a framework for interpretable multi-output conformal prediction based on Cartesian products of low-dimensional regions, together with calibration strategies that guarantee joint coverage and can achieve asymptotic balance. We also introduced the Rectangular Envelope score to obtain efficient hyperrectangular sets. Empirically, our method provides more informative sets than hyperrectangular baselines, while remaining interpretable, enabling reliable uncertainty quantification in high dimensions.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, but without time complexity
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). No.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). No

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Supplementary Material

A On the Calibration Strategies

We discuss in this section the properties of the three different calibration strategies presented in the PICP methodology.

A.1 Max-aggregation

Working with the Max-aggregation CS is a simple and practical strategy, because it outputs a same threshold for all groups. However, this makes it sensitive to the noisiest sub-score that may have a higher variability than the others, causing overcoverage and poor balance. Many papers highlight this issue and find strategies to normalize the scores like in Fan and Sesia (2025); Sampson and Chan (2024) where they focus on normalizing the residuals so that they live on a same scale. For instance, if we normalize the sub-scores $\frac{s^j-\mu^j}{\sigma_j}$ before taking the max, then each prediction would take the form: $\{s^j \leq \sigma^j \hat{Q}_\alpha + \mu_j\}$ which would make the volume of each sub-prediction set vary across the groupings.

A.2 Bonferroni Calibration

Bonferroni corrections are widely used for multiple testing (Sedgwick, 2012) and control the family-wise error rate (FWER). It is well known that Bonferroni-type corrections can be overly conservative under positive dependence between hypotheses (Vovk et al., 2022; Benjamini, 2010).

Because for each sub-prediction set we select the $1-\alpha_j$ empirical quantile, the set $\mathcal{C}^j_{\text{Bonf}}$ has the $1-\alpha_j$ coverage guarantee by Theorem (2.1). For the joint coverage to hold, it suffices to have $\sum_j \alpha_j \leq \alpha$ via a union bound argument:

$$\mathbb{P}\left(Y_{n+1} \in \prod_{j=1}^{m} \mathcal{C}^j_\alpha(X_{n+1})\right) = 1 - \mathbb{P}\left(\bigcap_{j=1}^{m} (Y_{n+1} \notin \mathcal{C}^j_\alpha(X_{n+1})\right) \geq 1 - \sum_{j=1}^{m} \alpha_j \geq 1 - \alpha.$$

In practice, we take $\alpha_j = \alpha/m$.

A.3 Max-Rank Correction

Max-Rank (Timans et al., 2025) is a carefully designed approach that efficiently controls FWER. The authors of the paper demonstrate that $r_{\max}$ is equal to $r_{(\lceil (1-\alpha)(n+1) \rceil)}$ where $r_i := \max_j \text{Rank}(s^j_i), i = 1, \ldots, n$. They also prove that Max-Rank correction is always better than Bonferroni in case of both independence (ID) and positive dependence (PD).

Denote $r_{\max}$ be the rank selected by the Max-Rank correction on a calibration sample of size n. Then each sub-prediction set is defined as

$$\mathcal{C}^j_\alpha(x) := \{s^j(x,u) \leq s^j_{(r_{\max})}\} = \{\text{Rank}(s^j(x,u)) \leq r_{\max}\}$$

A.4 Balance of Bonf-PICP.

Let n denote the calibration set size and fix $j \in [m]$. Denote the calibration sub-scores by $s^j_1, \ldots, s^j_n$ and the test sub-score by s^j_{n+1}. Define the (no-ties) rank

$$R^j := 1 + \sum_{i=1}^{n} \mathbf{1}\{s^j_i < s^j_{n+1}\} \in \{1, \ldots, n+1\}.$$

By exchangeability and continuity, R^j is uniform on $\{1, \dots, n+1\}$. If a method accepts group j whenever $R^j \leq r$, with r independent of R^j, then

$$\text{Cov}_j = \mathbb{P}(R^j \leq r) = \mathbb{E}\left[\frac{r}{n+1}\right] \qquad \forall j. \tag{*}$$

Bonf-PICP uses $r = \lceil (n+1)(1-\alpha/m) \rceil$.

Therefore Cov_j is the same for all j, hence the balance is 0 (and thus asymptotically 0). This proves proposition 3.1.

For MR-PICP, we do not have a direct proof of balance, since equation (*) does not apply when the rank $r_{\max}$ is data-dependent. Nevertheless, the strong balance observed in our experiments suggests that (*) may hold approximately in practice.

B Proofs Of Results

B.1 PICP coverage

Fix a partition $\mathcal{G} = \{G_1, \dots, G_m\}$. Since the calibration points (X_i, Y_i) are i.i.d., it follows that, for each fixed group G_j, the corresponding calibration sub-scores $\{s^j(X_i, Y_i^{G_j})\}_{i=1}^n$ are i.i.d. Then by construction of the calibration strategies, all MR-PICP, Max-PICP, Bonf-PICP methods achieve the finite-sample coverage guarantee (2).

B.2 Rectangular envelope

Proof of Proposition 4.1 Since all the superlevel-sets of p are hyperrectangles, we may choose a product representation

$$L_t := \prod_{i=1}^{d} I_i(t), \quad \text{and} \quad I_i(s) \subseteq I_i(t) \, \forall s > t, i \in [d]$$

with $I_i(t) \subset \mathbb{R}$.

Lemma B.1. *Let $\mathcal{X}$ be a set and let $f, g : \mathcal{X} \to [0, \infty)$. If*

$$\{x \in \mathcal{X} : \ f(x) \geq t\} = \{x \in \mathcal{X} : \ g(x) \geq t\} \qquad \forall t \geq 0,$$

then $f = g$ on X.

Proof. Fix $x \in X$. Let $t < g(x)$. Then $g(x) \geq t$, hence $x \in \{g \geq t\} = \{f \geq t\}$, so $f(x) \geq t$. Since this holds for all $t < g(x)$, we get $f(x) \geq g(x)$. By symmetry, for any $t < f(x)$ we have $x \in \{f \geq t\} = \{g \geq t\}$, hence $g(x) \geq t$. Letting $t \uparrow f(x)$ yields $g(x) \geq f(x)$.

Combining both inequalities gives $f(x) = g(x)$. As x was arbitrary, $f = g$ on $\mathcal{X}$. □

Now to prove the proposition we define for each $i \in [d]$ and $x \in \mathbb{R}$,

$$p_i(x) := \sup\{t \geq 0 : \ x \in I_i(t)\}, \qquad (\sup \varnothing := 0),$$

and set

$$q(z) := \min_{i \in [d]} p_i(z_i), \qquad z = (z_1, \dots, z_d) \in \mathbb{R}^d.$$

Fix i and $t > 0$. Then we can show that

$$\{x \in \mathbb{R} : \ p_i(x) \geq t\} = I_i(t).$$

Indeed, if $x \in I_i(t)$, $p_i(x) = \sup\{s : \ x \in I_i(s)\} \geq t$. Conversely, if $p_i(x) \geq t$, we distinguish two cases.

- **Case 1** ($p_i(x) > t$): Pick any $s \in (t, p_i(x))$. By definition of the supremum, $x \in I_i(s)$. By nesting $I_i(s) \subseteq I_i(t)$, the result follows.
- **Case 2** ($p_i(x) = t$): Pick $t_n \nearrow t$ with $x \in I_i(t_n)$. We have $L_t = \bigcap_{s<t} L_s$. Thus, since each L_s is a hyperrectangle, $I_i(t) = \bigcap_{s<t} I_i(s)$. For any $s < t$, eventually $t_n > s$, so $x \in I_i(t_n) \subseteq I_i(s)$. Hence, $x \in \bigcap_{s<t} I_i(s) = I_i(t)$.

Consequently, for any $t \geq 0$,

$$\begin{aligned}\{z : \ q(z) \geq t\} &= \Big\{z : \ \min_i p_i(z_i) \geq t\Big\} = \bigcap_{i=1}^{d} \{z : \ p_i(z_i) \geq t\} \\ &= \prod_{i=1}^{d} \{x : \ p_i(x) \geq t\} = \prod_{i=1}^{d} I_i(t) = L_t = \{z : \ p(z) \geq t\}.\end{aligned}$$

So the level sets of p and q coincide, so we conclude by lemma B.1 that both functions are equal.

Proof of Proposition 4.2. Fix $t > 0$ and set

$$L_t := \{z \in \mathbb{R}^d : \ p(z) \geq t\}.$$

Recall

$$p^R(z) := \min_{i \in [d]} p_i^R(z_i), \qquad p_i^R(x) := \sup_{z_{-i} \in \mathbb{R}^{d-1}} p(x, z_{-i}),$$

and define $I_i(t) := \{x \in \mathbb{R} : \ p_i^R(x) \geq t\}$, so that

$$R_t^\star := \{z : \ p^R(z) \geq t\} = \prod_{i=1}^{d} I_i(t).$$

By definition of the supremum, all $p_i^R(z_i) \geq p(z)$ hence $p^R(z) \geq p(z)$. Therefore, $\{p \geq t\} \subset \{p^R \geq t\} = R_t^\star$ so the rectangular set $R_t^\star$ contains L_t.

To prove that it is smaller in volume than any other hyperrectangle containing L_t, let

$$H = \prod_{i=1}^{d} J_i$$

be any hyperrectangle with $L_t \subseteq H$.

Fix i and $x \in I_i(t)$. Then $p_i^R(x) \geq t$, i.e. $\sup_{z_{-i}} p(x, z_{-i}) \geq t$. Pick $z_{-i}^{(n)}$ such that $p(x, z_{-i}^{(n)}) \geq t - 2^{-n}$. Then eventually we have $p(x, z_{-i}^{(n)}) \geq t/2$, meaning that $(x, z_{-i}^{(n)}) \in L_{t/2}$ for a sufficiently large n. By compactness of the superlevel-sets, we can extract a sub-sequence $(x, z_{-i}^{(n_k)})$ converging to $(x, z_{-i}^*) \in L_{t/2}$. Thus, by continuity of p,

$$p(x, z_{-i}^*) = \lim_{k \to \infty} p(x, z_{-i}^{(n_k)}) \geq t.$$

Hence $(x, z_{-i}^*) \in L_t \subseteq H$, and since $H = \prod_i J_i$ we obtain $x \in J_i$. So $I_i(t) \subseteq J_i$ for all i, and therefore

$$R_t^\star = \prod_{i=1}^{d} I_i(t) \subseteq \prod_{i=1}^{d} J_i = H.$$

Thus $R_t^\star$ contains L_t and is contained in every hyperrectangle containing L_t, so it is the smallest such hyperrectangle; in particular, $\lambda(R_t^\star) \leq \lambda(H)$ for all $H \supseteq L_t$. □

B.3 Rectangular envelope method

Explicit form of the pairwise RE score. To find the supremum of $p(y_1, y_2)$ over y_1, we can use the product rule $p(y_1, y_2) = p(y_2)p(y_1|y_2)$. Since $p(y_1|y_2)$ is $\mathcal{N}(\mu_{1|2}, \sigma^2_{1|2})$, its maximum value is $1/\sqrt{2\pi\sigma^2_{1|2}}$. Thus:

$$\begin{aligned} p_2^R(y_2) := \sup_{y_1} p(y_1, y_2) &= p(y_2) \cdot \max_{y_1} p(y_1|y_2) \\ &= \frac{1}{\sqrt{2\pi\sigma_{22}}} \exp\left(-\frac{(y_2-\mu_2)^2}{2\sigma_{22}}\right) \cdot \frac{1}{\sqrt{2\pi\sigma^2_{1|2}}} \\ &= \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{(y_2-\mu_2)^2}{2\sigma_{22}}\right) \end{aligned}$$

The prediction set writes $\{-\log(p^R(y_1, y_2|x) \leq \hat{q}\} = \{p^R(y_1, y_2|x) \geq e^{-\hat{q}}\} = \{p_1^R(y_1, |x) \geq e^{-\hat{q}}\} \times \{p_2^R(y_2|x) \geq e^{-\hat{q}}\}$ where $\hat{q}$ denotes the an empirical quantile resulting of a calibration strategy for the calibration scores $\{-\log p^R(X_i, Y_i)\}_{i=1}^n$. This means that the final rectangle per pair is

$$R(x) := \prod_{i=1}^{2} \left[\mu_i(x) - \sqrt{2\sigma_{ii}(x)(\hat{q} - \ln(2\pi\sqrt{|\Sigma(x)|}))}, \mu_i(x) + \sqrt{2\sigma_{ii}(x)(\hat{q} - \ln(2\pi\sqrt{|\Sigma(x)|}))}\right]$$

Remark on DR-CP prediction set volume. *The volume of the high-dimensional prediction set of the DR-CP method is typically estimated through a Monte Carlo estimation (Dheur et al., 2025). In our experimental setup where we consider a MVG model, we fix this problem because the DR-CP prediction set has explicit volume because it is an d-dimensional ellipsoid.*

C Related Work

We distinguish two categories of multi-output CP methods, each yielding prediction sets of different structure and geometry.

Hyperrectangular regions. These methods construct prediction sets as a Cartesian product of intervals for each individual output, offering the highest level of interpretability. While early work by Neeven and Smirnov (2018) applied this to weather forecasting, naive approaches like applying univariate SCP per dimension with Bonferroni correction often yield overly conservative regions. To improve this approach, Schlembach et al. (2025) formulates coverage allocation as an optimization problem to improve efficiency, without taking balance into account. Sampson and Chan (2024) proposes a method without covariance estimation that uses dimension-adaptive scores, which achieves asymptotic balance under strong assumptions. Fan and Sesia (2025); Zhou et al. (2024) propose a similar framework, but don't consider balance. To better capture output dependencies, Messoudi et al. (2021) fits copulas to the NCSs, which in practice is less efficient than Max-Rank correction (Timans et al., 2025). These approaches usually lack efficiency and fail to capture the dependencies of the dimensional outputs.

Flexible Regions. Many multi-output CP methods yielding flexible regions typically follow density-based or sample-based approaches. Density-based methods utilize highest density regions (Izbicki et al., 2020; Dheur et al., 2024), but their scoring functions often become intractable as dimensionality increases. Sample-based methods leverage generative models (Feldman et al., 2023; Wang et al., 2022; Plassier et al., 2024) or optimal-transport approaches (Thurin et al., 2025; Klein et al., 2025), yet these can make the prediction set random and are often computationally intensive and hard to interpret. While (Dheur et al., 2025) provides a comprehensive overview of these techniques, most produce complex sets lacking an explicit analytical form. This lack of structural simplicity often renders them less interpretable and difficult to apply in practical settings.

Ellipse and Convex sets. These methods use the empirical covariance matrices to shape prediction regions (Johnstone and Cox, 2021; Messoudi et al., 2022; Henderson et al., 2024) or fit convex set to the residuals (Tumu et al., 2024; Braun et al., 2025a). These methods remain restricted to convex shapes and assume an underlying elliptical structure. Although they can be more interpretable because the volume and shape of the prediction is explicit, they are still not interpretable when the dimension exceeds 3.

D Additional experimental results

D.1 Experimental setup

We evaluate on a collection of real-world multi-output regression datasets commonly used in multi-output CP regression: one dataset from (Camehl et al., 2025), two datasets from (Tsoumakas et al., 2011), 1 dataset fro m (Cevid et al., 2022), 2 datasets from (Feldman et al., 2023) and one dataset from (Wang et al., 2022). The characteristics of the datasets are provided in the table below:

Table 1: Characteristics of each dataset considered in this study

Source	**Dataset**	**Nb instances**	**Nb features** p	**Nb targets** d
Camehl	households	7207	4	4
Mulan	rf1	9005	64	8
Mulan	rf2	9005	64	8
Cevid	air	10000	15	6
Feldman	house	21613	14	2
Feldman	bio	45730	8	2
Wang	taxi	50000	4	2

We apply a preprocessing pipeline inspired by prior work on multi-output conformal regression (Dheur et al., 2025), and normalize our data using a quantile transform as in (Braun et al., 2025b).

After preprocessing, we limit the maximum size of each dataset to 10000, which we split each into 50% training, 10% validation, 20% calibration, and 20% test. For the conditional multivariate gaussian model, we use MLP backbone with hidden dimension 256 and 3 residual layers, and AdamW with learning rate 10^{-3}, batch size 32, and train for 150 epochs. The details of the model can be found in (Braun et al., 2025b).

All methods use the same conditional multivariate Gaussian (MVG) model $p_\theta(y|x)$. Depending on the method, we derive either point predictions (the conditional mean), quantile-based scores from marginal Gaussian quantiles, or density-based scores from the negative log-likelihood.

D.2 Metrics

We report standard evaluation metrics for multi-output conformal prediction.

Empirical coverage. For a prediction set $\mathcal{C}_\alpha(X)$ at level α, the empirical coverage over a test sample $\{(X_i, Y_i)\}_{i=1}^{n_{\text{test}}}$ is

$$\widehat{\text{Cov}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\{Y_i \in \mathcal{C}_\alpha(X_i)\}.$$

To assess balance, we also compute coverage over sub-prediction sets. Let $\mathcal{C}_\alpha^j(X)$ denote the sub-prediction set associated with group (or coordinate) G_j. The empirical coverage of G_j is

$$\widehat{\text{Cov}}_j = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\{Y^{G_j} \in \mathcal{C}_\alpha^j(X_i)\}.$$

For hyperrectangular baselines, each group is a singleton (i.e., per-dimension coverage). For our methods, groups correspond to output pairs (e.g., $(1, 2)$ and $(3, 4)$).

Log-volume. To assess efficiency, we measure the size of the prediction set via its volume. For hyperrectangular sets

$$\mathcal{C}_\alpha(X_i) = \prod_{j=1}^{d} [\ell_{ij}, u_{ij}],$$

the volume is

$$\mathrm{Vol}_i = \prod_{j=1}^{d} (u_{ij} - \ell_{ij}),$$

and we report the median log-volume across test points:

$$\mathrm{Quantile}_{0.5}\left(\log \mathrm{Vol}_i\right).$$

Balance. To quantify how evenly coverage is distributed across groups, we compute

$$\mathrm{Balance} = \max_g \left| \widehat{\mathrm{Cov}}_g - \overline{\mathrm{Cov}} \right|,$$

where $\widehat{\mathrm{Cov}}_g$ denotes the empirical coverage of group g and $\overline{\mathrm{Cov}}$ is the average coverage across groups. Lower values indicate more uniform (balanced) coverage.

D.3 Additional results

We report tables of mean logvolume and empirical coverage and balance for all datasets for $\alpha \in \{0.1, 0.2\}$ across 10 seeds. Results highlighted in bold denote smallest volume, underline means second best.

As expected DR-CP is always the smallest, then our variants come next, especially MR-PICP-PR then Bonf-PICP-PR, and both show small groupwise balance compared to other methods.

Table 2: Results on `RF1` for $\alpha = 0.1$ and $\alpha = 0.2$ (mean ± std).

	$\alpha = 0.1$			$\alpha = 0.2$		
Method	Coverage	LogVol.	Balance	Coverage	LogVol.	Balance
Bonf-PICP-PR	0.911 ± 0.009	-5.436 ± 0.225	0.006 ± 0.003	0.835 ± 0.012	-7.078 ± 0.156	0.008 ± 0.005
Bonferroni	0.914 ± 0.008	-2.682 ± 0.450	0.006 ± 0.002	0.840 ± 0.013	-5.446 ± 0.303	0.007 ± 0.002
CHR	0.901 ± 0.015	-3.661 ± 0.508	0.013 ± 0.006	0.806 ± 0.020	-6.354 ± 0.436	0.018 ± 0.004
CHR-CQR	0.902 ± 0.008	-4.550 ± 0.216	0.013 ± 0.005	0.801 ± 0.012	-6.261 ± 0.187	0.022 ± 0.007
DCP-mt	0.904 ± 0.008	-3.552 ± 0.344	0.011 ± 0.004	0.830 ± 0.013	-5.763 ± 0.292	0.015 ± 0.006
DR-CP	0.897 ± 0.010	**-8.145 ± 0.197**	–	0.801 ± 0.013	**-9.640 ± 0.135**	–
Empirical Copula	0.902 ± 0.011	-3.320 ± 0.384	0.007 ± 0.002	0.801 ± 0.015	-6.318 ± 0.301	0.009 ± 0.002
MR-PICP-PR	0.902 ± 0.010	-5.683 ± 0.175	0.006 ± 0.002	0.801 ± 0.009	-7.612 ± 0.127	0.008 ± 0.005
M-CP	0.899 ± 0.008	-3.974 ± 0.258	0.031 ± 0.006	0.799 ± 0.010	-5.607 ± 0.184	0.063 ± 0.009
Max-PICP-PR	0.900 ± 0.008	-5.503 ± 0.161	0.022 ± 0.006	0.799 ± 0.013	-7.216 ± 0.170	0.055 ± 0.007
Max-Rank	0.898 ± 0.012	-3.461 ± 0.363	0.007 ± 0.002	0.798 ± 0.015	-6.369 ± 0.289	0.009 ± 0.002

Table 3: Results on `RF2` for $\alpha = 0.1$ and $\alpha = 0.2$ (mean ± std).

	$\alpha = 0.1$			$\alpha = 0.2$		
Method	Coverage	LogVol.	Balance	Coverage	LogVol.	Balance
Bonf-PICP-PR	0.907 ± 0.008	-5.345 ± 0.240	0.006 ± 0.002	0.831 ± 0.012	-6.877 ± 0.203	0.008 ± 0.003
Bonferroni	0.911 ± 0.012	-2.845 ± 0.274	0.006 ± 0.001	0.834 ± 0.011	-5.330 ± 0.340	0.008 ± 0.003
CHR	0.899 ± 0.009	-3.439 ± 0.230	0.011 ± 0.002	0.800 ± 0.015	-6.104 ± 0.443	0.018 ± 0.005
CHR-CQR	0.896 ± 0.007	-4.356 ± 0.640	0.014 ± 0.005	0.797 ± 0.018	-6.006 ± 0.568	0.022 ± 0.010
DCP-mt	0.899 ± 0.011	-3.491 ± 0.264	0.010 ± 0.004	0.825 ± 0.013	-5.568 ± 0.353	0.017 ± 0.005
DR-CP	0.900 ± 0.008	**-7.943 ± 0.172**	–	0.797 ± 0.014	**-9.505 ± 0.158**	–
Empirical Copula	0.900 ± 0.011	-3.350 ± 0.308	0.006 ± 0.001	0.795 ± 0.017	-6.184 ± 0.368	0.010 ± 0.003
MR-PICP-PR	0.899 ± 0.007	-5.569 ± 0.266	0.006 ± 0.002	0.797 ± 0.015	-7.413 ± 0.183	0.010 ± 0.004
M-CP	0.897 ± 0.010	-3.845 ± 0.712	0.031 ± 0.007	0.802 ± 0.016	-5.402 ± 0.659	0.059 ± 0.007
Max-PICP-PR	0.893 ± 0.012	-5.409 ± 0.217	0.022 ± 0.006	0.798 ± 0.013	-6.976 ± 0.158	0.054 ± 0.007
Max-Rank	0.896 ± 0.010	-3.469 ± 0.335	0.006 ± 0.001	0.793 ± 0.016	-6.230 ± 0.370	0.010 ± 0.004

Table 4: Results on `bio` for $\alpha = 0.1$ and $\alpha = 0.2$ (mean ± std).

	$\alpha = 0.1$			$\alpha = 0.2$		
Method	Coverage	LogVol.	Balance	Coverage	LogVol.	Balance
Bonf-PICP-PR	0.896 ± 0.007	0.482 ± 0.094	0.000 ± 0.000	0.796 ± 0.009	-0.020 ± 0.109	0.000 ± 0.000
Bonferroni	0.902 ± 0.009	0.588 ± 0.037	0.003 ± 0.003	0.825 ± 0.009	0.075 ± 0.044	0.006 ± 0.003
CHR	0.893 ± 0.010	0.519 ± 0.060	0.006 ± 0.003	0.794 ± 0.013	-0.085 ± 0.066	0.007 ± 0.003
CHR-CQR	0.892 ± 0.010	0.529 ± 0.081	0.008 ± 0.008	0.798 ± 0.013	0.019 ± 0.089	0.019 ± 0.013
DCP-mt	0.900 ± 0.008	0.574 ± 0.032	0.006 ± 0.004	0.821 ± 0.009	0.060 ± 0.044	0.013 ± 0.007
DR-CP	0.895 ± 0.009	**0.287 ± 0.089**	–	0.797 ± 0.014	**-0.256 ± 0.123**	–
Empirical Copula	0.893 ± 0.008	0.521 ± 0.027	0.004 ± 0.003	0.793 ± 0.008	-0.089 ± 0.052	0.006 ± 0.004
MR-PICP-PR	0.896 ± 0.007	0.482 ± 0.094	0.000 ± 0.000	0.796 ± 0.009	-0.020 ± 0.109	0.000 ± 0.000
M-CP	0.893 ± 0.012	0.503 ± 0.089	0.008 ± 0.007	0.799 ± 0.014	-0.002 ± 0.096	0.021 ± 0.011
Max-PICP-PR	0.896 ± 0.007	0.482 ± 0.094	0.000 ± 0.000	0.796 ± 0.009	-0.020 ± 0.109	0.000 ± 0.000
Max-Rank	0.893 ± 0.008	0.519 ± 0.027	0.004 ± 0.003	0.793 ± 0.008	-0.090 ± 0.051	0.006 ± 0.004

Table 5: Results on `Households` for $\alpha = 0.1$ and $\alpha = 0.2$ (mean ± std).

	$\alpha = 0.1$			$\alpha = 0.2$		
Method	Coverage	LogVol.	Balance	Coverage	LogVol.	Balance
Bonf-PICP-PR	0.897 ± 0.012	4.188 ± 0.124	0.005 ± 0.005	0.809 ± 0.016	3.500 ± 0.088	0.005 ± 0.005
Bonferroni	0.904 ± 0.011	4.566 ± 0.080	0.006 ± 0.003	0.813 ± 0.014	3.864 ± 0.090	0.009 ± 0.006
CHR	0.893 ± 0.015	4.471 ± 0.111	0.011 ± 0.005	0.788 ± 0.023	3.694 ± 0.110	0.022 ± 0.010
CHR-CQR	0.891 ± 0.012	4.237 ± 0.152	0.014 ± 0.006	0.789 ± 0.014	3.460 ± 0.133	0.021 ± 0.006
DCP-mt	0.894 ± 0.008	4.502 ± 0.070	0.014 ± 0.008	0.806 ± 0.016	3.804 ± 0.083	0.021 ± 0.014
DR-CP	0.891 ± 0.018	**3.711 ± 0.143**	–	0.781 ± 0.022	**2.987 ± 0.098**	–
Empirical Copula	0.897 ± 0.012	4.498 ± 0.079	0.006 ± 0.003	0.788 ± 0.016	3.702 ± 0.084	0.010 ± 0.005
MR-PICP-PR	0.892 ± 0.014	4.141 ± 0.128	0.004 ± 0.005	0.791 ± 0.014	3.386 ± 0.076	0.005 ± 0.005
M-CP	0.892 ± 0.013	4.305 ± 0.123	0.016 ± 0.004	0.793 ± 0.017	3.569 ± 0.102	0.036 ± 0.009
Max-PICP-PR	0.893 ± 0.013	4.203 ± 0.115	0.013 ± 0.007	0.790 ± 0.016	3.424 ± 0.078	0.034 ± 0.007
Max-Rank	0.895 ± 0.012	4.481 ± 0.078	0.006 ± 0.003	0.785 ± 0.016	3.688 ± 0.082	0.010 ± 0.005

Table 6: Results on `air` for $\alpha = 0.1$ and $\alpha = 0.2$ (mean ± std).

	$\alpha = 0.1$			$\alpha = 0.2$		
Method	Coverage	LogVol.	Balance	Coverage	LogVol.	Balance
Bonf-PICP-PR	0.914 ± 0.011	8.107 ± 0.131	0.004 ± 0.003	0.833 ± 0.015	7.219 ± 0.088	0.006 ± 0.004
Bonferroni	0.917 ± 0.009	8.375 ± 0.081	0.006 ± 0.002	0.842 ± 0.016	7.541 ± 0.095	0.009 ± 0.003
CHR	0.899 ± 0.015	8.152 ± 0.149	0.006 ± 0.002	0.810 ± 0.019	7.261 ± 0.132	0.011 ± 0.003
CHR-CQR	0.902 ± 0.011	7.973 ± 0.104	0.008 ± 0.003	0.800 ± 0.014	6.950 ± 0.089	0.011 ± 0.004
DCP-mt	0.907 ± 0.012	8.274 ± 0.093	0.009 ± 0.003	0.833 ± 0.015	7.460 ± 0.082	0.014 ± 0.004
DR-CP	0.899 ± 0.010	**6.459 ± 0.094**	–	0.799 ± 0.015	**5.616 ± 0.069**	–
Empirical Copula	0.903 ± 0.013	8.176 ± 0.098	0.007 ± 0.002	0.803 ± 0.016	7.179 ± 0.102	0.009 ± 0.003
MR-PICP-PR	0.900 ± 0.012	7.922 ± 0.115	0.005 ± 0.003	0.801 ± 0.016	6.934 ± 0.078	0.006 ± 0.004
M-CP	0.903 ± 0.009	7.912 ± 0.082	0.006 ± 0.002	0.798 ± 0.016	6.919 ± 0.064	0.009 ± 0.003
Max-PICP-PR	0.901 ± 0.012	7.925 ± 0.121	0.005 ± 0.003	0.801 ± 0.017	6.940 ± 0.081	0.008 ± 0.004
Max-Rank	0.901 ± 0.012	8.155 ± 0.097	0.007 ± 0.002	0.802 ± 0.017	7.170 ± 0.102	0.009 ± 0.003

Table 7: Results on `taxi` for $\alpha = 0.1$ and $\alpha = 0.2$ (mean ± std).

	$\alpha = 0.1$			$\alpha = 0.2$		
Method	Coverage	LogVol.	Balance	Coverage	LogVol.	Balance
Bonf-PICP-PR	0.899 ± 0.009	2.554 ± 0.037	0.000 ± 0.000	0.796 ± 0.013	2.106 ± 0.047	0.000 ± 0.000
Bonferroni	0.908 ± 0.009	2.606 ± 0.029	0.002 ± 0.002	0.819 ± 0.010	2.205 ± 0.031	0.004 ± 0.003
CHR	0.897 ± 0.012	2.556 ± 0.058	0.005 ± 0.004	0.799 ± 0.009	2.117 ± 0.043	0.006 ± 0.004
CHR-CQR	0.899 ± 0.009	2.554 ± 0.034	0.004 ± 0.003	0.796 ± 0.011	2.081 ± 0.040	0.006 ± 0.003
DCP-mt	0.903 ± 0.009	2.593 ± 0.031	0.010 ± 0.005	0.814 ± 0.008	2.189 ± 0.031	0.018 ± 0.014
DR-CP	0.898 ± 0.010	**2.472 ± 0.047**	–	0.796 ± 0.013	**1.960 ± 0.041**	–
Empirical Copula	0.898 ± 0.009	2.560 ± 0.037	0.003 ± 0.003	0.797 ± 0.010	2.109 ± 0.042	0.005 ± 0.004
MR-PICP-PR	0.899 ± 0.009	2.554 ± 0.037	0.000 ± 0.000	0.796 ± 0.013	2.106 ± 0.047	0.000 ± 0.000
M-CP	0.899 ± 0.010	2.553 ± 0.038	0.003 ± 0.002	0.797 ± 0.012	2.084 ± 0.043	0.005 ± 0.004
Max-PICP-PR	0.899 ± 0.009	2.554 ± 0.037	0.000 ± 0.000	0.796 ± 0.013	2.106 ± 0.047	0.000 ± 0.000
Max-Rank	0.897 ± 0.009	2.556 ± 0.036	0.003 ± 0.003	0.797 ± 0.010	2.108 ± 0.039	0.005 ± 0.004

Table 8: Results on `house` for $\alpha = 0.1$ and $\alpha = 0.2$ (mean ± std).

	$\alpha = 0.1$			$\alpha = 0.2$		
Method	Coverage	LogVol.	Balance	Coverage	LogVol.	Balance
Bonf-PICP-PR	0.903 ± 0.008	1.414 ± 0.060	0.000 ± 0.000	0.801 ± 0.010	0.929 ± 0.055	0.000 ± 0.000
Bonferroni	0.909 ± 0.010	1.853 ± 0.056	0.005 ± 0.003	0.824 ± 0.015	1.368 ± 0.046	0.004 ± 0.004
CHR	0.900 ± 0.013	1.806 ± 0.064	0.005 ± 0.004	0.803 ± 0.015	1.272 ± 0.050	0.009 ± 0.008
CHR-CQR	0.904 ± 0.008	1.440 ± 0.061	0.008 ± 0.004	0.799 ± 0.011	0.930 ± 0.058	0.014 ± 0.008
DCP-mt	0.904 ± 0.010	1.832 ± 0.051	0.009 ± 0.006	0.817 ± 0.012	1.353 ± 0.041	0.017 ± 0.010
DR-CP	0.902 ± 0.009	**1.237 ± 0.071**	–	0.803 ± 0.013	**0.742 ± 0.040**	–
Empirical Copula	0.901 ± 0.012	1.804 ± 0.063	0.004 ± 0.003	0.802 ± 0.012	1.262 ± 0.052	0.004 ± 0.004
MR-PICP-PR	0.903 ± 0.008	1.414 ± 0.060	0.000 ± 0.000	0.801 ± 0.010	0.929 ± 0.055	0.000 ± 0.000
M-CP	0.903 ± 0.005	1.466 ± 0.065	0.006 ± 0.003	0.799 ± 0.012	0.966 ± 0.063	0.013 ± 0.006
Max-PICP-PR	0.903 ± 0.008	1.414 ± 0.060	0.000 ± 0.000	0.801 ± 0.010	0.929 ± 0.055	0.000 ± 0.000
Max-Rank	0.901 ± 0.012	1.803 ± 0.062	0.004 ± 0.003	0.801 ± 0.012	1.261 ± 0.053	0.004 ± 0.004

references

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and trends® in machine learning*, 16(4):494–591, 2023.

Yoav Benjamini. Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal*, 52(6):708–721, 2010.

Sacha Braun, Liviu Aolaritei, Michael I. Jordan, and Francis Bach. Minimum volume conformal sets for multivariate regression, 2025a. URL https://arxiv.org/abs/2503.19068.

Sacha Braun, Eugène Berta, Michael I Jordan, and Francis Bach. Multivariate conformal prediction via conformalized gaussian scoring. *arXiv preprint arXiv:2507.20941*, 2025b.

Annika Camehl, Dennis Fok, and Kathrin Gruber. On superlevel sets of conditional densities and multivariate quantile regression. *Journal of Econometrics*, 249:105807, 2025.

Domagoj Cevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022.

Victor Dheur, Tanguy Bosser, Rafael Izbicki, and Souhaib Ben Taieb. Distribution-free conformal joint prediction regions for neural marked temporal point processes. *Machine Learning*, 113(9):7055–7102, 2024.

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. In *The 42nd International Conference on Machine Learning*, 2025.

Yunjie Fan and Matteo Sesia. Interpretable multivariate conformal prediction with fast transductive standardization, 2025. URL https://arxiv.org/abs/2512.15383.

Shai Feldman, Stephen Bates, and Yaniv Romano. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

Iain Henderson, Adrien Mazoyer, and Fabrice Gamboa. Adaptive inference with random ellipsoids through conformal conditional linear expectation. *arXiv preprint arXiv:2409.18508*, 2024.

Rafael Izbicki, Gilson Shimizu, and Rafael Stern. Flexible distribution-free conditional predictive bands using density estimators. In *International Conference on Artificial Intelligence and Statistics*, pages 3068–3077. PMLR, 2020.

Chancellor Johnstone and Bruce Cox. Conformal uncertainty sets for robust optimization. In *Conformal and Probabilistic Prediction and Applications*, pages 72–90. PMLR, 2021.

Michal Klein, Louis Bethune, Eugene Ndiaye, and Marco Cuturi. Multivariate conformal prediction using optimal transport. *arXiv preprint arXiv:2502.03609*, 2025.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

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

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Ellipsoidal conformal inference for multi-target regression. In *Conformal and Probabilistic Prediction with Applications*, pages 294–306. PMLR, 2022.

Jelmer Neeven and Evgueni Smirnov. Conformal stacked weather forecasting. In *Conformal and Probabilistic Prediction and Applications*, pages 220–233. PMLR, 2018.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pages 345–356. Springer, 2002.

Vincent Plassier, Alexander Fishkov, Mohsen Guizani, Maxim Panov, and Eric Moulines. Probabilistic conformal prediction with approximate conditional validity. *arXiv preprint arXiv:2407.01794*, 2024.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Max Sampson and Kung-Sik Chan. Conformal multi-target hyperrectangles. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 17(5):e11710, 2024.

Filip Schlembach, Evgueni Smirnov, and Mark HM Winands. Dynamic conformal prediction for multi-target regression: Optimising informational efficiency under joint validity. In *Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2025)*, pages 193–213. PMLR, 2025.

Philip Sedgwick. Multiple significance tests: the bonferroni correction. *Bmj*, 344, 2012.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9 (3), 2008.

Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. *arXiv preprint arXiv:2501.18991*, 2025.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, Christian A Naesseth, and Eric Nalisnick. Max-rank: Efficient multiple testing for conformal prediction. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research*, 12:2411–2414, 2011.

Renukanandan Tumu, Matthew Cleaveland, Rahul Mangharam, George Pappas, and Lars Lindemann. Multimodal conformal prediction regions by optimizing convex shape templates. In *6th Annual Learning for Dynamics & Control Conference*, pages 1343–1356. PMLR, 2024.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Vladimir Vovk, Bin Wang, and Ruodu Wang. Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, 50(1):351–375, 2022.

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David M Blei. Probabilistic conformal prediction using conditional random samples. *arXiv preprint arXiv:2206.06584*, 2022.

Yanfei Zhou, Lars Lindemann, and Matteo Sesia. Conformalized adaptive forecasting of heterogeneous trajectories. *arXiv preprint arXiv:2402.09623*, 2024.