Optimal Transport-based Conformal Prediction

Gauthier Thurin, Kimia Nadjahi, Claire Boyer

Optimal Transport-based Conformal Prediction

Gauthier Thurin*[1], Kimia Nadjahi[1], and Claire Boyer[2]

[1]*CNRS, ENS Paris, France*
[2]*LMO, Université Paris-Saclay, Orsay, France ; Institut universitaire de France*

Abstract

Conformal Prediction (CP) is a principled framework for quantifying uncertainty in black-box learning models, by constructing prediction sets with finite-sample coverage guarantees. Traditional approaches rely on scalar nonconformity scores, which fail to fully exploit the geometric structure of multivariate outputs, such as in multi-output regression or multiclass classification. Recent methods addressing this limitation impose predefined convex shapes for the prediction sets, potentially misaligning with the intrinsic data geometry. We introduce a novel CP procedure handling multivariate score functions through the lens of optimal transport. Specifically, we leverage Monge-Kantorovich vector ranks and quantiles to construct prediction region with flexible, potentially non-convex shapes, better suited to the complex uncertainty patterns encountered in multivariate learning tasks. We prove that our approach ensures finite-sample, distribution-free coverage properties, similar to typical CP methods. We then adapt our method for multi-output regression and multiclass classification, and also propose simple adjustments to generate adaptive prediction regions with asymptotic conditional coverage guarantees. Finally, we evaluate our method on practical regression and classification problems, illustrating its advantages in terms of (conditional) coverage and efficiency.

1 Introduction

In various domains, including high-stakes applications, state-of-the-art performances are often achieved by black-box machine learning models. As a result, accurately quantifying the uncertainty of their predictions has become a critical priority. Conformal Prediction (CP, Vovk et al., 2005) has emerged as a compelling framework to address this need, by generating prediction sets with *coverage guarantees* (ensuring they contain the true outcome with a specified confidence level) regardless of the model or data distribution. Most CP methods are thus model-agnostic and distribution-free while easy to implement, which explain their growing popularity in recent years.

The main idea of CP is to convert a set of *non-conformity scores* into reliable uncertainty sets using *quantiles*. Non-conformity scores are empirical measurements of how unusual a prediction is. For example, in regression, the score can be defined as $|\hat{y} - y|$, where $\hat{y} \in \mathbb{R}$ is the model's prediction and $y \in \mathbb{R}$ the true response (Lei et al., 2018). These scores are central to the CP framework as they encapsulate the uncertainty stemming from both the model and the data, directly influencing the size and shape of the resulting prediction sets. Therefore, the quality of the prediction sets hinges on the relevance of the chosen non-conformity score: while a poorly designed score may still achieve the required coverage guarantee, it often leads to overly conservative or inefficient prediction

*Corresponding author: gthurin@mail.di.ens.fr

sets, failing to capture the complex patterns of the underlying data distribution (Angelopoulos and Bates, 2023).

Most CP approaches rely on *scalar* non-conformity scores (*e.g.*, Angelopoulos and Bates, 2023; Romano et al., 2020; Cauchois et al., 2021; Sesia and Romano, 2021; Lei et al., 2018). Although conceptually simple, such one-dimensional representations can be too restrictive or poorly suited in applications that require multivariate prediction sets. To circumvent this, recently-proposed CP methods seek to incorporate correlations despite the use of scalar scores, by leveraging techniques such as copulas (Messoudi et al., 2021) or ellipsoids (Johnstone and Cox, 2021; Messoudi et al., 2022; Henderson et al., 2024). Nevertheless, these approaches either lack finite-sample coverage guarantees or impose restrictive modeling assumptions that prescribe the shape of the prediction region. Feldman et al. (2023) recently proposed a CP method able to construct more adaptive prediction regions with non-convex shapes, establishing a connection with multivariate quantiles. However, their method cannot be directly applied to a black-box model, thus fails to meet one of the key desiderata of standard CP.

Contributions. In this work, we introduce a novel general CP framework that accommodates multivariate scores, enabling more expressive representations of prediction errors. The core idea is to leverage *Monge-Kantorovich (MK) quantiles* (Chernozhukov et al., 2017; Hallin et al., 2021), a multivariate extension of traditional scalar quantiles rooted in optimal transport theory. MK quantiles are constructed by mapping multidimensional scores onto a reference distribution. The resulting CP framework, called OT-CP for Optimal Transport-based Conformal Prediction, effectively captures the structure and dependencies within multivariate data while ensuring distribution-free ranks, thanks to the distinctive properties of MK quantiles Deb and Sen (2023); Hallin et al. (2021). This distribution-freeness property allows us to establish a multivariate extension of the quantile lemma, a key result in standard CP theory. Building on this, we demonstrate that OT-CP constructs prediction regions with finite-sample coverage guarantees. These hold for any choice of multivariate score function, which makes OT-CP a robust and practical tool to address complex uncertainty quantification task.

After presenting the general OT-CP methodology with its theoretical guarantees (Section 2), we apply it on two typical learning tasks: multi-output regression (Section 3) and classification (Section 4). For each of these, we use multivariate score functions which, when integrated in OT-CP, yield prediction regions that effectively capture correlations between the score dimensions. In the context of regression, we also develop an extension of OT-CP that conditionally adapts to input covariates, further enhancing the flexibility of our method. Moreover, we show that this adaptive version provably reaches asymptotic conditional coverage. These two case studies serve a dual purpose: they highlight the versatility and user-friendliness of OT-CP while offering concrete frameworks to evaluate its benefits over existing methods through numerical experiments. In doing so, we believe this lays a solid foundation for future explorations of OT-CP across a wider range of applications.

2 Methodology

2.1 Setting

We consider a pre-trained black-box model $\hat{f} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ respectively denote the input and output spaces of the learning task. Assume we have access to a set of n exchangeable observations $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, not used during the training of $\hat{f}$ and referred to as the *calibration set*. Consider a *score function* $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+^d$ that produces $d \geq 1$ *non-conformity scores*, measuring the discrepancies between the target Y_i and the prediction $\hat{f}(X_i)$.

Considering a multivariate score in the context of CP departs from typical strategies, which rely on scalar scores. Such multivariate scores can particularly be useful for quantifying uncertainties, as described in the examples below.

Example 1 (Multi-output regression)**.** *In multi-output regression, both the response Y and prediction $\hat{f}(X)$ take values in $\mathbb{R}^d$. One can consider multivariate scores $s(Y, \hat{f}(X))$ corresponding to component-wise prediction errors (see Section 3), without the need of aggregating them into a single value (e.g., by considering the mean squared error). Figure 1(a) illustrates 2-dimensional scores in a context of bivariate regression.*

Example 2 (Multiclass classification)**.** *Consider a classification setting with $K \geq 3$ classes and let $\hat{\pi}(x) = \{\hat{\pi}_k(x)\}_{k=1}^K$ be the estimated class probabilities returned by a classifier for some input x. Denote by $\bar{y} = \{\mathbb{1}_{k=y}\}_{k=1}^K$ the one-hot encoding of a label y. A multivariate score can be formed as the component-wise absolute difference*

$$s(x, y) = |\hat{\pi}(x) - \bar{y}| \in \mathbb{R}_+^K \,. \tag{1}$$

This score retains K-dimensional predictive information, allowing for the exploration of correlations between its components. For instance, when $K = 3$, consider two inputs x_1 and x_2 with output probabilities $\hat{\pi}(x_1) = (0.6, 0.4, 0)$ and $\hat{\pi}(x_2) = (0, 0.4, 0.6)$. For both predictions, assessing the conformity of $y = 2$ with a typical score $1 - \hat{\pi}_y(x)$ used in CP for classification would return the same value of 0.6. This potentially ignores that co-occurrences between labels 1 and 2 might be more frequent than between 2 and 3. In contrast, the multivariate alternative (1) distinguishes these two probability profiles, as $s(x_1, y) \neq s(x_2, y)$. This can be more helpful to capture the underlying confusion patterns of the predictor across different label modalities.

In the rest of the paper, we denote by $\{S_i\}_{i=1}^n = \{s(X_i, Y_i)\}_{i=1}^n$ the scores computed on the calibration set.

2.2 Optimal transport toolbox

In the context of conformal prediction, dealing with multivariate scores implies defining an adequate notion of multivariate quantiles. To do so, we view the non-conformity scores $\{S_i\}_{i=1}^n$ through the empirical distribution $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{S_i}$ and leverage optimal transport (OT) tools, more specifically, *Monge-Kantorovich quantiles.*

Definition 2.1 (Empirical Monge-Kantorovich ranks, Chernozhukov et al. (2017); Hallin et al. (2021))**.** *Consider the reference rank vectors $\{U_i\}_{i=1}^n$ given by*

$$\forall i \in \{1, \ldots, n\}, \;\; U_i = \frac{i}{n}\theta_i \,, \tag{2}$$

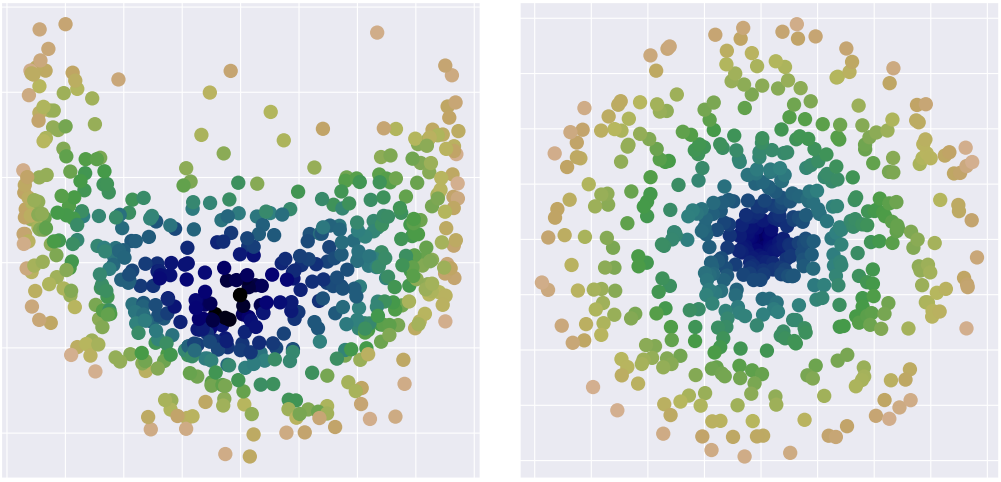
where θ_i are i.i.d. random vectors drawn uniformly on the Euclidean sphere $\mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$. The Monge-Kantorovich rank map *is defined for any score $s \in \mathbb{R}^d$ as*

$$\mathbf{R}_n(s) = \underset{U_i : 1 \leq i \leq n}{\operatorname{argmax}} \left\{ \langle U_i, s \rangle - \psi_n(U_i) \right\}, \tag{3}$$

with ψ_n the potential solving the dual of Kantorovich's OT problem, i.e.,

$$\psi_n = \underset{\varphi}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \varphi(U_i) + \frac{1}{n}\sum_{i=1}^n \varphi^*(S_i),$$

where the optimization is performed over the set of lower-semicontinuous convex functions, and $\varphi^*(x) = \sup_u\{\langle x, u\rangle - \varphi(u)\}$ is the Legendre transform of a convex function φ.



(a) Multivariate scores $\{S_i\}_{i=1}^n$ (b) Reference rank vectors $\{U_i\}_{i=1}^n$

Figure 1: Ranking multivariate scores using optimal transport. The colormap encodes how the 2-dimensional scores $\{S_i\}_{i=1}^n$ in (a) are transported onto the reference rank vectors $\{U_i\}_{i=1}^n$ in (b).

Note that $\mathbf{R}_n$ verifies $\mathbf{R}_n(S_i) = U_{\sigma_n(i)}$ for $i \in \{1, \ldots, n\}$, where σ_n is the solution of the assignment problem

$$\sigma_n = \underset{\sigma \in P_n}{\operatorname{argmin}} \sum_{i=1}^{n} \|S_i - U_{\sigma(i)}\|^2, \tag{4}$$

for P_n the set of all permutations of $\{1, \cdots, n\}$.

This transport-based rank map echoes the one-dimensional case, with $\{U_i\}_{i=1}^n$ replacing traditional ranks $\{1, 2, \ldots, n\}$ based on univariate quantile levels $\{\frac{1}{n}, \frac{2}{n}, \ldots, 1\}$. By definition, $\|\mathbf{R}_n(s)\| \in \{\frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ for any $s \in \mathbb{R}^d$. This allows to introduce a specific ordering of $\mathbb{R}^d$, namely

$$s_1 \leq_{\mathbf{R}_n} s_2 \text{ if, and only if, } \|\mathbf{R}_n(s_1)\| \leq \|\mathbf{R}_n(s_2)\|.$$

This multivariate ordering is illustrated in Figure 1. A main virtue is its ability to capture the shape of the underlying probability distribution. Namely, it serves as the basis for defining the following quantile region of level $\beta \in [0, 1]$,

$$\widehat{\mathcal{Q}}_n(\beta) = \left\{ s : \|\mathbf{R}_n(s)\| \leq \frac{\lceil \beta n \rceil}{n} \right\}. \tag{5}$$

Another notable advantage of Definition 2.1 is that the ranks are *distribution-free*: by construction, $\|\mathbf{R}_n(S_1)\|, \ldots, \|\mathbf{R}_n(S_n)\|$ correspond to a random permutation of $\{\frac{1}{n}, \frac{2}{n}, \cdots, 1\}$, regardless of the distribution of the non-conformity scores $\{S_i\}_{i=1}^n$. Therefore, the quantile region given by (5) captures a β-proportion of the scores $\{S_i\}_{i=1}^n$ in an appropriate multivariate manner. This is the building block of our CP proposal, enabling the derivation of coverage guarantees through distribution-freeness.

Remark 2.2. The choice of reference rank vectors $\{U_i\}_{i=1}^n$ in Definition 2.1 is flexible and can be tailored to specific needs, provided that $\{S_i\}_{i=1}^n$ and $\{U_i\}_{i=1}^n$ remain independent (Ghosal and Sen, 2022, Remark 3.11). The convention adopted in Definition 2.1 has the merit to fix the ideas and to be appropriate for regression tasks.

2.3 OT-based conformal prediction (OT-CP)

We present a new methodology, OT-CP, that leverages optimal transport to perform (split) conformal prediction with multivariate scores. Unlike traditional CP approaches, our method relies on a multivariate perspective to quantify uncertainties and construct prediction regions through Monge-Kantorovich vector quantiles. Given a confidence level $\alpha \in [0,1]$, the proposed framework consists of three steps:

1. **Multivariate score computation:** Compute the multivariate scores $(S_i)_{i=1}^n = \big(s(X_i,Y_i)\big)_{i=1}^n$ on the calibration set $(X_i,Y_i)_{i=1}^n$,
2. **Quantile region construction:** Construct the quantile region $\widehat{\mathcal{Q}}_n\big((1+\frac{1}{n})\alpha\big)$ as in (5),
3. **Prediction set computation:** For a test input X_{test} following the same distribution as the calibration set, return the prediction region $\widehat{\mathcal{C}}_\alpha(X_{\text{test}}) = \Big\{y \in \mathcal{Y} : s(X_{\text{test}},y) \in \widehat{\mathcal{Q}}_n\Big(\Big(1+\frac{1}{n}\Big)\alpha\Big)\Big\}$.

The key novelty of OT-CP lies in the use of multivariate scores (step 1) along with the construction of an OT-based confidence region (step 2). By definition, this region leverages multivariate quantiles of the empirical distribution $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{S_i}$, accounting for marginal correlations within $(S_1,\dots,S_n)$. Our methodology enables the construction of confidence regions without predefined shapes, thereby aligning better with the underlying data distribution. In step 3, the prediction set for a new input X_{test} is evaluated through the preimage by the score function of the quantile region. This generalizes the one-dimensional case, where classical quantiles are used to construct prediction sets in the form of intervals.

Remark 2.3 (Computational aspects)**.** Our approach requires solving an optimal transport problem between two discrete distributions, each consisting of n points. While the exact solution via linear programming has a computational complexity of $O(n^3)$, efficient approximation methods can reduce it to $O(n^2)$ (Peyré et al., 2019). In the case of univariate scores, this OT problem simplifies to a sorting operation, thus one recovers the standard $O(n\log n)$ cost.

2.4 Coverage guarantees

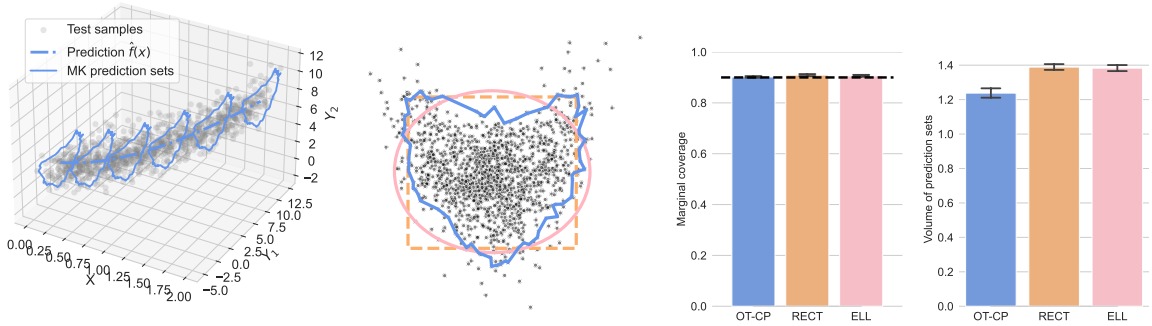
Next, we show that the prediction regions constructed with OT-CP are valid, meaning they satisfy the coverage property.

Theorem 2.4 (Coverage guarantee)**.** *Suppose $(X_i,Y_i)_{i=1}^n \cup (X_{\text{test}},Y_{\text{test}})$ are exchangeable. Let $\alpha \in (0,1)$ such that $\lceil \alpha(n+1)\rceil \le n$. The prediction region $\widehat{\mathcal{C}}_\alpha$ constructed on $(X_i,Y_i)_{i=1}^n$ satisfies*

$$\alpha \;\le\; \mathbb{P}\big(Y_{\text{test}} \in \widehat{\mathcal{C}}_\alpha(X_{\text{test}})\big) \;\le\; \alpha + \frac{2}{n+1}\,, \tag{6}$$

where the probability is taken over the joint distribution of $(X_i,Y_i)_{i=1}^n \cup (X_{\text{test}},Y_{\text{test}})$.

We present two proof strategies for Theorem 2.4 in Appendices A.1 and A.2. While they differ in the way the argumentation is carried out, they are similar in essence. In particular, both approaches extend the quantile lemma (Lei et al., 2018; Vovk et al., 2005), originally established for univariate scores: the rank of a new sample score among calibration scores follows a uniform distribution, ensuring valid prediction regions without distributional assumptions. The key innovation here is that MK quantiles enable this property in a fully multivariate setting, eliminating the need to rely on univariate quantiles. This stems directly from the distribution-freeness of MK quantiles, a



(a) Prediction $x \mapsto \hat{f}(x)$ and prediction sets $\widehat{\mathcal{C}}_\alpha(x)$ (b) Scores (residuals) with different quantile regions (c) Coverage (d) Volumes

Figure 2: Conformal multi-output regression on simulated data

property that has led to numerous applications in rank-based statistical testing (Deb and Sen, 2023; Ghosal and Sen, 2022), which shares connections with CP (Kuchibhotla, 2020).

Building upon this extended quantile lemma, Theorem 2.4 ensures that, for a given coverage level $\alpha \in (0, 1)$, the true label Y_{test} belongs to the OT-based prediction region $\widehat{\mathcal{C}}_\alpha(X_{\text{test}})$ with probability at least α. Moreover, this coverage probability is shown to be of the order of α, being upper-bounded by $\alpha + 2/(n+1)$ with n the size of the calibration set. The factor 2 in this upper bound naturally arises from the use of MK quantiles, which introduces ties in the ranking procedure: there exists $i \in \{1, \ldots, n\}$ such that $\|\mathbf{R}_n(s(X_{\text{test}}, y))\| = \|\mathbf{R}_n(S_i)\| \in \{\frac{1}{n}, \frac{2}{n}, \ldots, 1\}$.

While OT-CP can be applied to any model and score function, the next sections focus on specific settings to clearly demonstrate its advantages over existing CP strategies.

3 Multi-Output Regression

This section examines the application of OT-CP for multi-output regression. First, we demonstrate how this approach accommodates arbitrary score distributions, enabling the creation of diverse and data-tailored prediction region shapes. Next, we introduce an extension of our method, called OT-CP+, which incorporates conditional adaptivity to input covariates. We demonstrate its effectiveness both empirically and theoretically, establishing an asymptotic coverage guarantee for OT-CP+.

3.1 OT-CP can output non-convex prediction sets

For any feature vector $X \in \mathbb{R}^p$ and response vector $Y \in \mathbb{R}^d$, we aim to conformalize the prediction $\hat{f}(X)$ returned by a given black-box regressor.

CP methods for multi-output regression. To further motivate OT-CP in this context, we first review existing conformal strategies. One could consider vanilla CP relying on a univariate aggregated score, $s(x, y) = \|y - \hat{f}(x)\|$. This yields spherical prediction regions $\{\hat{f}(x)\} + \text{Ball}_{\|\cdot\|}(\tau_\alpha)$[1] where $\text{Ball}_{\|\cdot\|}(\tau_\alpha)$ is the Euclidean ball of radius $\tau_\alpha > 0$. One can also treat the d components of $Y \in \mathbb{R}^d$ separately to produce prediction regions based on hyperrectangles, $\{\hat{f}(x)\} + \prod_{i=1}^d [a_i, b_i]$

[1]The expression involves the Minkowski sum between two sets: for two sets A and B, $A+B = \{a+b,\, a \in A,\, b \in B\}$.

(Neeven and Smirnov, 2018). However, these approaches are often ill-suited to accurately capture the geometry of multivariate distributions. In particular, the output prediction sets (whether spherical or hyperrectangles) can be too large when handling anisotropic uncertainty that varies across different output dimensions. To mitigate this, prior works have introduced scores that account for anisotropy and correlations among the residual dimensions, as with ellipsoidal prediction sets (Messoudi et al., 2020; Johnstone and Cox, 2021; Henderson et al., 2024). Still, this implicitly assumes an elliptical distribution for the non-conformity score, thereby compromising the distribution-free nature of the method.

OT-CP for multi-output regression. Our strategy consists in applying OT-CP with a multivariate residual as the score,

$$s(x, y) = y - \hat{f}(x) \in \mathbb{R}^d \,, \tag{7}$$

and yields the following prediction regions

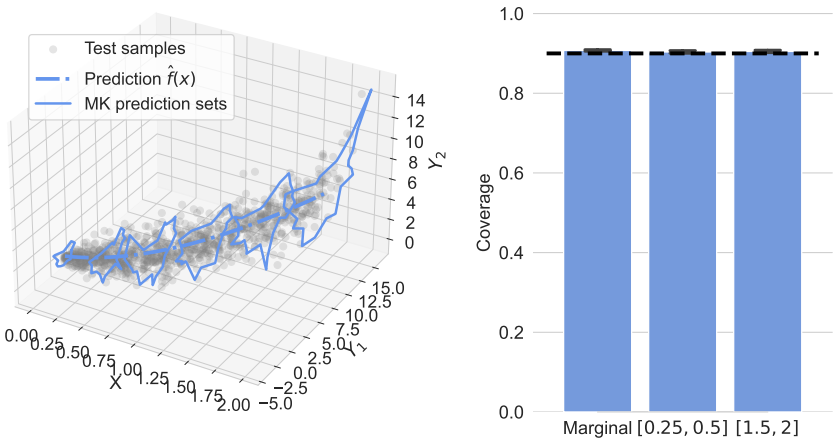
$$\forall x \in \mathbb{R}^p, \quad \widehat{\mathcal{C}}_\alpha(x) = \{\hat{f}(x)\} + \widehat{\mathcal{Q}}_n\Big(\Big(1 + \frac{1}{n}\Big)\alpha\Big) \,. \tag{8}$$

These sets can take on flexible, arbitrary shapes, that adapt to the calibration error distribution and the underlying data geometry. This key advantage is illustrated concretely in our numerical experiments below.

Numerical experiments. In what follows, we study a practical regression problem and compare several CP methods described above: OT-CP for forming prediction regions as in (8), a CP approach producing ellipses (`ELL`, Johnstone and Cox, 2021), and a simple method creating hyperrectangle (`REC`, Neeven and Smirnov, 2018), with the miscoverage level adjusted by the Bonferroni correction. We simulate univariate inputs $X \sim \mathrm{Unif}([0, 2])$ with responses $Y \in \mathbb{R}^2$, and we assume that we are given a pre-trained predictor $\hat{f}(x) = (2x^2, (x + 1)^2)$, $x \in \mathbb{R}$. We interpret the score $s(X, Y) = Y - \hat{f}(X)$ as a random vector ζ distributed from a mixture of Gaussians and independent of X, meaning that the distribution of $s(X, Y)$ remains unchanged when conditioned on X. Quantile regions for $\alpha = 0.9$ are constructed using $n = 1000$ calibration instances. More implementation details can be found in Appendix B. As expected, OT-CP prediction regions exhibit superior adaptability to the distribution of residuals, whereas hyperrectangles and ellipses tend to be overly conservative (Figures 2(a) and 2(b)). We also compare the methods in terms of empirical coverage on test data (Figure 2(c)) and efficiency (volume of prediction regions, Figure 2(d)). While all approaches adhere to the α-coverage guarantee OT-CP achieves greater efficiency, producing smaller and more precise prediction regions. This highlights that MK quantiles help effectively address uncertainty quantification challenges for multi-output regression.

3.2 OT-CP+: an adaptive version

So far, the form of the constructed prediction regions (8) does not depend on the input X, as illustrated in Figure 2(a). This uniformity stems from computing quantile regions over the distribution of scores $(S_i)_{i=1}^n$ *marginalized* over $(X_i, Y_i)_{i=1}^n$. In other words, $(S_i)_{i=1}^n$ are treated as i.i.d. realizations of $S = Y - \hat{f}(X)$. As a result, while the quantile regions provided by OT-CP effectively capture the global geometry of the scores, they do not adapt to variations in X. This lack of adaptivity is inadequate in applications where prediction uncertainties vary between input examples, as discussed by Chernozhukov et al. (2021); Foygel Barber et al. (2020).



(a) Adaptive prediction regions (b) Conditional coverage

Figure 3: Adaptive conformal regression through quantile regression on conditional scores $s(X, Y)|X = x$

Methodology. To account for input-dependent uncertainty in the predictions, we introduce OT-CP+, a conformal procedure that computes *adaptive* MK quantile region by leveraging multiple-output quantile regression del Barrio et al. (2024). Given a test point $X_{\text{test}} = x$ and a coverage level $\alpha \in (0, 1)$, OT-CP+ selects the k-nearest neighbors of x in the calibration set and solves an OT problem between the k associated scores and reference vectors $(U_i)_{i=1}^k$ accordingly to Definition 2.1. This gives rise to the *conditional* empirical MK rank map, $\mathbf{R}_k(\cdot|x)$ as defined in del Barrio et al. (2024). Similarly to (5), the quantile region based on $\mathbf{R}_k(\cdot|x)$ is

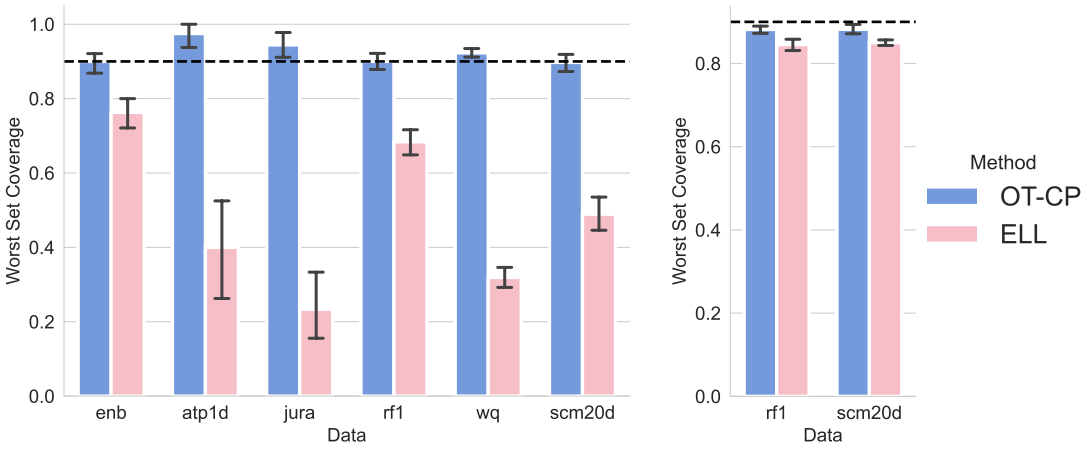
$$\widehat{\mathcal{Q}}_k\Big(\Big(1+\frac{1}{k}\Big)\alpha|x\Big) = \Big\{s : \|\mathbf{R}_k(s|x)\| \leq \frac{\lceil (k+1)\alpha \rceil}{k}\Big\}.$$

Hence, OT-CP+ follows the same procedure as OT-CP but operates on the distribution of $s(X, Y)$ given X, which is approximated on a neighborhood of X. The prediction regions returned by OT-CP+ are thus given by

$$\forall x \in \mathbb{R}^p,\ \widehat{C}_{\alpha,k}(x) = \{\hat{f}(x)\} + \widehat{\mathcal{Q}}_k\Big(\Big(1+\frac{1}{k}\Big)\alpha|x\Big). \tag{9}$$

Experiments on simulated data. We first consider a similar setting as that of Section 3.1, where the score $s(X, Y)$ is now distributed as $\sqrt{X}\zeta$. Consequently, the variance of the residual increases with X, which suggests that wider quantiles should be constructed for larger values of X. Figure 3 confirms that OT-CP+ effectively constructs adaptive prediction regions with the desired α-coverage. To quantify this more precisely, we evaluate the empirical coverage conditionally on X: Figure 3(b) reports box plots of $\mathbb{P}(Y_{\text{test}} \in \widehat{\mathcal{C}}_\alpha(X_{\text{test}})|X_{\text{test}} \in \mathcal{I})$ for several choices of subsets $\mathcal{I} \subset [0, 2]$. Our results show that OT-CP+ satisfies the conditional coverage guarantee.

Experiments on real data. Next, we evaluate OT-CP+ on real datasets sourced from Mulan (Tsoumakas et al., 2011). We also implement a concurrent CP method (Messoudi et al., 2022), that is an adaptive extension of the previous ellipsoidal approach (Johnstone and Cox, 2021). Specifically,



(a) Low data regime (≤ 1500 total samples)

(b) Medium data regime (9000 total samples)

Figure 4: Conditional coverage on real datasets of two conformal procedures for multi-output regression

Messoudi et al. (2022) construct ellipsoidal prediction sets that account for local geometry, by estimating the covariance of $Y|X$ with the k-nearest neighbors (kNN) of X.

We split each dataset into training, calibration, and testing subsets (50%–25%–25% ratio) and train a random forest model as the regressor. Both methods use a kNN step that selects 10% of the calibration set as neighbors for each test point X_{test}. As a coverage metric, we consider the *worst-set coverage*, $\min_{j\in\{1,\dots,J\}} \mathbb{P}(Y_{\text{test}} \in \widehat{\mathcal{C}}_\alpha(X_{\text{test}})|X_{\text{test}} \in \mathcal{A}_j)$, with $\{\mathcal{A}_j\}_{j\in\{1,\dots,J\}}$ a partition of the input space tailored to the test data. This metric is conceptually similar to the *worst-slab coverage* (Cauchois et al., 2021), which considers specific partitions in the form of slabs. In our approach, we obtain $J = 5$ regions $\{\mathcal{A}_j\}_{j\in\{1,\dots,5\}}$ by clustering, *i.e.,* employing (i) a random selection of centroids, and (ii) a kNN procedure ensuring that each region contains 10% of the test samples. Empirical results presented in Figure 4 provide evidence supporting the approximate conditional coverage achieved by OT-CP+. Indeed, the worst-set coverage of OT-CP+ remains consistently close to the target level $\alpha = 0.9$ across all datasets, regardless of the sample size. This contrasts with the adaptive ellipsoidal approach, which does not achieve such α-coverage and exhibits greater variability.

Asymptotic conditional coverage. In the one-dimensional case ($d = 1$), Lei et al. (2018) established the inherent limitation of achieving exact distribution-free conditional coverage in finite samples. However, asymptotic conditional coverage remains attainable under regularity assumptions (Lei et al., 2018; Chernozhukov et al., 2021). OT-CP+ benefits from such a guarantee, leveraging asymptotic properties of quantile regression for MK quantiles (del Barrio et al., 2024). The following assumption is needed.

Assumption 3.1. *Suppose that* $(X_1, Y_1), \dots, (X_n, Y_n)$, $(X_{\text{test}}, Y_{\text{test}})$ *are i.i.d. Assume that for almost every* x, *the distribution* $\mathbb{P}_{S|X=x}$ *of* $s(X_{\text{test}}, Y_{\text{test}})$ *given* $X_{\text{test}} = x$ *is Lebesgue-absolutely continuous on its convex support* $\text{Supp}(\mathbb{P}_{S|X=x})$. *For any* $R > 0$, *suppose that its density* $p(\cdot|x)$ *verifies for all* $s \in \text{Supp}(\mathbb{P}_{S|X=x}) \cap \text{Ball}_{\|\cdot\|}(R)$, $\lambda_R^x \leq p(s|x) \leq \Lambda_R^x$.

Theorem 3.2. *Let k be the number of nearest neighbors used to estimate $\mathbf{R}_k(\cdot|x)$. Assume that $k \to +\infty$ as $n \to +\infty$ and $k/n \to 0$. Under Assumption 3.1, the following holds for any $\alpha \in [0,1]$,*

$$\lim_{n,k\to+\infty} \mathbb{P}\big(Y_{\text{test}} \in \widehat{C}_{\alpha,k}(X_{\text{test}}) \big| X_{\text{test}}\big) = \alpha\,. \tag{10}$$

where $\widehat{C}_{\alpha,k}(x) = \{\hat{f}(x)\} + \widehat{\mathcal{Q}}_k\Big(\big(1+\frac{1}{k}\big)\alpha|x\Big)$ depends on $\hat{f}$ previously learned on (fixed) training data.

4 Classification

In this section, we apply OT-CP to multiclass classification. Each data point consists of a feature-label pair $(X,Y) \in \mathbb{R}^p \times \{1,\dots,K\}$, with $K \geq 3$ the number of classes. The given black-box classifier outputs, for any input $X \in \mathbb{R}^p$, a vector $\hat{\pi}(X)$ of estimated class probabilities, where the k-th component $\hat{\pi}_k(X)$ is the probability estimate that X belongs to class k (hence, $\sum_{k=1}^K \hat{\pi}_k(X) = 1$).

CP methods for classification. Commonly used scores for multiclass classification include the Inverse Probability (IP), $s(x,y) = 1 - \hat{\pi}_y(x)$ and the Margin Score (MS), $s(x,y) = \max_{y'\neq y} \hat{\pi}_{y'}(x) - \hat{\pi}_y(x)$ (Johansson et al., 2017). IP only considers the probability estimate for the correct class label ($\hat{\pi}_y(x)$), whereas MS also involves the most likely incorrect class label ($\max_{y'\neq y} \hat{\pi}_{y'}(x)$). More adaptive options argue in favor of incorporating more class labels in the score function (Romano et al., 2020; Angelopoulos et al., 2021; Melki et al., 2024). The idea is to rank the labels from highest to lowest confidence (by sorting the probability estimates as $\hat{\pi}_{(y_1)}(x) \geq \hat{\pi}_{(y_2)}(x) \geq \cdots \geq \hat{\pi}_{(y_K)}(x)$), then return the labels such that the total confidence (*i.e.,* the cumulative sum) is at least α. It is worth noting that this strategy stems from a notion of *generalized conditional quantile function* (Romano et al., 2020), by analogy with $\inf_{c\in\mathbb{R}}\{\mathbb{P}(Y \leq c|X=x) \geq \alpha\}$.

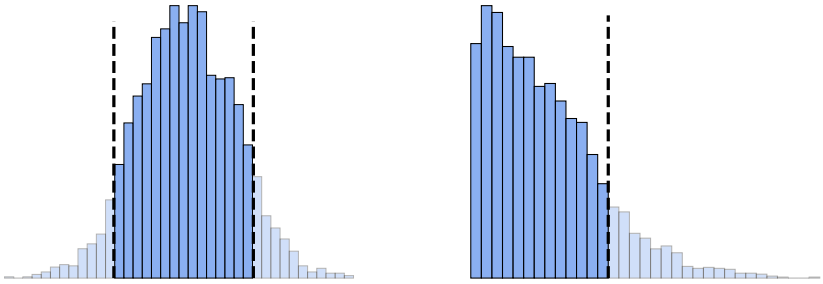
OT-CP for multiclass classification. As an alternative CP method for this problem, we propose using OT-CP with the following multivariate score,

$$s(X,Y) = |\bar{Y} - \hat{\pi}(X)| \in \mathbb{R}^K_+\,, \tag{11}$$

where the absolute value is taken component-wise and $\bar{Y} = (\mathbb{1}_{Y=k})_{k=1}^K$ denotes the one-hot encoding of Y. One can remark in passing that $\|s(x,y)\|_1 = 2(1-\hat{\pi}_y(x))$, which corresponds to the aforementioned IP scalar score. Our OT-CP procedure builds upon generalized quantiles to take into account *all* the components of $\hat{\pi}_y(x)$ (and not only the largest values) and to capture the correlations between them.

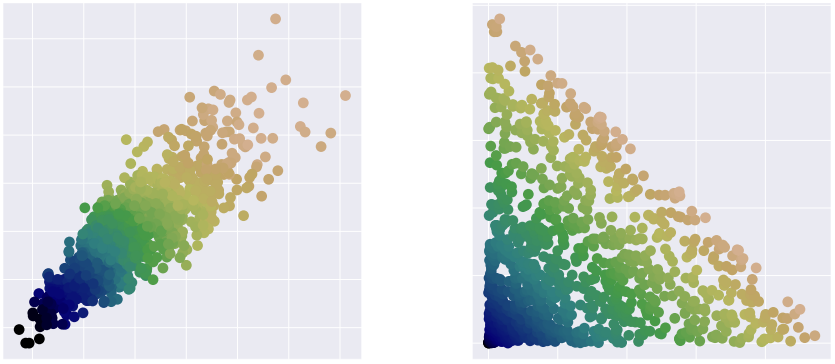
The score in (11) takes values in $\mathbb{R}^K_+$ and naturally induces a *left-to-right* ordering. This contrasts with the score function used in our previous application, multi-output regression, where the ordering is center-outward. To further clarify this difference, let us focus on a single component of the score, $s(x,y)_k$, for simplicity. A center-outward interval of the form of $[q_{\alpha/2}, q_{1-\alpha/2}]$ applied to $s(x,y)_k$ excludes lower values from $[0, q_{\alpha/2})$ (Figure 5(a)). This exclusion is problematic for the score structure induced by (11), since lower values of $s(x,y)_k$ indicate greater conformity between x and the ground-truth y. In this context, left-to-right ordering is more appropriate, as illustrated in Figure 5(b).

A left-to-right ordering can be easily achieved by making a slight adjustment to Definition 2.1: we choose the reference rank vectors as $U_i = \frac{i}{n}\theta_i^+$, where θ_i^+ is uniformly sampled in $\{\theta \in \mathbb{R}^d_+ : \|\theta\|_1 = 1\}$. As depicted in Figure 6, the resulting MK ranks reflect the desired left-to-right ordering. It is worth noting that this adjustment is fully compatible with the general definition of MK quantiles, which is flexible enough to accommodate arbitrary reference distributions (see Remark 2.2).



(a) $s_i = y_i - \hat{\pi}(x_i)$ (b) $s_i^+ = |y_i - \hat{\pi}(x_i)|$

Figure 5: Ordering must depend on the chosen scores: (a) Center-outward for signed errors, (b) Left-to-right for absolute errors



(a) Multivariate scores $\{S_i\}_{i=1}^n$ corresponding to absolute errors in $\mathbb{R}_+^d$ (b) Positive reference rank vectors $\{U_i\}_{i=1}^n$

Figure 6: Positive reference ranks for a left-to-right ordering. The colormap encodes how the 2-dimensional scores $\{S_i\}_{i=1}^n$ in (a) are transported onto the reference rank vectors $\{U_i\}_{i=1}^n$ in (b)

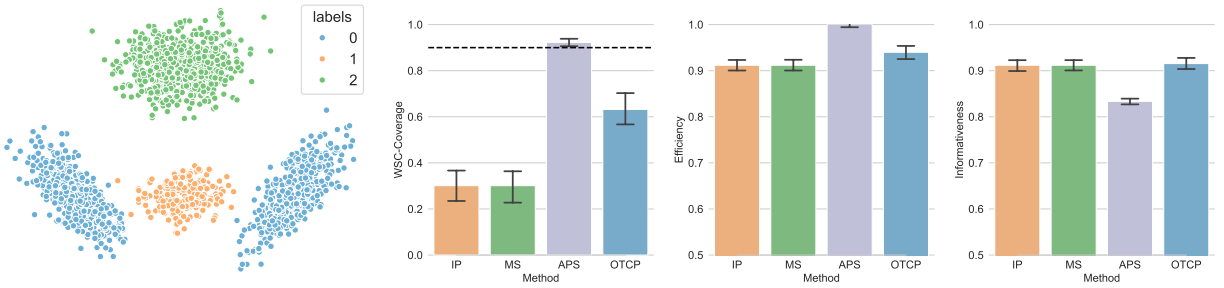
Based on this choice of score (11) and reference rank vectors, OT-CP generates the following prediction sets,

$$\widehat{\mathcal{C}}_\alpha(x) = \left\{ \bar{y} \in \{0,1\}^K : |\bar{y} - \hat{\pi}(x)| \in \widehat{\mathcal{Q}}_n\left(\left(1 + \frac{1}{n}\right)\alpha\right) \right\},$$

where the region $\widehat{\mathcal{Q}}_n(\cdot)$ is constructed from $\{U_i\} = \{\frac{i}{n}\theta_i^+\}$.

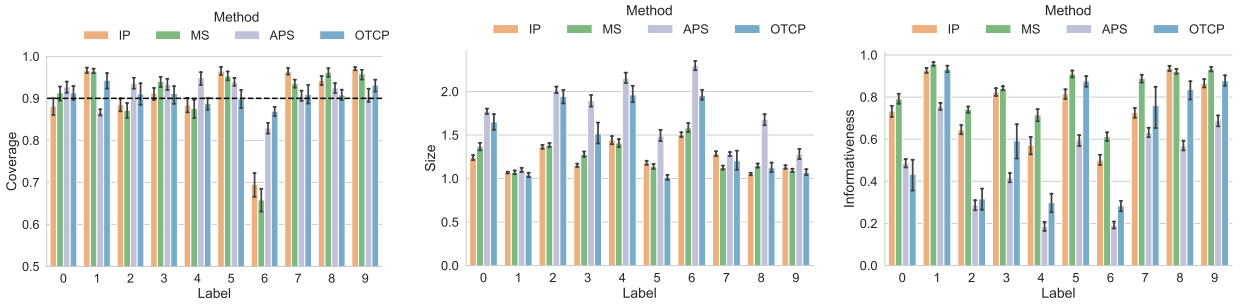
Numerical experiments. We compare OT-CP against IP, MS and APS scores in terms of worst-case coverage (WSC, measuring conditional coverage, as proposed in Romano et al. (2020)), efficiency (average size of the predicted set) and informativeness (average number of predicted singletons). More implementation details are given in Appendix B.

We start by simulating data according to a Gaussian mixture model, represented in Figure 7(a) and we consider a pre-trained classifier based on Quadratic Discriminant Analysis. Figures 7(b) to 7(d) outline that OT-CP successfully retains the efficiency and informativeness—hallmarks of IP and MS—while simultaneously enhancing conditional coverage on X, akin to the improvements achieved by APS. These results highlight that OT-CP effectively handles arbitrary probability profiles by leveraging the entire softmax output, rather than relying solely on its sum, to construct more informative and meaningful prediction sets.



(a) Simulated data (b) WSC coverage (c) Average size (d) Informativeness

Figure 7: Conformal classification by Quadratic Discriminant Analysis on simulated data



(a) Coverage (b) Average size (c) Informativeness

Figure 8: Label-wise results on $K = 10$ classes of Fashion-MNIST

The relevance of OT-CP is also confirmed on real datasets. In Figure 8, we present the results for Fashion-MNIST for a random forest. Additional numerical experiments on MNIST and CIFAR-10 datasets are provided in Appendix B. Interestingly, despite not being explicitly designed for this purpose, OT-CP achieves conditional coverage with respect to the label on par with APS, where IP and MS fall short. In addition, OT-CP maintains the efficiency and informativeness of IP and MS, offering a convenient balance across all the considered metrics. We finally emphasize that the numerical experiments were designed as prototypes to demonstrate how OT-CP can be seamlessly and effectively adapted to typical classification tasks. The focus is on demonstrating a useful application of our general framework, which already shows several benefits while remaining conceptually simple.

5 Conclusion and perspectives

We have introduced a general and versatile framework for conformal prediction grounded in optimal transport theory. This approach not only revisits classical CP methods based on scalar scores, but also extends easily to handle multivariate scores in a novel and robust manner, thanks to the inherent properties of Monge-Kantorovich quantiles. The OT-CP methodology is flexible, enabling the construction of prediction regions tailored to diverse scenarios, besides being well-suited to capture complex uncertainty structures.

This methodology can pave the way for future developments, with potential adaptations to new learning tasks. One might think of multi-label classification where the multivariate score (11) immediately applies by replacing the one-hot encoding by a multi-hot encoding, see *e.g.*, Katsios

and Papadopoulos (2024) for related ellipsoidal inference.

Future work could explore the development of more sophisticated multivariate scores, potentially building on existing univariate alternatives, see, *e.g.,* (Tumu et al., 2024; Wang et al., 2023; Plassier et al., 2024) for regression, and Angelopoulos et al. (2021); Melki et al. (2024) for classification. Indeed, our numerical experiments demonstrate that basic multivariate scores can outperform classical univariate counterparts, providing a strong foundation and motivation for pursuing into this direction.

References

Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591.

Angelopoulos, A. N., Bates, S., Jordan, M., and Malik, J. (2021). Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*.

Cauchois, M., Gupta, S., and Duchi, J. C. (2021). Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22(81):1–42.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223 – 256.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118.

Deb, N. and Sen, B. (2023). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, 118(541):192–207.

del Barrio, E., Sanz, A. G., and Hallin, M. (2024). Nonparametric multiple-output center-outward quantile regression. *Journal of the American Statistical Association*, pages 1–15.

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

Foygel Barber, R., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2020). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482.

Ghosal, P. and Sen, B. (2022). Multivariate ranks and quantiles using optimal transport: Consistency, rates, and nonparametric testing. *The Annals of Statistics*, 50(2):1012–1037.

Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139 – 1165.

Henderson, I., Mazoyer, A., and Gamboa, F. (2024). Adaptive inference with random ellipsoids through conformal conditional linear expectation. *arXiv preprint arXiv:2409.18508*.

Johansson, U., Linusson, H., Löfström, T., and Boström, H. (2017). Model-agnostic nonconformity functions for conformal classification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2072–2079. IEEE.

Johnstone, C. and Cox, B. (2021). Conformal uncertainty sets for robust optimization. In *Conformal and Probabilistic Prediction and Applications*, pages 72–90. PMLR.

Katsios, K. and Papadopoulos, H. (2024). Multi-label conformal prediction with a mahalanobis distance nonconformity measure. *Proceedings of Machine Learning Research*, 230:1–14.

Kuchibhotla, A. K. (2020). Exchangeability, conformal prediction, and rank tests. *arXiv preprint arXiv:2005.06095*.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Melki, P., Bombrun, L., Diallo, B., Dias, J., and Da Costa, J.-P. (2024). The penalized inverse probability measure for conformal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3512–3521.

Messoudi, S., Destercke, S., and Rousseau, S. (2020). Conformal multi-target regression using neural networks. In *Conformal and Probabilistic Prediction and Applications*, pages 65–83. PMLR.

Messoudi, S., Destercke, S., and Rousseau, S. (2021). Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101.

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

Neeven, J. and Smirnov, E. (2018). Conformal stacked weather forecasting. In *Conformal and Probabilistic Prediction and Applications*, pages 220–233. PMLR.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Plassier, V., Fishkov, A., Guizani, M., Panov, M., and Moulines, E. (2024). Probabilistic conformal prediction with approximate conditional validity. *arXiv preprint arXiv:2407.01794*.

Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. *Advances in neural information processing systems*, 32.

Romano, Y., Sesia, M., and Candes, E. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591.

Sesia, M. and Romano, Y. (2021). Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315.

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., and Vlahavas, I. (2011). Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research*, 12:2411–2414.

Tumu, R., Cleaveland, M., Mangharam, R., Pappas, G., and Lindemann, L. (2024). Multi-modal conformal prediction regions by optimizing convex shape templates. In *6th Annual Learning for Dynamics & Control Conference*, pages 1343–1356. PMLR.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Wang, Z., Gao, R., Yin, M., Zhou, M., and Blei, D. (2023). Probabilistic conformal prediction using conditional random samples. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8814–8836. PMLR.

A Proofs

A.1 Proof of Theorem 2.4 (marginal coverage guarantee)

The proof of Theorem 2.4 consists in extending the reasoning for the traditional quantile lemma (*e.g.*, Lemma 2 in Romano et al. (2019)) to a multivariate setting. In particular, we leverage the distribution-freeness of Monge-Kantorovich ranks and the associated multivariate ordering $\leq_{\mathbf{R}_n}$ in $\mathbb{R}^d$.

Proof of Theorem 2.4. We begin with a rewriting of the quantile region $\widehat{\mathcal{Q}}_n(\alpha)$ in (5). Define the order statistics $\{S_{(i,n)}\}_{i=1}^n$ by relabeling such that

$$S_{(1,n)} \leq_{\mathbf{R}_n} S_{(2,n)} \leq_{\mathbf{R}_n} \cdots \leq_{\mathbf{R}_n} S_{(n,n)}. \tag{12}$$

By definition of $\leq_{\mathbf{R}_n}$ and by construction of our grid $\{U_i\}_{i=1}^n$, it is easy to see that $\mathbf{R}_n(S_{(i,n)}) = U_i$ for every i. Consequently,

$$\begin{aligned} s \leq_{\mathbf{R}_n} S_{(\lceil \alpha n \rceil, n)} \quad &\iff \quad \|\mathbf{R}_n(s)\| \leq \|U_{\lceil \alpha n \rceil}\|, \\ &\iff \quad \|\mathbf{R}_n(s)\| \leq \frac{\lceil \alpha n \rceil}{n} \\ &\iff \quad s \in \widehat{\mathcal{Q}}_n(\alpha). \end{aligned}$$

Thus, the quantile region $\widehat{\mathcal{Q}}_n(\alpha)$ can be rewritten with

$$S_{\text{test}} \in \widehat{\mathcal{Q}}_n(\alpha) \quad \iff \quad S_{\text{test}} \leq_{\mathbf{R}_n} S_{(\lceil \alpha n \rceil, n)}. \tag{13}$$

Now, denote by $i_0 \in \{1, \ldots, n\}$ the index such that $\mathbf{R}_n(S_{\text{test}}) = U_{i_0}$, so that S_{test} and S_{i_0} have the same multivariate rank, *i.e.*, $S_{(i_0)} =_{\mathbf{R}_n} S_{\text{test}}$. Denote by $S_{(k,n+1)}$ the k-th smallest value (with respect to $\leq_{\mathbf{R}_n}$) within $S_1, \cdots, S_n, S_{\text{test}}$. The terms $\{S_{(k,n+1)}\}_{k=1}^{n+1}$ correspond to

$$S_{(1,n)} \leq_{\mathbf{R}_n} \cdots \leq_{\mathbf{R}_n} S_{(i_0,n)} \leq_{\mathbf{R}_n} S_{\text{test}} \leq_{\mathbf{R}_n} S_{(i_0+1,n)} \leq_{\mathbf{R}_n} \cdots \leq_{\mathbf{R}_n} S_{(n,n)}. \tag{14}$$

Here, we make the arbitrary choice $S_{(i_0,n+1)} = S_{(i_0,n)}$ and $S_{(i_0+1,n+1)} = S_{\text{test}}$. Note that the opposite choice would not change the incoming arguments. As a direct byproduct of (14), one can see that

- for $k < i_0$, $S_{(k,n+1)} = S_{(k,n)}$,
- for $k \in \{i_0, i_0 + 1\}$, $S_{(k,n+1)}$ equals either S_{test} or $S_{(i_0,n)}$. In both cases, $S_{(k,n+1)} \leq_{\mathbf{R}_n} S_{(k,n)}$.
- for $k > i + 1$, $S_{(k,n+1)} = S_{(k-1,n)} \leq_{\mathbf{R}_n} S_{(k,n)}$.

Thus, $\forall k \in \{1, \ldots, n\}$, $S_{(k,n+1)} \leq_{\mathbf{R}_n} S_{(k,n)}$. Hence, if $S_{\text{test}} \leq_{\mathbf{R}_n} S_{(k,n+1)}$ then $S_{\text{test}} \leq_{\mathbf{R}_n} S_{(k,n)}$. We claim that the reciprocal also holds. Assume that $S_{\text{test}} \leq_{\mathbf{R}_n} S_{(k,n)}$. Regarding (14), it must be that $k \geq i_0 + 1$, in which case $S_{(k,n+1)}$ is the greater of S_{test} and $S_{(k-1,n)}$ (with respect to $\leq_{\mathbf{R}_n}$). Putting everything together, we showed that

$$S_{\text{test}} \leq_{\mathbf{R}_n} S_{(k,n)} \quad \iff \quad S_{\text{test}} \leq_{\mathbf{R}_n} S_{(k,n+1)}.$$

Thus, $\mathbb{P}\big(S_{\text{test}} \leq_{\mathbf{R}_n} S_{(k,n)}\big) = \mathbb{P}\big(S_{\text{test}} \leq_{\mathbf{R}_n} S_{(k,n+1)}\big)$. Hence, for any $\beta \in [0, 1]$,

$$\mathbb{P}\big(S_{\text{test}} \in \widehat{\mathcal{Q}}_n(\beta)\big) = \mathbb{P}\big(S_{\text{test}} \leq_{\mathbf{R}_n} S_{(\lceil \beta n \rceil, n+1)}\big). \tag{15}$$

By definition of $S_{(k,n+1)}$, the proportion of elements from $\{S_1, \cdots, S_n, S_{\text{test}}\}$ that is lower than $S_{(k,n+1)}$ (with respect to $\leq_{\mathbf{R}_n}$) is either $k/(n+1)$ or $(k+1)/(n+1)$, due to the tie $S_{(i_0+1,n+1)} = S_{\text{test}}$. Combining the above with (15) implies the following, for any $\beta \in [0,1]$,

$$\frac{\lceil \beta n \rceil}{n+1} \leq \mathbb{P}\big(S_{\text{test}} \in \widehat{\mathcal{Q}}_n(\beta)\big) \leq \frac{\lceil \beta n \rceil}{n+1} + \frac{1}{n+1}.$$

Finally, taking $\beta = (1 + \frac{1}{n})\alpha = \frac{n+1}{n}\alpha$ induces the result, as

$$\alpha \leq \frac{\lceil \beta n \rceil}{n+1} \leq \alpha + \frac{1}{n+1}.$$

□

A.2 Alternative proof of Theorem 2.4

We provide an alternative proof of Theorem 2.4, which is similar in essence to the previous one but based on another perspective. To this end, we recall a variant of the traditional quantile lemma adapted to our needs (*i.e.,* when there are ties in the ranks) and detail its proof for completeness.

Lemma A.1. *(Quantile lemma, Lei et al., 2018) Suppose $U_1, \ldots, U_{n+1}$ are exchangeable random variables in $\mathbb{R}$. Then, for any $\beta \in (0,1)$,*

$$\mathbb{P}(U_{n+1} \leq U_{(\lceil \beta(n+1) \rceil)}) \geq \beta \tag{16}$$

Additionally, suppose there exists $k \in \{1, \ldots, n+1\}$ such that $(U_1, \ldots, U_{n+1}) \backslash \{U_k\}$ are almost surely distinct, and $(U_1, \ldots, U_{n+1})$ are not (i.e., there exists $i \in \{1, \ldots, n+1\}$ with $i \neq k$ such that $U_i = U_k$ and $U_i \neq U_j$ for $j \in \{1, \ldots, n+1\}$, $j \neq i$). Then,

$$\mathbb{P}(U_{n+1} \leq U_{(\lceil \beta(n+1) \rceil)}) \leq \beta + \frac{2}{n+1} \tag{17}$$

The probabilities are taken over the joint distribution of $(U_1, \ldots, U_{n+1})$.

Proof. By exchangeability of $U_1, \ldots, U_{n+1}$, for any $i \in \{1, \ldots, n+1\}$,

$$\mathbb{P}(U_{n+1} \leq U_{(\lceil \beta(n+1) \rceil)}) = \mathbb{P}(U_i \leq U_{(\lceil \beta(n+1) \rceil)})\,. \tag{18}$$

Therefore,

$$\begin{aligned}
\mathbb{P}(U_{n+1} \leq U_{(\lceil \beta(n+1) \rceil)}) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}(U_i \leq U_{(\lceil \beta(n+1) \rceil)}) && (19)\\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}_{U_i \leq U_{(\lceil \beta(n+1) \rceil)}} \right] && (20)\\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}_{U_i < U_{(\lceil \beta(n+1) \rceil)}} + \mathbb{1}_{U_i = U_{(\lceil \beta(n+1) \rceil)}} \right] && (21)
\end{aligned}$$

Since $U_{(\lceil \beta(n+1) \rceil)}$ is the $\lceil \beta(n+1) \rceil$-th smallest value of $(U_1, \ldots, U_{n+1})$, then $\sum_{i=1}^{n+1} \mathbb{1}_{U_i \leq U_{(\lceil \beta(n+1) \rceil)}} \geq \lceil \beta(n+1) \rceil$, and by (20),

$$\begin{aligned}
\mathbb{P}(U_{n+1} \leq U_{(\lceil \beta(n+1) \rceil)}) &\geq \frac{1}{n+1} \mathbb{E}\left[\lceil \beta(n+1) \rceil \right] && (22)\\
&\geq \frac{\lceil \beta(n+1) \rceil}{n+1} \geq \beta\,. && (23)
\end{aligned}$$

Additionally, based on our assumption on the distinctness of $(U_1, \dots, U_{n+1})$, we distinguish three cases.

- If $U_k < U_{(\lceil \beta(n+1) \rceil)}$, then $\sum_{i=1}^{n+1} \mathbb{1}_{U_i < U_{(\lceil \beta(n+1) \rceil)}} = \lceil \beta(n+1) \rceil$ and $\sum_{i=1}^{n+1} \mathbb{1}_{U_i = U_{(\lceil \beta(n+1) \rceil)}} = 1$.
- If $U_k = U_{(\lceil \beta(n+1) \rceil)}$, then $\sum_{i=1}^{n+1} \mathbb{1}_{U_i < U_{(\lceil \beta(n+1) \rceil)}} = \lceil \beta(n+1) \rceil - 1$ and $\sum_{i=1}^{n+1} \mathbb{1}_{U_i = U_{(\lceil \beta(n+1) \rceil)}} = 2$.
- If $U_k > U_{(\lceil \beta(n+1) \rceil)}$, then $\sum_{i=1}^{n+1} \mathbb{1}_{U_i < U_{(\lceil \beta(n+1) \rceil)}} = \lceil \beta(n+1) \rceil - 1$ and $\sum_{i=1}^{n+1} \mathbb{1}_{U_i = U_{(\lceil \beta(n+1) \rceil)}} = 1$.

By considering all these cases in (21), we can conclude that

$$\mathbb{P}(U_{n+1} \leq U_{(\lceil \beta(n+1) \rceil)}) \leq \beta + \frac{2}{n+1}\,. \tag{24}$$

□

By using Lemma A.1 along with the properties of Monge-Kantorovich rank maps, we can prove Theorem 2.4 as follows.

Alternative proof of Theorem 2.4. By construction of the prediction region, we have

$$\begin{aligned}\left\{Y_{\text{test}} \in \widehat{\mathcal{C}}_\alpha(X_{\text{test}})\right\} &= \left\{S_{\text{test}} \in \widehat{\mathcal{Q}}_n\left((1+\frac{1}{n})\alpha\right)\right\} \\ &= \left\{\|\mathbf{R}_n(S_{\text{test}})\| \leq \frac{\lceil (n+1)\alpha \rceil}{n}\right\}.\end{aligned}$$

Therefore,

$$\mathbb{P}(Y_{\text{test}} \in \widehat{\mathcal{C}}_\alpha(X_{\text{test}})) = \mathbb{P}\left(\|\mathbf{R}_n(S_{\text{test}})\| \leq \frac{\lceil (n+1)\alpha \rceil}{n}\right). \tag{25}$$

For any $m \in \mathbb{N}^*$ and $k \in \{1, \dots, m\}$, denote by $S_{(k,m)}$ the k-th smallest value in $(S_1, \dots, S_m)$ according to our ordering $\leq_{\mathbf{R}_n}$ (4), *i.e.*,

$$\|\mathbf{R}_n(S_{(1,m)})\| \leq \|\mathbf{R}_n(S_{(2,m)})\| \leq \cdots \leq \|\mathbf{R}_n(S_{(m,m)})\|\,.$$

We know that for any $k \in \{1, \dots, n\}$, $\|\mathbf{R}_n(S_{\text{test}})\| \leq \|\mathbf{R}_n(S_{(k,n)})\|$ if, and only if, $\|\mathbf{R}_n(S_{\text{test}})\| \leq \|\mathbf{R}_n(S_{(k,n+1)})\|$ (*e.g.*, see the proof of Lemma 2 in Romano et al. (2019)).

By definition of our rank vectors, $\frac{\lceil (n+1)\alpha \rceil}{n} = \|\mathbf{R}_n(S_{(\lceil \alpha(n+1) \rceil, n)})\|$. Then, $\|\mathbf{R}_n(S_{\text{test}})\| \leq \frac{\lceil (n+1)\alpha \rceil}{n}$ if and only if $\|\mathbf{R}_n(S_{\text{test}})\| \leq \|\mathbf{R}_n(S_{(\lceil \alpha(n+1) \rceil, n+1)})\|$. By (25), we deduce that

$$\mathbb{P}(Y_{\text{test}} \in \widehat{\mathcal{C}}_\alpha(X_{\text{test}})) = \mathbb{P}(\|\mathbf{R}_n(S_{\text{test}})\| \leq \|\mathbf{R}_n(S_{(\lceil \alpha(n+1) \rceil, n+1)})\|)\,. \tag{26}$$

We conclude by applying the quantile lemma to $(\|\mathbf{R}_n(S_i)\|)_{i=1}^n \cup \{\|\mathbf{R}_n(S_{\text{test}})\|\}$, which is formally stated and proved in Lemma A.1 for completeness. □

A.3 Proof of Theorem 3.2 (asymptotic conditional coverage)

Proof of Theorem 3.2. As in Section 3.2, we denote by $\mathbf{R}_k(\cdot|x)$ the conditional empirical MK rank map and by $\widehat{\mathcal{Q}}_k(\alpha|x)$ the conditional MK quantile region of level $\alpha \in [0,1]$ (del Barrio et al., 2024). By assumption, k is a function of n satisfying $k \to +\infty$ as $n \to +\infty$ and $\frac{k}{n} \to 0$. For clarity, we omit the explicit dependence of k on n in our notation.

By definition of our prediction regions (8), the desired result (10) can also be written as

$$\lim_{n,k\to+\infty} \mathbb{P}\Big(s(X_{\text{test}}, Y_{\text{test}}) \in \widehat{\mathcal{Q}}_k\Big((1+\frac{1}{k})\alpha|X_{\text{test}}\Big)\Big|X_{\text{test}}\Big) = \alpha\,.$$

By applying Corollary 3.4 from del Barrio et al. (2024), we obtain

$$\forall \tau \in [0,1], \quad \lim_{n,k\to+\infty} \mathbb{P}\Big(s(X_{\text{test}}, Y_{\text{test}}) \in \widehat{\mathcal{Q}}_k(\tau|X_{\text{test}})\Big|X_{\text{test}}\Big) = \tau\,. \tag{27}$$

Therefore, the main technical challenge of our proof is to understand how this asymptotic convergence guarantee combines with a coverage level choice of $(1+\frac{1}{k})\alpha$.

For any $\tau \in [0,1]$, the limit in (27) can be equivalently written as

$$\forall \epsilon > 0,\, \exists N_2 \in \mathbb{N},\, \forall k \geq N_2 : \quad \tau - \epsilon \leq \mathbb{P}\Big(s(X_{\text{test}}, Y_{\text{test}}) \in \widehat{\mathcal{Q}}_k(\tau|X_{\text{test}})\Big|X_{\text{test}}\Big) \leq \tau + \epsilon\,. \tag{28}$$

Consider $N_1 \in \mathbb{N}$ large enough so that $(1+\frac{1}{2N_1})\alpha \leq 1$. By plugging $\tau = (1+\frac{1}{2N_1})\alpha$ and $\epsilon = \frac{\alpha}{2N_1}$ in (28), we obtain that for sufficiently large k (thus, for sufficiently large n as well),

$$\exists N_2 \in \mathbb{N},\, \forall k \geq N_2 : \quad \alpha \leq \mathbb{P}\Big(s(X_{\text{test}}, Y_{\text{test}}) \in \widehat{\mathcal{Q}}_k\Big(\Big(1+\frac{1}{2N_1}\Big)\alpha|X_{\text{test}}\Big)\Big|X_{\text{test}}\Big) \leq \alpha\Big(1+\frac{1}{N_1}\Big). \tag{29}$$

Recall that MK quantile regions (5) can be characterized by the following relation,

$$\forall \beta \in [0,1], \quad s \in \widehat{\mathcal{Q}}_k(\beta|X_{\text{test}}) \iff \|\mathbf{R}_k(s|X_{\text{test}})\| \leq \frac{\lceil \beta k \rceil}{k}.$$

As a consequence, they enjoy a monotonic embedding property, in the sense that they are nested (Chernozhukov et al., 2017; Hallin et al., 2021). Indeed, for any s such that $s \in \widehat{\mathcal{Q}}_k(\beta|X_{\text{test}})$, then $s \in \widehat{\mathcal{Q}}_k\Big(\beta'|X_{\text{test}}\Big)$ for $\beta' \geq \beta$, thus $\widehat{\mathcal{Q}}_k(\beta|X_{\text{test}}) \subseteq \widehat{\mathcal{Q}}_k(\beta'|X_{\text{test}})$.

As a byproduct, for sufficiently large $k \geq \max(N_2, 2N_1)$, the following holds

$$\widehat{\mathcal{Q}}_k(\alpha|X_{\text{test}}) \subseteq \widehat{\mathcal{Q}}_k\Big(\Big(1+\frac{1}{k}\Big)\alpha|X_{\text{test}}\Big) \subseteq \widehat{\mathcal{Q}}_k\Big(\Big(1+\frac{1}{2N_1}\Big)\alpha|X_{\text{test}}\Big)\,.$$

Combining this with (29), we obtain that for $k \geq \max(N_2, 2N_1)$,

$$\mathbb{P}\Big(s(X_{\text{test}}, Y_{\text{test}}) \in \widehat{\mathcal{Q}}_k\Big(\alpha|X_{\text{test}}\Big)\Big|X_{\text{test}}\Big) \leq \mathbb{P}\Big(s(X_{\text{test}}, Y_{\text{test}}) \in \widehat{\mathcal{Q}}_k\Big((1+\frac{1}{k})\alpha|X_{\text{test}}\Big)\Big|X_{\text{test}}\Big) \leq \alpha(1+\frac{1}{N_1})\,. \tag{30}$$

Taking the limit of (30) when $k, n \to +\infty$ yields

$$\alpha \leq \lim_{n,k+\infty} \mathbb{P}\Big(s(X_{\text{test}}, Y_{\text{test}}) \in \widehat{\mathcal{Q}}_k\Big((1+\frac{1}{k})\alpha|X_{\text{test}}\Big)\Big|X_{\text{test}}\Big) \leq \alpha(1+\frac{1}{N_1})\,, \tag{31}$$

where the left-hand side term of that inequality follows from applying (27).

We conclude by taking the limit of (31) when $N_1 \to +\infty$, which is permitted since the above reasoning was conducted for an arbitrary N_1 as long as $(1+\frac{1}{2N_1})\alpha \leq 1$ (*i.e.,* $(2N_1+1)\alpha \leq 2N_1$). □

B Experimental Details

B.1 Implementation details for regression

In Figure 2, the score $s(X,Y) = Y - \hat{f}(X)$ can be seen as a random vector ζ distributed as $\sum_{\ell=1}^{3} \pi_\ell \mathcal{N}(m_\ell, \Sigma_\ell)$, where $\pi_1 = \pi_2 = \frac{3}{8}$, $\pi_3 = \frac{1}{4}$, $m_1 = \binom{5}{0}$, $m_2 = -m_1$, $m_3 = \binom{0}{0}$, $\Sigma_1 = \left(\begin{smallmatrix} 4 & -3 \\ -3 & 4 \end{smallmatrix}\right)$, $\Sigma_2 = \left(\begin{smallmatrix} 4 & 3 \\ 3 & 4 \end{smallmatrix}\right)$, $\Sigma_3 = \left(\begin{smallmatrix} 3 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

For our real data experiments, we used datasets available in Tsoumakas et al. (2011). The following table specifies the number of observations and variables (for the features X and for the output Y) for each dataset.

Name	**#Instances**	**#Features**	**#Targets**
atp1d	337	411	6
rf1	9125	64	8
scm20d	8966	61	16
jura	359	15	3
wq	1060	16	14
enb	768	8	2

Table 1: Details of datasets used for multiple-output regression.

B.2 Implementation details for classification

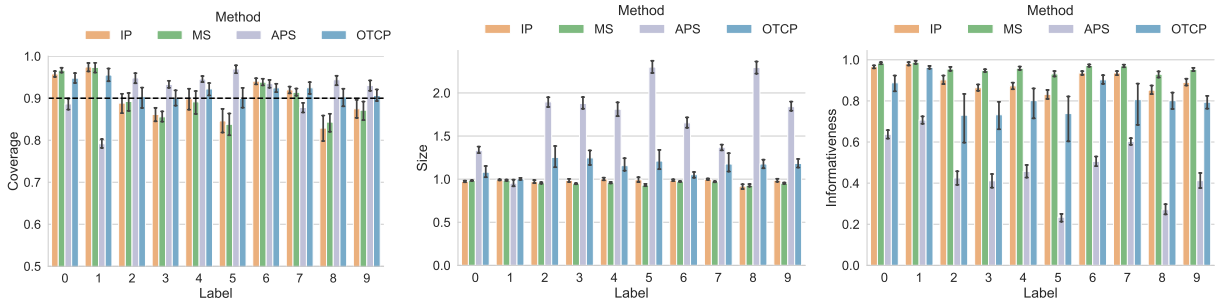
The implementation of the ARS score relies on codes made available with the original paper Romano et al. (2020).

Experiments on Fashion-MNIST in Figure 8 involve a random forest classifier implemented with the Python library scikit-learn. We used 20 000 data splitted in train/calibration/test with ratio 10%/80%/10%, since this is sufficient for the classifier to reach 90% accuracy and to ensure reasonable size for the test data. Metrics are computed and averaged over $N = 10$ repeated random draws.

B.3 Additional experiments

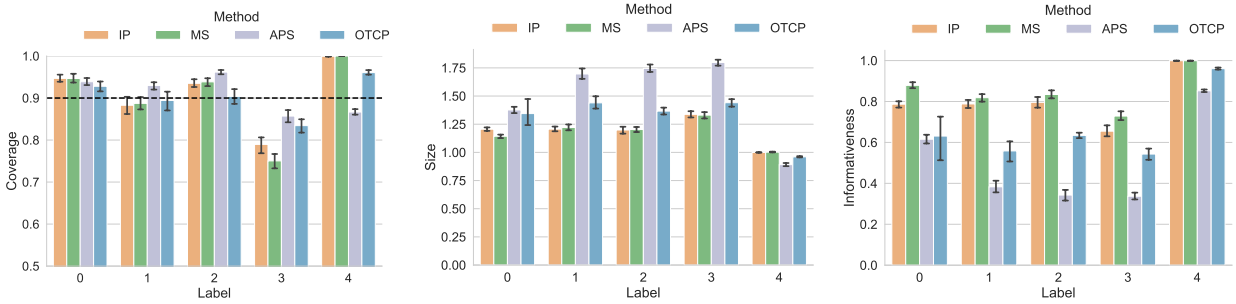
In Figure 9, we run on MNIST the same experiments than in Figure 8 with identical number of samples, number of repetitions and classifier.

In Figure 10, we conduct the same experiment than in Figure 8 but with a subset of $K = 5$ labels of the Fashion-MNIST dataset (T-shirt/top, Pullover, Coat, Shirt, Ankle boot). In Figure 11, the same experiment was run on $K = 5$ classes of CIFAR-10 and lead to similar conclusions on the appropriate behavior of the OT-CP methodology. For the CIFAR-10 dataset, the predictor is chosen as a neural network with two hidden layers of respective size $3000, 1000$ and ReLU activations. Both experiments of Figures 10 and 11 were repeated 10 times over 10000 randomly chosen observations splitted in train/calibration/test with ratio $50\%, 40\%, 10\%$. The performances of OT-CP are even better for $K = 5$ than when $K = 10$, reaching the efficiency of the scores IP / MS while improving adaptivity, akin to the score APS. This indicates on the relation between K and OT-CP when the score is the K-dimensional $|\bar{y} - \hat{\pi}(x)|$. In particular, for large K, OT-CP might benefit from further design of the score, inspired *e.g.,* by univariate penalized approaches (Angelopoulos et al., 2021; Melki et al., 2024).



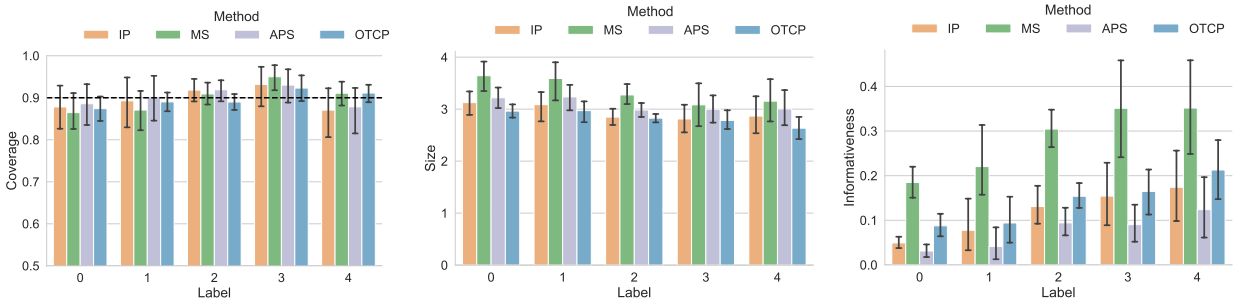
(a) Coverage (b) Average size (c) Informativeness

Figure 9: Label-wise results on $K = 10$ classes of MNIST



(a) Coverage (b) Average size (c) Informativeness

Figure 10: Label-wise results on $K = 5$ classes of Fashion-MNIST



(a) Coverage (b) Average size (c) Informativeness

Figure 11: Label-wise results on $K = 5$ classes of CIFAR-10