HAL open science

Selecting informative conformal prediction sets with false coverage rate control

Ulysse Gazin, Ruth Heller, Ariane Marandon, Etienne Roquain

To cite this version:

Ulysse Gazin, Ruth Heller, Ariane Marandon, Etienne Roquain. Selecting informative conformal prediction sets with false coverage rate control. 2024. hal-04539688

HAL Id: hal-04539688

https://hal.science/hal-04539688v1

Preprint submitted on 9 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Selecting informative conformal prediction sets with false coverage rate control

Ulysse Gazin

Université Paris Cité and Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75013 Paris, France.

E-mail: ugazin@lpsm.paris

Ruth Heller

Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel.

E-mail: ruheller@gmail.com

Ariane Marandon

The Alan Turing Institute, London, United Kingdom.

E-mail: marandon-carlhian@turing.ac.uk

Etienne Roquain

Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France.

E-mail: etienne.roquain@upmc.fr

Summary. In supervised learning, including regression and classification, conformal methods provide prediction sets for the outcome/label with finite sample coverage for any machine learning predictor. We consider here the case where such prediction sets come after a selection process. The selection process requires that the selected prediction sets be 'informative' in a well defined sense. We consider both the classification and regression settings where the analyst may consider as informative only the sample with prediction sets small enough, excluding null values, or obeying other appropriate 'monotone' constraints. We develop a unified framework for building such informative conformal prediction sets while controlling the false coverage rate (FCR) on the selected sample. While conformal prediction sets after selection have been the focus of much recent literature in the field, the new introduced procedures, called `InfoSP` and `InfoSCOP`, are to our knowledge the first ones providing FCR control for informative prediction sets. We show the usefulness of our resulting procedures on real and simulated data.

Keywords: classification, false discovery rate, label shift, prediction interval, regression, selective inference.

1. Introduction

In modern data analysis, machine learning algorithms are often used to make predictions and a main challenge is to measure the uncertainty of such methods. Conformal inference offers an elegant solution to this problem, by providing prediction sets that provably cover the true value with high probability, for any sample size, any predictive algorithm,

and any distribution of the data (Vovk et al., 2005). We consider the following classic split/inductive conformal prediction setting (Papadopoulos et al., 2002; Vovk et al., 2005; Lei et al., 2014). Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random vector with unknown distribution P_{XY}. Typically, $\mathcal{X}$ is a subset of $\mathbb{R}^d$ (real valued covariates), and $\mathcal{Y} = [K]$ (classification of among $K \geq 2$ classes) or $\mathcal{Y} = \mathbb{R}$ (regression, with a real valued outcome)†. In this setting, there are two independent samples of points (X_i, Y_i): the calibration sample $\{(X_j, Y_j), j \in [n]\}$; and the test sample $\{(X_{n+i}, Y_{n+i}), i \in [m]\}$.

While all measurements are observed in the calibration sample, only the X_i's are observed in the test sample and the aim is to provide prediction sets for the unobserved $(Y_{n+i}, i \in [m])$. We denote the prediction set for Y_{n+i} by $\mathcal{C}_{n+i}$. It is a subset of $[K]$ for classification and a subset (often an interval) of $\mathbb{R}$ for regression. The classic conformal prediction set for Y_{n+i}, denoted by $\mathcal{C}^{\alpha}_{n+i}$, is a function of $\{(X_j, Y_j), j \in [n]\}, X_{n+i}$, and α (Sadinle et al., 2019; Lei et al., 2018) and has the following coverage guarantee, assuming that (X_{n+i}, Y_{n+i}) and $(X_j, Y_j), j \in [n]$, are exchangeable pairs of observations from any distribution P_{XY}:

$$\forall \alpha \in (0, 1), \ \mathbb{P}\left(Y_{n+i} \in \mathcal{C}^{\alpha}_{n+i}\right) \geq 1 - \alpha. \tag{1}$$

In practice, the size of the test sample m is often large, encompassing hundreds or thousands of unlabeled examples. Inferring on all of them is unnecessary or inefficient in many applications (Jin and Candes, 2023; Bao et al., 2024). For example, in classification, if each image belongs to one of $[K]$ categories, the analyst is not interested in the examples where $\mathcal{C}^{\alpha}_{n+i} = [K]$. Thus, it is natural to assume that a subset of individuals will be selected. However, reporting their prediction sets constructed to have at least $1 - \alpha$ confidence is problematic, since conditional on being selected, the coverage may be much smaller (Benjamini and Yekutieli, 2005; Benjamini and Bogomolov, 2013), see also Figure 1 below.

Our focus is on the common setting where the analyst is only interested in reporting "interesting" or "informative" prediction sets, i.e., that cover only part of the $\mathcal{Y}$ space in a well defined sense to the analyst, which definition depends on the specific context. For concreteness, we start by providing typical examples of what can be the pre-specified collection $\mathcal{I}$ of informative subsets of $\mathcal{Y}$.

Example 1.1 (Informative prediction sets in regression, $\mathcal{Y} = \mathbb{R}$). *(a) Intervals excluding a range of values: $\mathcal{I} = \{I \text{ interval of } \mathbb{R} \ : \ I \cap \mathcal{Y}_0 = \emptyset\}$ for some subset of null values $\mathcal{Y}_0 \subset \mathcal{Y}$ that are considered as uninteresting for the user. The choice $\mathcal{Y}_0 = (-\infty, y_0]$ is related to the selection proposed in Jin and Candes (2023), for which a "normal" value for the outcome is a value below y_0.*

(b) Length-restricted intervals: $\mathcal{I} = \{I \text{ interval of } \mathbb{R} \ : \ |I| \leq 2\lambda_0\}$ for some $\lambda_0 > 0$, which are useful for only reporting prediction intervals that are accurate enough.

Example 1.2 (Informative prediction sets in classification, $\mathcal{Y} = [K]$). *(a) Excluding one class: $\mathcal{I} = \{C \subset [K] \ : \ y_0 \notin C\}$ for some null class $y_0 \in [K]$. It is suitable when the user does not want to report prediction sets for individuals that*

†Our theory can be applied for more general observation spaces (e.g., $\mathcal{Y} = \mathbb{R}^{d'}$) but we consider the most common settings $\mathcal{Y} \in \{[K], \mathbb{R}\}$ for simplicity.

are in class y_0. This can be extended to excluding several classes: $\mathcal{I} = \{C \subset [K] : C \cap \mathcal{Y}_0 = \emptyset\}$ for some label set $\mathcal{Y}_0 \subset [K]$.

(b) *Non-trivial classification: $\mathcal{I} = \{C \subset [K] : |C| \leq K - 1\}$. It is appropriate when the analyst wants a label set that is minimally informative. More generally, at most k_0-sized classification can be considered with $\mathcal{I} = \{C \subset [K] : |C| \leq k_0\}$.*

A common target error guarantee is that the inference on at most α examples among the selected is expected to be false. This is a classical error criterion in the selective inference literature, (Benjamini and Yekutieli, 2005; Benjamini and Bogomolov, 2013; Weinstein and Ramdas, 2020). It has been used, e.g., for novelty detection (Bates et al., 2023; Marandon et al., 2024), for classification (Rava et al., 2021; Zhao and Su, 2023; Jin and Candes, 2023), for regression (Bao et al., 2024), and for unsupervised clustering (Mary-Huard et al., 2022; Marandon et al., 2022). For selecting prediction sets, the target error guarantee is thus that at most α examples among the selected are expected to have prediction sets that do not cover their true outcome value. The false coverage proportion (FCP) for the procedure $\mathcal{R} = (\mathcal{C}_{n+i})_{i \in \mathcal{S}}$ is defined as the proportion of non-covered examples in the selected set $\mathcal{S}$:

$$\mathrm{FCP}(\mathcal{R}, Y) = \frac{\sum_{i \in \mathcal{S}} \mathbf{1}\{Y_{n+i} \notin \mathcal{C}_{n+i}\}}{1 \vee |\mathcal{S}|}, \tag{2}$$

and the target error is simply its expectation, which we refer to as the false coverage rate (FCR) as in Bao et al. (2024):

$$\mathrm{FCR}(\mathcal{R}) = \mathbb{E}[\mathrm{FCP}(\mathcal{R}, Y)]. \tag{3}$$

In this work, we consider two popular models for generating the samples‡. First, the *iid model* (both for regression and classification): the variables $(X_i, Y_i) \sim P_{XY}$, $i \in [n+m]$, are all independent and identically distributed (iid). This is the standard assumption classically used for conformal prediction (Vovk et al., 2005). The parameter of the model is in that case P_{XY}. Second, the *class-conditional model* (only the classification setting): all the labels $(Y_i, i \in [n+m])$ are deterministic and the covariates $(X_i, i \in [n+m])$ are mutually independent with $X_j \sim P_{X|Y=Y_j}$. It relaxes the exchangeable model assumptions of iid model, by only requiring that the distribution within each class is the same for the test and calibration sample (Sadinle et al., 2019; Ding et al., 2023). The target FCR is then conditional on the labels. The parameters of the model are then given by $(P_{X|Y=k})_{k \in [K]}$ and $(Y_i, i \in [n+m])$.

We now briefly summarize the contributions of our work. We first introduce a new method, called `InfoSP` (Informative selective prediction sets), that selects only informative prediction sets with a level α FCR guarantee on the selected (§ 3.1). Formally, this means that we achieve both $\mathrm{FCR}(\mathcal{R}) \leq \alpha$ and $\forall i \in \mathcal{S}$, $\mathcal{C}_{n+i} \in \mathcal{I}$, for $\mathcal{I}$ being the collection of informative subsets. The FCR control of `InfoSP` is established both in the iid model and class-conditional model (see Theorem 3.1).

‡In both models, the independence assumption can be relaxed. In the iid model, it is enough that all $n+m$ samples are exchangeable. In the class-conditional model, it is enough that the subset of $[n+m]$ of samples from the same class is exchangeable, for all classes in $[K]$. See § C for more details.

We introduce a second procedure, called `InfoSCOP` (Informative selective conditional prediction sets), that has the same theoretical properties as `InfoSP` in the iid model (§ 3.2) and that relies on an initial selection step that is aimed at eliminating (at least some of) the examples for which informative prediction sets cannot be constructed. Further selection then takes place in order to ensure that all reported prediction sets are informative. While the pre-processing step is inspired by Bao et al. (2024), their theoretical framework precludes this type of selection (see § A for more details).

Third, our main theoretical FCR control guarantees come from a single general theorem in § C.1 that generalizes the theoretical results in Benjamini and Yekutieli (2005); Benjamini and Bogomolov (2013), to include more general classes of p-values and selection rules. Importantly, our novel theory supports conformal p-values, and selecting only informative prediction sets, as specific examples.

We optimize the analysis pipeline for common informative selection rules in § 4 (regression) and § 5-§ 6 (classification), while providing additional theoretical results and appropriate numerical experiments. Finally, an application to directional FDR control is also provided in § B.

To provide an intuition for our approach, `InfoSP` is illustrated in Figure 1 for the classification case (for $K = 3$ classes). Left panels display a naive method reporting the marginal classical conformal prediction sets $\mathcal{C}^{\alpha}_{n+i}$ in (1) for all those that are informative, that is, such that $\mathcal{C}^{\alpha}_{n+i} \in \mathcal{I}$, without further correction. Since no error can occur when the prediction set is trivial, the selection always inflates the FCP values. The new procedure `InfoSP` is displayed in the right panels: prediction sets are made slightly larger to maintain a correct FCP value while the selection (red boxes) guarantees that only informative (i.e., non trivial) prediction sets are selected. This example is only for one data generation: it is presented here only for illustrative propose and more accurate in-expectation FCR values are given in § 6. The regression case is illustrated in Figures 2 and 3 (§ 4), for which the second procedure `InfoSCOP` is also displayed. It is worth to note that in some situations, the latter may even results in prediction sets that are smaller (!) than those of the naive method.

1.1. Relation to previous work

There are interesting connections between our approach and previous work of the literature: selecting confidence interval that satisfies specific notions of informativeness (Weinstein and Yekutieli, 2020; Weinstein and Ramdas, 2020); multi-class classification (Zhao and Su, 2023); and very recent works on selective conformal inference (Bao et al., 2024; Jin and Ren, 2023). Due to the limited space, the details are in§ A of the SM.

We also underline that our second procedure `InfoSCOP` relies on the approach of splitting the calibration sample, which has already been used in conformal literature in different contexts. The idea is to enable an additional data-driven tuning of the method by only paying the price of splitting the calibration sample. For instance, it has been provided in Marandon et al. (2024) for the task of building adaptive scores. In the present aim of controlling the FCR on a data-driven selection, we formulate a general statement in the SM, see Lemma F.4. It applies to any type of data-driven selection and error rate.

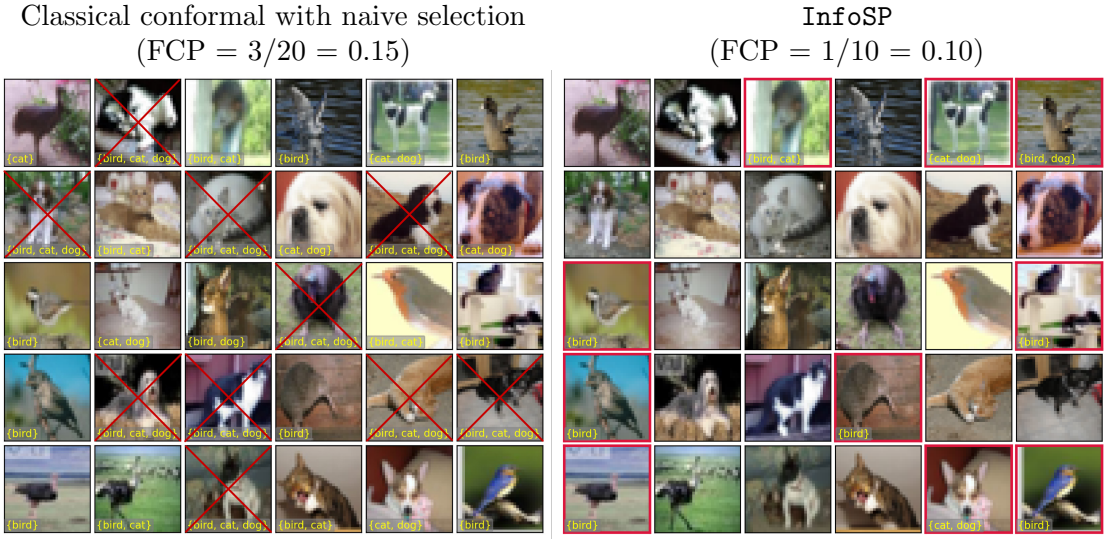


Fig. 1: Informative prediction sets in classification for CIFAR-10 dataset, restricted to the $K = 3$ classes "bird", "cat", and "dog" classes (iid setting). Informative prediction subsets are those of size smaller than $K - 1 = 2$ (i.e., non-trivial, Example 1.2 item 1). Selection by `InfoSP` are framed in red (right panel). $\alpha = 10\%$. See § 6 for more details.

2. Preliminaries

2.1. Notation

Expectations and probabilities are denoted for the iid model with $\mathbb{E}_{(X,Y)\sim P_{XY}}(\cdot)$ and $\mathbb{P}_{(X,Y)\sim P_{XY}}(\cdot)$, and for the class-conditional model with $\mathbb{E}_{X\sim P_{X|Y}}(\cdot)$ and $\mathbb{P}_{X\sim P_{X|Y}}(\cdot)$, respectively. If the data generation process is clear from the context, or if the expression is relevant for both models, then the subscript is omitted. In addition, $A \subset B$ means that the set A is included in the set B with a possible equality. For any subset $C \subset \mathbb{R}$, $|C|$ denotes the cardinality of C if the set C is finite, and the Lebesgue measure of C if C is an interval. Finally, for two samples $\mathcal{D}_1$ and $\mathcal{D}_2$, $\mathcal{D}_1 \cup \mathcal{D}_2$ denotes the new sample formed by concatenating the elements of $\mathcal{D}_1$ and $\mathcal{D}_2$.

2.2. Conformal prediction sets

The classical conformal prediction set for Y_{n+i} is given by

$$\mathcal{C}^{\alpha}_{n+i}(\mathbf{p}) = \{y \in \mathcal{Y} \,:\, p_i^{(y)} > \alpha\}, \quad i \in [m], \tag{4}$$

where $\mathbf{p} = (p_i^{(y)}, y \in \mathcal{Y}, i \in [m])$ is a collection of conformal p-values satisfying the following super-uniform guarantee:

$$\mathbb{P}(p_i^{(Y_{n+i})} \leq \alpha) \leq \alpha, \quad i \in [m], \tag{5}$$

where the above probability is computed either in the iid (Y_{n+i} random) or conditional (Y_{n+i} fixed) model. Super-uniformity (5) implies directly that $\mathcal{C}^{\alpha}_{n+i}(\mathbf{p})$ in (4) provides

$1-\alpha$ coverage for Y_{n+i}, that is,

$$\mathbb{P}(Y_{n+i} \in \mathcal{C}^{\alpha}_{n+i}(\mathbf{p})) \geq 1-\alpha, \quad i \in [m], \quad \alpha \in (0,1), \tag{6}$$

which is generally referred to as marginal coverage.

The p-value family is built from the calibration and test sample by using non-conformity score functions $S_y : x \in \mathbb{R}^d \mapsto \mathbb{R}$, $y \in \mathcal{Y}$, measuring the inadequacy between y and the prediction at point x. Importantly, we follow a split/inductive conformal approach, where the score functions have been computed from an independent training data sample so that they can be considered as fixed here§ (and all probabilities/expectations are taken conditional on that training sample).

Assumption 1. *The score functions $S_y(\cdot)$, $y \in \mathcal{Y}$, have been computed from an independent training sample and the computed scores $(S_{Y_i}(X_i), i \in [n+m])$ have no ties almost surely. When $\mathcal{Y} = \mathbb{R}$, the score function is regular in the following sense: for every $x \in \mathbb{R}^d$, the function $y \in \mathbb{R} \mapsto S_y(x) \in \mathbb{R}$ is right continuous with left limits.*

Assumption 1 is a very mild assumption, which is typically satisfied. For instance, for the regression case, a classical choice is the locally weighted residual function $S_y(x) = |y - \mu(x)|/\sigma(x)$ where $\mu(x) \in \mathcal{Y}$ is the predicted outcome at point x and $\sigma(x)$ is the predicted standard deviation of Y given $X = x$ (Lei et al., 2018). Another common example is the quantile-based score function $S_y(x) = \max(q_{\beta_0}(x) - y, y - q_{\beta_1}(x))$ where $q_\beta(x)$ is the predicted β-quantile of Y given $X = x$, which corresponds to the so-called conformalized quantile regression (Romano et al., 2019; Sesia and Romano, 2021) (for some prespecified $0 < \beta_0 < \beta_1 < 1$). More generally, we refer the reader to Gupta et al. (2022) for a general framework giving rise to a large class of score functions. In the classification case, the typical score is the residual function $S_y(x) = 1 - \pi_y(x)$ where $\pi_y(x)$ is an estimator of the probability to generate label y at point x.

Formally, the p-value family is given as follows:

- full-calibrated p-values: $\overline{\mathbf{p}} = (\overline{p}_i^{(y)}, i \in [m], y \in \mathcal{Y})$, both for the regression and classification cases, with

$$\overline{p}_i^{(y)} = \frac{1}{n+1}\Big(1 + \sum_{j=1}^{n} \mathbf{1}\{S_{Y_j}(X_j) \geq S_y(X_{n+i})\}\Big), \quad i \in [m], y \in \mathcal{Y}. \tag{7}$$

- class-calibrated p-values: $\widetilde{\mathbf{p}} = (\tilde{p}_i^{(y)}, i \in [m], y \in \mathcal{Y})$, only for the classification case $\mathcal{Y} = [K]$, with

$$\tilde{p}_i^{(y)} = \frac{1}{|\mathcal{D}^{(y)}_{\text{cal}}| + 1}\Big(1 + \sum_{j \in \mathcal{D}^{(y)}_{\text{cal}}} \mathbf{1}\{S_y(X_j) \geq S_y(X_{n+i})\}\Big), \quad i \in [m], y \in \mathcal{Y}, \tag{8}$$

  where $\mathcal{D}^{(y)}_{\text{cal}} = \{j \in [n] \ : \ Y_j = y\}$ corresponds to the elements of the calibration sample that have label $y \in \mathcal{Y}$.

§Note that we can relax slightly this assumption: our theory also allows this function to depend on the calibration plus test samples in an exchangeable way, see Assumption 6.

Both p-values $\overline{p}_i^{(y)}$ and $\widetilde{p}_i^{(y)}$ are computed by examining how extreme the score $S_y(X_{n+i})$ is among the scores of the true labels in the calibration sample. Full-calibrated p-values and class-calibrated p-values satisfy the super-uniformity (5) in the iid model and class-conditional model, respectively, by using an exchangeability argument, see Vovk et al. (2005); Romano and Wolf (2005); Bates et al. (2023). This means that prediction set $\mathcal{C}_{n+i}^{\alpha}(\mathbf{p})$ in (4) satisfies the marginal coverage (6) in each context. As a result, the false coverage rate (3) for the full selection $\mathcal{S} = [m]$ is controlled at level α, that is,

$$\mathbb{E}\left[\frac{\sum_{i\in[m]} \mathbf{1}\{Y_{n+i} \notin \mathcal{C}_{n+i}^{\alpha}(\mathbf{p})\}}{m}\right] = m^{-1}\sum_{i\in[m]} \mathbb{P}(Y_{n+i} \notin \mathcal{C}_{n+i}^{\alpha}(\mathbf{p})) \leq \alpha. \tag{9}$$

Remark 2.1. *The prediction set $\mathcal{C}_{n+i}^{\alpha}(\mathbf{p})$ in (4) can be described as a score level set, with a threshold depending on the calibration scores. Formally, we have*

$$\mathcal{C}_{n+i}^{\alpha}(\mathbf{p}) = \{y \in \mathcal{Y} \,:\, p_i^{(y)} > \alpha\} = \{y \in \mathcal{Y} \,:\, S_y(X_{n+i}) \leq \hat{s}_\alpha\} \tag{10}$$

where the score threshold $\hat{s}_\alpha$ is $S_{(\lceil(1-\alpha)(n_{cal}+1)\rceil)}$ with $S_{(1)} \leq \cdots \leq S_{(n_{cal})}$ being the ordered calibration scores (and with the convention $S_{(n_{cal}+1)} = +\infty$). For full-calibrated p-values, the $n_{cal} = n$ calibration scores are $S_{Y_j}(X_j)$, $j \in [n]$. For class-conditional p-values, the $n_{cal} = |\mathcal{D}_{cal}^{(y)}|$ calibration scores are $S_y(X_j)$, $j \in \mathcal{D}_{cal}^{(y)}$ ($\hat{s}_\alpha$ depends on y).

2.3. $\mathcal{I}$-adjusted p-values

Our theory relies on the following assumption.

Assumption 2. *The subset collection $\mathcal{I}$ is monotone in the following sense: for the considered p-value collection $\mathbf{p}$ (either $\overline{\mathbf{p}}$ or $\widetilde{\mathbf{p}}$), we have*

(i) If a prediction set is informative, then all the prediction sets it contains are also informative, that is, for all $\mathcal{C}, \mathcal{C}'$ (subsets of $[K]$ for classification, intervals of $\mathbb{R}$ for regression) with $\mathcal{C}' \subset \mathcal{C}$, $\mathcal{C} \in \mathcal{I}$ implies $\mathcal{C}' \in \mathcal{I}$.

(ii) Almost surely, the function $\alpha \in (0,1] \mapsto \mathbf{1}\{\mathcal{C}_{n+i}^{\alpha}(\mathbf{p}) \in \mathcal{I}\} \in \{0,1\}$ is right-continuous.

(iii) (for regression) For all $\alpha \in (0,1)$, almost surely, $\mathcal{C}_{n+i}^{\alpha}(\mathbf{p})$ is an interval of $\mathbb{R}$.

Note that Assumption 2 implies that $\alpha \in (0,1] \mapsto \mathbf{1}\{\mathcal{C}_{n+i}^{\alpha}(\mathbf{p}) \in \mathcal{I}\} \in \{0,1\}$ is both right-continuous and nondecreasing: if $\alpha \leq \alpha'$, it follows that $\mathcal{C}_{n+i}^{\alpha'}(\mathbf{p}) \subset \mathcal{C}_{n+i}^{\alpha}(\mathbf{p})$ from (4) and thus $\mathcal{C}_{n+i}^{\alpha}(\mathbf{p}) \in \mathcal{I}$ implies that $\mathcal{C}_{n+i}^{\alpha'}(\mathbf{p}) \in \mathcal{I}$ by Assumption 2 (i) (iii).

As a result, we can define the $\mathcal{I}$-adjusted p-value vector by $\mathbf{q} = (q_i)_{i\in[m]}$ where

$$q_i = \min\{\alpha \in (0,1] \,:\, \mathcal{C}_{n+i}^{\alpha}(\mathbf{p}) \in \mathcal{I}\}, \tag{11}$$

with by convention $q_i = 1$ if the set is empty. Assumption 2 can be easily checked for Examples 1.1 and 1.2, with an explicit expression for q_i's.

Example 2.1 (Example 1.1 continued). *For regression (see § 4 for more details): $q_i = \sup_{y\in[a,b]} p_i^{(y)}$ for intervals excluding $\mathcal{Y}_0 = [a,b]$; $q_i = (n+1)^{-1}(1 +$*

$\sum_{j=1}^{n} \mathbf{1}\{S_{Y_j}(X_j) > A\}\big)$ *for length-restricted intervals* I *(with* $|I| \leq 2\lambda_0$*), where* $A = \lambda_0/\sigma(X_{n+i})$ *for* $S_y(x) = |y - \mu(x)|/\sigma(x)$ *and* $A = \lambda_0 - (q_{\beta_1}(X_{n+i}) - q_{\beta_0}(X_{n+i}))/2$ *for* $S_y(x) = \max(q_{\beta_0}(x) - y, y - q_{\beta_1}(x))$.

Example 2.2 (Example 1.2 continued). *For classification:* $q_i = p_i^{(y_0)}$ *for excluding one-class;* $q_i = \max_{y \in \mathcal{Y}_0} p_i^{(y)}$ *for excluding several classes;* $q_i = \min_{y \in [K]} p_i^{(y)}$ *for non trivial classification;* $q_i =$ *the* $(K - k_0)$*-th smallest element in the set* $\{p_i^{(y)}, y \in [K]\}$ *for at most* k_0*-sized classification.*

Example 2.3 (Combining informative subset collections). *Let two subset collections* $\mathcal{I}_1$ *and* $\mathcal{I}_2$ *that satisfy Assumption 2 with adjusted p-values given by* $(q_{1,i})_{i \in [m]}$ *and* $(q_{2,i})_{i \in [m]}$*, respectively. Then we can easily check that the intersected collection* $\mathcal{I} := \{I_1 \cap I_2, I_1 \in \mathcal{I}_1, I_2 \in \mathcal{I}_2\}$ *also satisfies Assumption 2 with adjusted p-values given by* $q_i = \max\{q_{1,i}, q_{2,i}\}$*,* $i \in [m]$*. This is useful to combine the constraints imposed by the informativeness. For instance, in the regression case (Examples 1.2 and 2.2), we can declare a subset as informative if it excludes a null class while it is of cardinality at most one (see § B for a concrete application).*

From its definition in (11), it follows that q_i can be seen as a p-value to test whether Y_{n+i} lies in an informative set or not, that is, to test

$$H_{0,i}\text{: “}Y_{n+i} \notin \cup_{C \in \mathcal{I}} C\text{” versus } H_{1,i} : \text{“}Y_{n+i} \in \cup_{C \in \mathcal{I}} C\text{”}. \tag{12}$$

In cases where being informative is linked to particular values in $\mathcal{Y}$, this testing problem takes an especially meaningful form. In classification, for excluding one class in classification: $q_i = p_i^{(y_0)}$ tests $H_{0,i}$: “$Y_{n+i} = y_0$” versus $H_{1,i}$: “$Y_{n+i} \neq y_0$”. In regression, for excluding $\mathcal{Y}_0 = [a, b]$: q_i tests $H_{0,i}$: “$Y_{n+i} \in [a, b]$” versus $H_{1,i}$: “$Y_{n+i} \notin [a, b]$”.

Note that in case where informative sets are those with small size (e.g., non-trivial classification or length-restricted intervals), we have $\cup_{C \in \mathcal{I}} C = \mathcal{Y}$, so the null hypothesis is always false, and the testing problem (12) is not interesting. However, in case the small size is just one of the criteria for being informative, as in Example 2.3, then the testing problem (12) may still be meaningful.

Remark 2.2. *Assumption 2 implies that the* $\mathcal{I}$*-adjusted p-value vector* $\mathbf{q}$ *is a nondecreasing function of the p-value collection, see Lemma C.5; Assumption 2 (iii) is always satisfied up to taking as prediction sets* $\mathcal{C}_{n+i}^{\alpha}(\mathbf{p})$ *the convex hull of the set in the right-hand-side of* (4)*. It is also satisfied without modifying* $\mathcal{C}_{n+i}^{\alpha}(\mathbf{p})$ *for any score function such that all the sets* $\{y \in \mathbb{R} \ : \ S_y(x) \leq s\}$ *are intervals, which is often the case, see Remark 2.1 and § 4.*

Remark 2.3. *In a context of building online confidence intervals, Weinstein and Ramdas (2020) have proposed to report only intervals that are “good” in the sense that they “localize” the signal. Formalizing their proposal in our regression setting and with our notation, this corresponds to consider the informative collection* $\mathcal{I} = \{I \text{ interval of } \mathbb{R} \ : \ I \subset \mathcal{C}_\ell \text{ for one } \ell \in [L]\}$*, where* $\mathcal{C}_\ell$*,* $\ell \in [L]$*, are pre-specified disjoint subsets of* $\mathbb{R}$*. This collection satisfies Assumption 2 and the corresponding* $\mathcal{I}$*-adjusted p-values are given by* $q_i = \min_{\ell \in [L]} \sup_{y \notin \mathcal{C}_\ell} p_i^{(y)}$.

2.4. Aim: informative selection with FCR guarantees

The general inferential task is to produce prediction sets for examples of interest in the test sample, that is, after selection. The selection process is driven by the requirement that the prediction sets be informative, and the requirement of a relevant error control.

A *selective prediction set procedure* is a function of the observations $\{(X_j, Y_j), j \in [n]\}$, $\{X_{n+i}, i \in [m]\}$ of the form $\mathcal{R} = (\mathcal{C}_{n+i})_{i\in\mathcal{S}}$ where $\mathcal{S}$ is the selected subset of $[m]$ for which prediction sets are constructed, and $\mathcal{C}_{n+i} \subset \mathcal{Y}$ is the prediction set for $Y_{n+i}, i \in \mathcal{S}$.

A selective prediction set procedure $\mathcal{R} = (\mathcal{C}_{n+i})_{i\in\mathcal{S}}$ *is said to be $\mathcal{I}$-informative (or informative)* if the selection $\mathcal{S}$ is a subset of $[m]$ that imposes that $\mathcal{C}_{n+i}$ is informative, that is, $\forall i \in \mathcal{S}$, $\mathcal{C}_{n+i} \in \mathcal{I}$.

The FCR in the iid model and class-conditional model, respectively, are

$$\mathrm{FCR}(\mathcal{R}, P_{XY}) = \mathbb{E}_{(X,Y)\sim P_{XY}}[\mathrm{FCP}(\mathcal{R}, Y)]; \tag{13}$$

$$\mathrm{FCR}(\mathcal{R}, P_{X|Y}, Y) = \mathbb{E}_{X\sim P_{X|Y}}[\mathrm{FCP}(\mathcal{R}, Y)], \tag{14}$$

for the FCP in (2). The FCR expression in (14) for quantifying the errors among the selected is classical in the selective inference literature: since $(Y_{n+i})_{i\in[m]}$ is fixed, this is the false coverage rate on the parameters (Benjamini and Yekutieli, 2005; Benjamini and Bogomolov, 2013). On the other hand, in (13), the false coverage rate is on random outcomes (i.e., $(Y_{n+i})_{i\in[m]}$ in our setting) rather than on parameters, which is the usual setting in selective conformal inference (see references in § 1) and is related to the Bayes FDR criterion in the multiple testing literature, see, e.g., Efron et al. (2001).

We will focus on finding selective prediction set procedures $\mathcal{R} = \mathcal{R}_\alpha$ with either of the two following controls:

$$\sup_{P_{X,Y}} \{\mathrm{FCR}(\mathcal{R}, P_{X,Y})\} \leq \alpha \ ; \tag{15}$$

$$\sup_{P_{X|Y},Y} \{\mathrm{FCR}(\mathcal{R}, P_{X|Y}, Y)\} \leq \alpha \ . \tag{16}$$

Obviously, the class-conditional control (16) (considered only for classification) is stronger than the unconditional control (15) (considered both for classification and regression).

To balance with FCR control, we also consider the *resolution adjusted power*:

$$\mathrm{Pow}(\mathcal{R}) = \mathbb{E}\left[\sum_{i\in\mathcal{S}} \frac{\mathbf{1}\{Y_{n+i} \in \mathcal{C}_{n+i}\}}{|\mathcal{C}_{n+i}|}\right]. \tag{17}$$

Hence, for the same selection set, a decision with a smaller covering decision set $\mathcal{C}_{n+i}$ yields higher power. Our aim is to maximize the resolution adjusted power (i.e., informally, to select as many examples as possible that are informative, with as narrow as possible a prediction set for the selected examples), while controlling the FCR at a pre-specified level α.

Finally, considering the multiple testing problem (12), that test if Y_{n+i} lies in an informative set or not, we can also quantify the error amount of a given selection procedure $\mathcal{S}$ (by itself, without quantifying the non-coverage errors of the attached prediction sets), by letting (Benjamini and Hochberg, 1995)

$$\mathrm{FDP}(\mathcal{S}, Y) = \frac{\sum_{i\in\mathcal{S}} \mathbf{1}\{Y_{n+i} \notin \cup_{C\in\mathcal{I}} C\}}{1 \vee |\mathcal{S}|}. \tag{18}$$

For instance, when one wants to exclude a given label set $\mathcal{Y}_0$ in classification/regression, we have $\mathrm{FDP}(\mathcal{S}, Y) = (\sum_{i\in\mathcal{S}} \mathbf{1}\{Y_{n+i} \in \mathcal{Y}_0\})/(1 \vee |\mathcal{S}|)$. The false discovery rates in the iid model and class-conditional model are the corresponding expectations

$$\mathrm{FDR}(\mathcal{S}, P_{XY}) = \mathbb{E}_{(X,Y)\sim P_{XY}}[\mathrm{FDP}(\mathcal{S}, Y)]; \tag{19}$$

$$\mathrm{FDR}(\mathcal{S}, P_{X|Y}, Y) = \mathbb{E}_{X\sim P_{X|Y}}[\mathrm{FDP}(\mathcal{S}, Y)], \tag{20}$$

respectively. The following lemma holds.

LEMMA 2.1. *For any selective prediction set procedure $\mathcal{R} = (\mathcal{C}_{n+i})_{i\in\mathcal{S}}$ that is $\mathcal{I}$-informative, we have* $\mathrm{FDP}(\mathcal{S}, Y) \leq \mathrm{FCP}(\mathcal{R}, Y)$.

It comes directly from the fact that, for $i \in \mathcal{S}$, $Y_{n+i} \in \mathcal{C}_{n+i}$ implies $Y_{n+i} \in \cup_{C\in\mathcal{I}} C$ because $\mathcal{C}_{n+i} \in \mathcal{I}$. As a result, producing an informative selective prediction set procedure that controls the FCR at level α immediately ensures that the attached selection procedure controls the FDR at level α and thus comes with a relevant interpretation.

3. Main results

3.1. Informative selective prediction sets (`InfoSP`)

In order to have a level α FCR guarantee on the selected examples, it is necessary to correct the threshold α in $\mathcal{C}^{\alpha}_{n+i}(\mathbf{p})$ (4). A standard approach in the selective inference literature (Benjamini and Yekutieli, 2005; Benjamini and Bogomolov, 2013) is to use the reduced level $\alpha|\mathcal{S}|/m$ for a selection set $\mathcal{S}$. In order for the selective prediction set procedure to be $\mathcal{I}$ informative, the selection rule needs to be carefully selected. We shall use the following basic observation: selection by a thresholding rule on the family $\mathbf{q} = (q_i, i \in [m])$ given by (11) will result in selected examples that are $\mathcal{I}$ informative for prediction sets that are constructed at a level that is at least at the selection threshold, since $\mathcal{C}^t_{n+i} \in \mathcal{I}$ if and only if $q_i \leq t$ for all $t \in (0, 1]$ by the definition of q_i. Combining the standard approach for FCR control with this basic observation, leads us to use the following selection thresholding rule which is necessarily $\mathcal{I}$-informative: all examples with q_i at most

$$\max\left\{ t : t \leq \alpha \frac{\left(\sum_{j=1}^m \mathbf{1}\{q_j \leq t\}\right)}{m} \right\}.$$

This is exactly the BH selection rule (Benjamini and Hochberg, 1995) on the adjusted p-value vector $\mathbf{q} = (q_i, i \in [m])$. In practice, let us recall that the BH procedure $\mathrm{BH}(\mathbf{q})$ can be obtained as

$$\mathrm{BH}(\mathbf{q}) = \{i \in [m] \ : \ q_i \leq \alpha\hat{\ell}/m\}, \tag{21}$$

where $\hat{\ell} = |\mathrm{BH}(\mathbf{q})| = \max\{\ell \in [m] \ : \ q_{(\ell)} \leq \alpha\ell/m\}$ (with $\hat{\ell} = 0$ if the set is empty) and where $q_{(1)} \leq \cdots \leq q_{(m)}$ denote the ordered q_i's.

DEFINITION 1. *The informative selective prediction set procedure (`InfoSP`) based on a p-value family $\mathbf{p} = (p_i^{(y)}, y \in \mathcal{Y}, i \in [m])$, is defined as $\mathcal{R}^{\mathrm{InfoSP}}_{\alpha}(\mathbf{p}) = (\mathcal{C}^{\alpha|\mathrm{BH}(\mathbf{q})|/m}_{n+i}(\mathbf{p}))_{i\in\mathrm{BH}(\mathbf{q})}$, that is, is given as follows:*

(a) *Apply the BH procedure on the corresponding adjusted vector* $\mathbf{q} = (q_i, i \in [m])$ *given by* (11),(21) *and select* $\mathcal{S}(\mathbf{p}) = \mathrm{BH}(\mathbf{q}) \subset [m]$*;*

(b) *For each* $i \in \mathcal{S}(\mathbf{p})$*, consider the prediction set* $\mathcal{C}_{n+i}^{\alpha|\mathcal{S}(\mathbf{p})|/m}(\mathbf{p})$ *for* Y_{n+i}*, computed according to* (4)*.*

The key theoretical difficulty in proving that the FCR of `InfoSP` is at most α, is that the conformal p-values are dependent, and that the selection step is also p-value-based (by contrast with Bao et al., 2024). Thus, the error rate in (13) for the iid model and (14) for the class-conditional model may not be controlled. Bates et al. (2023); Marandon et al. (2024) showed that, for outlier detection, the special positive dependency between the conformal p-values is such that the BH procedure remains valid for FDR control. Their results are on a different set of conformal p-values, but we develop a similar result for our problem, which enables us to establish the desired error guarantee for various selection rules.

THEOREM 3.1. *Consider score functions satisfying Assumption 1, an informative subset collection* $\mathcal{I}$ *satisfying Assumption 2 and a p-value collection* $\mathbf{p}$ *being either* $\overline{\mathbf{p}}$ *(full-calibrated) or* $\widetilde{\mathbf{p}}$ *(class-calibrated), then the* $\mathcal{I}$*-informative selective prediction set procedure* $\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\mathbf{p})$ *(Definition 1) controls the FCR at level* α*, respectively in the iid model (*$\mathbf{p} = \overline{\mathbf{p}}$*) with the control* (15) *or the class-conditional model (*$\mathbf{p} = \widetilde{\mathbf{p}}$*) with the control* (16)*.*

A proof is provided in § E.1, which relies on a more general result, see Theorem C.1. Theorem C.1 provides the FCR guarantee to more general p-value collections and more general selection rules than those satisfying Assumptions 4 and 5. In particular, our result applies beyond informative selection (e.g., selection by p-value thresholding) and for the two p-value collections $\overline{\mathbf{p}}$ and $\widetilde{\mathbf{p}}$ under less restrictive conditions that those considered in Theorem 3.1: the independence assumption can be relaxed to an exchangeable assumption (Propositions C.2 and C.3), while the score function can take a general form that can provide an extra improvement (Assumption 6).

We apply Lemma 2.1 to obtain the following FDR guarantee for `InfoSP`.

COROLLARY 3.2. *In the setting of Theorem 3.1, the selection rule* $\mathrm{BH}(\mathbf{q})$ *of* $\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\mathbf{p})$ *provides level* α *FDR control, given either by* (19) *in the iid model (*$\mathbf{p} = \overline{\mathbf{p}}$*) or* (20) *in the class-conditional model (*$\mathbf{p} = \widetilde{\mathbf{p}}$*).*

REMARK 3.1. *For* `InfoSP`*, we can avoid the computation of* $\mathbf{q}$ *by using that the BH procedure is the iterative limit of a recursion (Gao et al., 2023). Indeed, since* $\mathcal{C}_{n+i}^{\alpha k/m} \in \mathcal{I}$ *if and only if* $q_i \leq \alpha k/m$*, the selection* $S = \mathrm{BH}(\mathbf{q})$ *can be obtained as follows: Step 1:* $\mathcal{S}_1 = \{i \in [m] \,:\, \mathcal{C}_{n+i}^{\alpha} \in \mathcal{I}\}$*; Step* $t \geq 1$*:* $\mathcal{S}_t = \{i \in \mathcal{S}_{t-1} \,:\, \mathcal{C}_{n+i}^{\alpha|\mathcal{S}_{t-1}|/m} \in \mathcal{I}\}$*; Consider* t_0 *the first* t *where* $\mathcal{S}_t = \mathcal{S}_{t-1}$ *and let* $\mathcal{S} = \mathcal{S}_{t_0}$*.*

3.2. Informative selective conditional prediction sets (`InfoSCOP`)

Throughout this section, we consider the iid model. It turns out that `InfoSP` can be too conservative in some contexts; this is manifest in the true FCR being much smaller

than the nominal α level, see illustrations in § 4 and § 5. To avoid this, we can adapt the conditional approach of Bao et al. (2024) to our framework and start by selecting test samples and calibration samples. We would like, following the initial selection, to have as few as possible test samples for which $\mathcal{I}$-informative prediction sets cannot be constructed. For this purpose, we further split the calibration sample, which is a classical trick in conformal literature in order to preserve exchangeability with calibration samples after the initial selection (see Lemma F.4). Specifically, we follow the following steps

(a) Split the calibration sample $((X_j, Y_j), j \in [n])$ into two samples $((X_j, Y_j), j \in [r])$ and $((X_j, Y_j), j \in [r+1, n])$ for some $r \in [n-1]$.

(b) Apply an initial conformal selection rule $\mathcal{S}^{(0)} = \mathcal{S}^{(0)}((X_j, Y_j)_{j\in[r]}, (X_j)_{j\in[r+1,n+m]}) \subset [r+1, n+m]$ that considers as calibration sample $((X_j, Y_j), j \in [r])$ and test sample $((X_j, Y_j), j \in [r+1, n+m])$.

(c) For $i + n \in \mathcal{S}^{(0)} \cap [n+1, n+m]$, compute the conformal p-values using calibration set $\{(X_j, Y_j), j \in \mathcal{S}^{(0)} \cap [r+1, n]\}$ (i.e., using the conditional empirical distribution, post initial selection):

$$\overline{p}_i^{(0,y)} = \frac{1}{|\mathcal{S}^{(0)} \cap [r+1, n]| + 1} \left(1 + \sum_{j \in \mathcal{S}^{(0)} \cap [r+1,n]} \mathbf{1}\{S_{Y_j}(X_j) \geq S_y(X_{n+i})\} \right). \quad (22)$$

We assume that the initial selection $\mathcal{S}^{(0)}$ satisfies the following permutation preserving assumption.

Assumption 3. *For any permutation σ of $[r+1, n+m]$,*

$$\mathcal{S}^{(0)}((X_j, Y_j)_{j\in[r]}), (X_{\sigma(j)})_{j\in[r+1,n+m]}) = \sigma\big(\mathcal{S}^{(0)}((X_j, Y_j)_{j\in[r]}, (X_j)_{j\in[r+1,n+m]})\big).$$

The initial selection $\mathcal{S}^{(0)}$ is typically the result of a multiple testing procedure that is applied to the examples in $[r+1, n+m]$, that uses p-values computed with $((X_j, Y_j), j \in [r])$ as calibration sample and $(X_i, i \in [r+1, n+m])$ as covariate test sample, which immediately satisfies Assumption 3. We can use $\mathrm{BH}(\mathbf{q})$ as an initial selection stage (where the q_i's are computed with the aforementioned sample split) so that selected examples are likely to correspond to Y_{n+i} where an informative prediction set can be built. For excluding a null range in regression or a null class in classification, another choice is to use an appropriate BH procedure for testing that the examples from $[r+1, n+m]$ are from that null, see examples in § 4.1 and § 5.2.

Definition 2. *The informative selective conditional prediction set procedure pre-processed with the initial selection rule $\mathcal{S}^{(0)}$, called* `InfoSCOP`*, is defined as the* `InfoSP` *procedure of Definition 1 applied with the pre-processed p-value family $\overline{\mathbf{p}}^0$ (22), that is, $\mathcal{R}_\alpha^{\mathrm{InfoSCOP}}(\overline{\mathbf{p}}) = \mathcal{R}_\alpha^{\mathrm{InfoSP}}(\overline{\mathbf{p}}^0) = (\mathcal{C}_{n+i}^{\alpha^0}(\overline{\mathbf{p}}^0))_{i\in\mathrm{BH}(\mathbf{q}^0)}$, where $\alpha^0 = \alpha|\mathrm{BH}(\mathbf{q}^0)|/|\mathcal{S}^{(0)} \cap [n+1, n+m]|$ and the $\mathcal{I}$-adjusted p-values $\mathbf{q}^0$ are computed via (11) from the pre-processed p-values $\overline{\mathbf{p}}^0$.*

THEOREM 3.3. *Consider the iid model (both for regression and classification), score functions satisfying Assumption 1, an informative subset collection $\mathcal{I}$ satisfying Assumption 2 and any initial selection rule $\mathcal{S}^{(0)} \subset [r+1, n+m]$ that satisfies Assumption 3. Then the* `InfoSCOP` *procedure of Definition 2 is such that* $\mathrm{FCR}(\mathcal{R}_\alpha^{\mathrm{InfoSCOP}}(\overline{\mathbf{p}})) \leq \alpha$. *In addition, the associated selection rule* $\mathrm{BH}(\mathbf{q}^0)$ *controls the FDR* (19) *at level* α.

The proof, provided in § E.2, follows directly from the general calibration splitting trick Lemma F.4.

As we will see in the next sections, while it maintains the FCR guarantee, `InfoSCOP` can greatly improve over `InfoSP`. The main reason is that adjusting for selection is cheaper after the initial selection step: by reducing the fraction of examples in the test sample for which it is not possible to construct $\mathcal{I}$-informative prediction sets, the correction term α^0 is expected to be close to α (or not much smaller than α). Another reason is that the selection-conditional p-values $\overline{\mathbf{p}}^0$ will be better in settings where the initial selection step tends to remove the examples from the calibration set that have large non-conformity scores, see § 4, § G, and § H for such cases.

In general, the way `InfoSCOP` can improve over `InfoSP` depends on the context. For instance, for excluding a null range in classification, a null class in regression or for length restriction in regression, we show in § 4 and § 5, respectively, the great potential advantage of using the initial selection. On the other hand, we also show that for non-trivial classification, there may be no advantage of initial selection (in fact, there can be a slight disadvantage since the calibration sample after initial selection is smaller, as demonstrated in § 5.3).

4. Application to regression

This section is devoted to the regression case (that is, $\mathcal{Y} = \mathbb{R}$), as already introduced in Example 1.1. Throughout the section, we work in the iid model. Illustrations for other informative selections can be found in § G in the SM.

4.1. Excluding $[a, b]$ from the prediction interval

We focus here on the case where the user wants to build prediction intervals only for outcomes such that $Y_{n+i} < a$ or $Y_{n+i} > b$, which corresponds to excluding $\mathcal{Y}_0 = [a, b]$ from the prediction interval, where $a < b$ are two benchmark values. This corresponds to common practice where users are interested only in reporting prediction intervals for individuals with "abnormal" outcomes. Setting $a = -\infty$ recovers the case where we only want to report prediction intervals for examples such that $Y_{n+i} > b$, which is the selection considered in Jin and Candes (2023). We focus on two-sided prediction intervals here (the case of one-sided prediction intervals is postponed to § G). The choice of score function defines `InfoSP` and entails all the desired inferential guarantees. We formalize this for the locally weighted score function in Corollary 4.1. Using other score functions is also possible, e.g., the score function that corresponds to conformal quantile regression, see Remark 4.1.

COROLLARY 4.1. *Consider the iid model in the regression case, the locally weighted score function* $S_y(x) = |\mu(x) - y|/\sigma(x)$ *and suppose that Assumption 1 holds. Then the*

following holds for `InfoSP` *with informative collection* $\mathcal{I} = \{I \text{ interval of } \mathbb{R} : I \cap [a,b] = \emptyset\}$ *and full-calibrated p-value collection* $\overline{\mathbf{p}}$ (7)*:*

(i) `InfoSP` *selects* $\mathcal{S} = \mathrm{BH}(\mathbf{q})$ *with*

$$q_i = \overline{p}_i^{(a)} \mathbf{1}\{\mu(X_{n+i}) < a\} + \overline{p}_i^{(b)} \mathbf{1}\{\mu(X_{n+i}) > b\} + \mathbf{1}\{a \leq \mu(X_{n+i}) \leq b\}. \tag{23}$$

(ii) The selection $\mathcal{S}$ *of* `InfoSP` *controls the FDR at level* α *in the following sense:*

$$\sup_{P_{XY}} \mathbb{E}_{(X,Y)\sim P_{XY}} \left[\frac{\sum_{i\in\mathcal{S}} \mathbf{1}\{Y_{n+i} \in [a,b]\}}{1 \vee |\mathcal{S}|} \right] \leq \alpha.$$

(iii) The selected prediction intervals do not contain $[a,b]$ *and are of the form* $\mathcal{C}_{n+i} = [\mu(x) - S_{(n_\alpha(\mathbf{p}))}\sigma(x), \mu(x) + S_{(n_\alpha(\mathbf{p}))}\sigma(x)]$*, where* $S_{(1)} \leq \cdots \leq S_{(n)}$ *are the ordered calibration scores* $S_{Y_j}(X_j)$*,* $1 \leq j \leq n$ *(with* $S_{(n+1)} = +\infty$*), and* $n_\alpha(\mathbf{p}) = \lceil (1 - \alpha|\mathcal{S}(\mathbf{p})|/m)(n+1) \rceil$*.*

(iv) These prediction intervals control the FCR at level α *in the sense of* (15)*.*

Proof. Point (i) follows from Example 2.1 because $\min_{y\in[a,b]} S_y(x)$ is 0 if $\mu(x) \in [a,b]$, $S_a(x)$ if $\mu(x) < a$, and $S_b(x)$ if $\mu(x) > b$. Point (ii) follows from Lemma 2.1, (iii) from Remark 2.1 and the fact that $q_i \leq \alpha|\mathcal{S}(\mathbf{p})|/m$ iff $\mathcal{C}_{n+i}$ does not contain $[a,b]$. Point (iv) follows from Theorem 3.1.

Hence, our method, in addition to providing an FCR control on the selected (iv) ensures that the obtained prediction intervals are informative in the sense that they do not include benchmark values (i.e., values in $[a,b]$). This ensures that the selection method is meaningful for the considered aim, which formally entails the FDR control (ii). Obviously, since `InfoSCOP` is an `InfoSP` method for preprocessed p-values (22), a similar result holds for `InfoSCOP`, for any initial selection step $\mathcal{S}^{(0)} \subset [r+1, n+m]$ that satisfies the permutation preserving Assumption 3.

Corollary 4.1 is illustrated on Figure 2 when $\mathcal{S}^{(0)}$ is taken here as $\mathrm{BH}(\mathbf{q})$ at level 2α (with the score $S_y(x) = |\mu(x) - y|/\sigma(x)$). In the first row, errors are larger further away from $[a,b]$. Hence, while the marginal prediction intervals control the FCR at level α when selecting all the covariates (as granted by (9)), the FCR is inflated for a naive selection that selects each example with a prediction interval at level α not intersecting $[a,b]$ (that is, naive selection is given by $\mathcal{S}_1$ in the recursion of Remark 3.1). To maintain the FCR control at level $\alpha = 0.1$, `InfoSP` adjusts the width of the prediction interval in an accurate way to accommodate the informative constraint. `InfoSCOP` is roughly the same as `InfoSP` in this case, because the largest scores are kept in the calibration sample after initial selection. This is the most unfavorable setting for `InfoSCOP`, but it nevertheless performs similarly to `InfoSP`. In the second row, errors are smaller further away from $[a,b]$, which makes the FCR of the naive selection and `InfoSP` far too conservative. By contrast, the initial selection of `InfoSCOP` removes the largest scores of the calibration sample, resulting in much narrower prediction intervals, and thus in a much larger resolution-adjusted power (even better than the naive procedure).

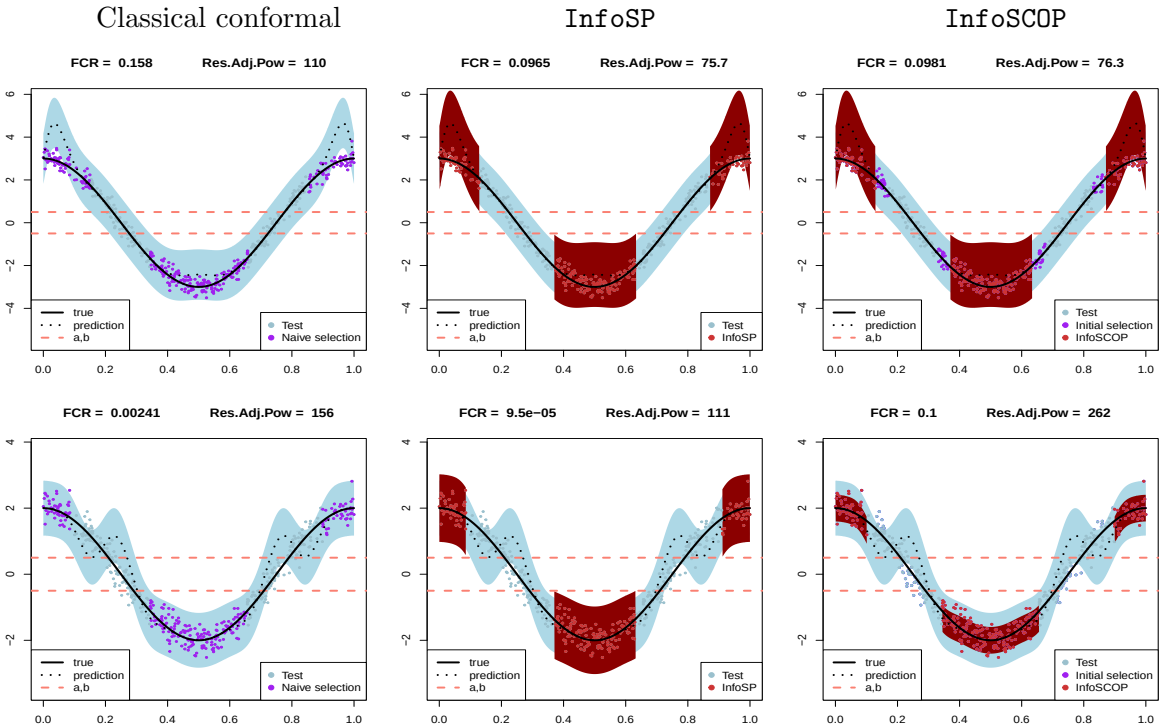


Fig. 2: Informative prediction intervals when excluding $[a, b]$ (homoscedastic Gaussian regression model with perfect variance prediction), see text. The predictor μ (dotted line) does not approximate well the true $\mu^*(x) = \mathbb{E}[Y|X = x]$ (solid line) in the selection area (top row) and out of the selection area (bottom row). The marginal and informative prediction intervals (`InfoSP` and `InfoSCOP`) are depicted in light blue and red, respectively. While the plot corresponds to one data generation, the FCR and adjusted power computed in the title of each panel are computed with 100 Monte-Carlo simulations. $n = 1000$, $m = 500$, $\alpha = 0.1$.

Remark 4.1. *In Corollary 4.1, we consider the locally weighted score function* $S_y(x) = |\mu(x) - y|/\sigma(x)$ *for simplicity of exposition, but we can use any score function satisfying Assumption 1. For instance, for the quantile-based score function* $S_y(x) = \max(q_{\beta_0}(x) - y, y - q_{\beta_1}(x))$*, the corresponding* q_i *have the expression* $q_i = \overline{p}_i^{(a)}\mathbf{1}\{\mu(X_{n+i}) < a\} + \overline{p}_i^{(b)}\mathbf{1}\{\mu(X_{n+i}) > b\} + \overline{p}_i^{(\mu(X_{n+i}))}\mathbf{1}\{a \leq \mu(X_{n+i}) \leq b\}$*, where* $\mu(x) = (q_{\beta_0}(x) + q_{\beta_1}(x))/2$ *and where the* $\overline{p}_i^{(y)}$*'s are computed by using this score function. This leads to the prediction intervals* $\mathcal{C}_{n+i} = [q_{\beta_0}(X_{n+i}) - S_{(n_\alpha(\mathbf{p}))}, q_{\beta_1}(X_{n+i}) + S_{(n_\alpha(\mathbf{p}))}]$ *(which do not contain* $[a, b]$*), by using the notation of Corollary 4.1.*

4.2. Length-restricted prediction intervals

In this section, we consider the situation where the user only wants to report prediction intervals that are accurate enough, which corresponds to consider $\mathcal{I} = \{[a, b] \subset \mathbb{R} : 0 <$

$b - a \leq 2\lambda_0\}$ as the informative subset collection, for some size $\lambda_0 > 0$.

COROLLARY 4.2. *Consider the iid model in the regression case, consider the locally weighted score function $S_y(x) = |\mu(x) - y|/\sigma(x)$ and suppose that Assumption 1 holds. Then the following holds for* `InfoSP` *with informative collection $\mathcal{I} = \{[a, b] \subset \mathbb{R} : 0 < b - a \leq 2\lambda_0\}$ and full-calibrated p-value collection $\overline{\mathbf{p}}$ (7):*

(i) `InfoSP` *selects $\mathcal{S} = \mathrm{BH}(\mathbf{q})$ with q_i given by the formula of Example 2.1.*

(ii) The selected prediction intervals are of length at most $2\lambda_0$ and are of the form $\mathcal{C}_{n+i} = [\mu(x) - S_{(n_\alpha(\mathbf{p}))}\sigma(x), \mu(x) + S_{(n_\alpha(\mathbf{p}))}\sigma(x)]$, where $S_{(1)} \leq \cdots \leq S_{(n)}$ are the ordered calibration scores $S_{Y_j}(X_j)$, $1 \leq j \leq n$ (with $S_{(n+1)} = +\infty$), and $n_\alpha(\mathbf{p}) = \lceil (1 - \alpha|\mathcal{S}(\mathbf{p})|/m)(n + 1)\rceil$.

(iii) These prediction intervals control the FCR at level α in the sense of (15).

A similar result holds for `InfoSCOP`. In Corollary 4.2 (ii), the length of the prediction interval on the selection is always granted to be (at most) of the correct size $2\lambda_0$, even if adjusting the level is necessary to account for selection (which de facto enlarge the prediction interval). Thanks to the $\mathrm{BH}(\mathbf{q})$ selection the size-adjustment is automatic, while maintaining the FCR control.

PROOF. By Remark 2.1, we have $q_i \leq \alpha$ iff $\sigma(X_{n+i})S_{(\lceil(1-\alpha)(n+1)\rceil)} \leq \lambda_0$. This implies the expression of q_i and that $|\mathcal{C}_{n+i}| = 2\sigma(X_{n+i})S_{(\lceil(1-\alpha)(n_\alpha(\mathbf{p})+1)\rceil)} \leq 2\lambda_0$ since $q_i \leq \alpha|\mathcal{S}(\mathbf{p})|/m$.

Figure 3 displays length-restricted informative prediction intervals in particular settings. In the first row, errors are more likely to occur on the selection (due to under-estimation of the variance), while in the second row, errors are less likely to occur on the selection (due to over-estimation of the variance). Hence, the comment is similar to the previous section: `InfoSP` and `InfoSCOP` are similar in the first situation but `InfoSCOP` improves `InfoSP` in the second.

REMARK 4.2. *Corollary 4.2 easily extends to the case of conformalized quantile regression, by considering the quantile-based score function $S_y(x) = \max(q_{\beta_0}(x) - y, y - q_{\beta_1}(x))$. In that case, $q_i \leq \alpha$ iff $2S_{(\lceil(1-\alpha)(n+1)\rceil)} + q_{\beta_1}(X_{n+i}) - q_{\beta_0}(X_{n+i}) \leq 2\lambda_0$ which leads to the formula of Example 2.1 for the q_i's and to the prediction intervals $\mathcal{C}_{n+i} = [q_{\beta_0}(X_{n+i}) - S_{(n_\alpha(\mathbf{p}))}, q_{\beta_1}(X_{n+i}) + S_{(n_\alpha(\mathbf{p}))}]$ (of length at most $2\lambda_0$), with the notation of Corollary 4.2.*

5. Application to classification

We consider the classification case $\mathcal{Y} = [K]$, for both the iid model and the class-conditional model. Importantly, in classification, any selective prediction set procedure $\mathcal{R} = (\mathcal{C}_{n+i})_{i\in\mathcal{S}}$ is post-processed by setting, for $i \in \mathcal{S}$, $\mathcal{C}_{n+i} = \arg\min_{k\in[K]}\{S_k(X_{n+i})\}$ whenever $\mathcal{C}_{n+i} = \emptyset$ (that is, if empty take the smallest non-conformity score). Clearly, this operation can only decrease the FCP while it can only increase the adjusted power, so it should always be preferred in the classification case. In this paper, `InfoSP` and `InfoSCOP` always refer to the post-processed procedures in the classification case.

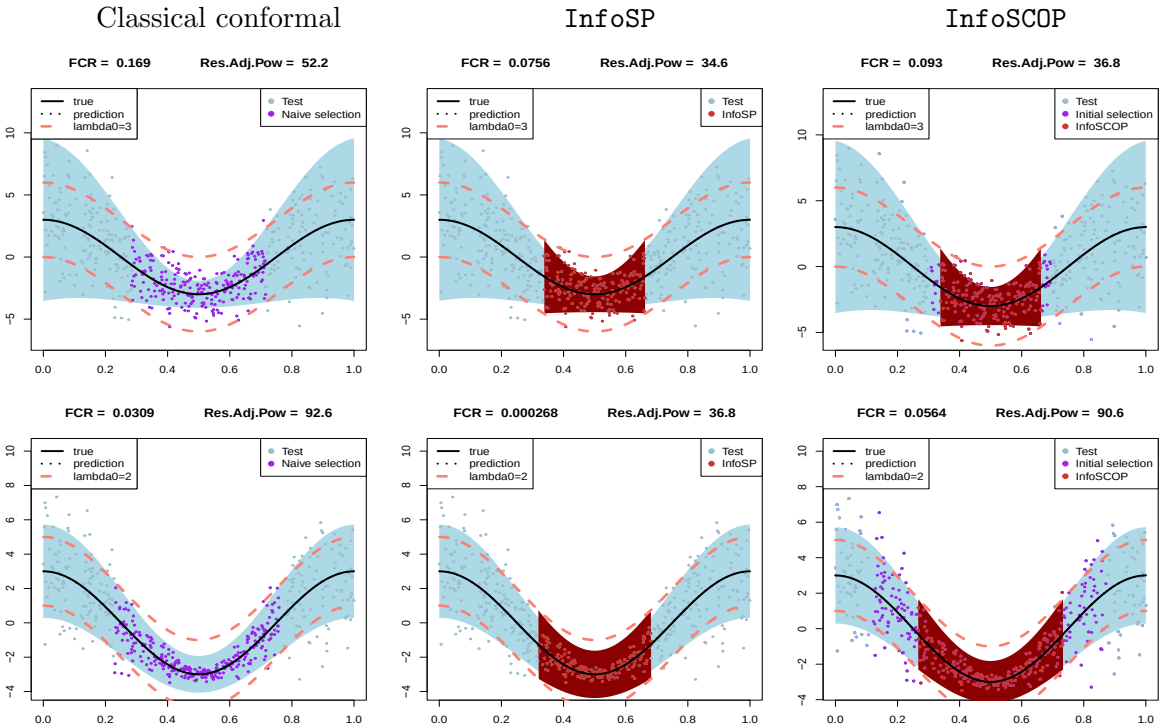


Fig. 3: Informative prediction intervals when length-restricted (heteroscedastic Gaussian regression model with perfect mean prediction), see text. The predictor σ underestimates (top row) and over-estimates (bottom row) the true $\sigma^*(x) = \mathbb{V}^{1/2}[Y|X = x]$ in the selection area. The marginal and informative prediction intervals (`InfoSP` and `InfoSCOP`) are depicted in light-blue and red, respectively. While the plot corresponds to one data generation, the FCR and adjusted power computed in the title of each panel are computed with 100 Monte-Carlo simulations. $n = 1000$, $m = 500$, $\alpha = 0.1$.

5.1. Choosing the appropriate p-value collection in classification

While the family of full-calibrated p-values are only valid for the iid model, the family of class-calibrated p-values are valid both in the iid and in the class-conditional model.

Thus, for the iid model, we can in principle use either full-calibrated p-values or class-calibrated p-values. In the applications we consider next, it appears that using full-calibrated p-values in `InfoSP` is best. We support this claim by theory for non-trivial classification in § 5.3, and by numerical experiments in § 5.2, § 5.3, and § H. We note that for the initial selection step in `InfoSCOP`, class-calibrated p-values can be useful, as demonstrated in § 5.2.

For the class-conditional model, there can be a label shift from the calibration to the test sample. So the full-calibrated p-values are not valid, and class-calibrated p-values must be used. In § D we consider more generally weighted class-calibrated p-values, where the weights are functions of estimators of the proportion of labels in each class in the test sample.

5.2. An illustrative example: prediction sets excluding a null class

Suppose the analyst is interested in reporting prediction sets that exclude a null class, say class $y_0 = 1$ (see first items of Examples 1.2 and 2.2). We consider the following novel procedures, in addition to the naive procedure using the classic conformal procedure, that reports $\mathcal{C}^{\alpha}_{n+i}$ only if $\mathcal{C}^{\alpha}_{n+i}$ does not contain the null class: first, `InfoSP` on full-calibrated p-values, denoted by $\mathcal{R}^{\mathrm{InfoSP}}_{\alpha}(\overline{\mathbf{p}})$. Second, `InfoSCOP` on full-calibrated p-values, denoted by $\mathcal{R}^{\mathrm{InfoSCOP}}_{\alpha}(\overline{\mathbf{p}})$, with initial selection step $\mathcal{S}^{(0)} \subset [r+1, n+m]$ being the BH procedure applied to the class-calibrated adaptive p-value family $(\tilde{p}^{(1)}_{i,\mathrm{adapt}}, i \in [r+1, n+m])$, given by $\tilde{p}^{(1)}_{i,\mathrm{adapt}} = \hat{\pi}_1 \tilde{p}^{(1)}_i$, $i \in [r+1, n+m]$, using $\{(X_j, Y_j), j \in [r]\}$ and $\{(X_j, Y_j), j \in [r+1, n+m]\}$ as calibration and test samples, respectively, and with calibrated-based estimator $\hat{\pi}_1 = (r+1)^{-1}\big(\sum_{i=1}^{r} \mathbf{1}\{Y_i = 1\} + 1\big)$. Third, `InfoSP` on class-calibrated p-values, denoted by $\mathcal{R}^{\mathrm{InfoSP}}_{\alpha}(\widetilde{\mathbf{p}})$. Note that `InfoSCOP` above uses class-calibrated p-values for initial selection (because these are better to detect examples from the non-null class), and full-calibrated p-values on the selected examples from $[r+1, n+m]$ in the second step (because these are better p-values for building prediction sets in the iid model).

We consider a Gaussian mixture model with $K = 3$ components, where each component is bivariate normal. The centers for the three components are (0,0), (SNR,0), and (SNR,SNR). So the overlap between components is larger as SNR decreases. We consider the case of balanced classes in the calibration sample, as well as the case of unbalanced classes where the null class is much larger than the others. Specifically, the balanced case has class probabilities 0.33, 0.33, and 0.34, and the unbalanced case has class probabilities 0.15, 0.10, and 0.75 (the null class). In the balanced case, we consider the iid setting where the test sample has the same class probabilities as the calibration sample (depicted in Figure 10 for an SNR value of 3), as well as class-conditional setting where the test sample has class probabilities 0.2,0.2, and 0.6 (the null class), so the label shift is large. We estimate the probability of being in each class with a support vector classifier implemented by the *e1071* R package Meyer et al. (2023).

Figure 4 shows the FCR and resolution-adjusted power of all procedures considered. As expected, the classic conformal procedure does not control the FCR for any data generating model (it uses the full-calibrated and class-calibrated p-values in the iid and class-conditional setting, respectively). All other procedures control the FCR.

For the iid model (Figure 4 left and middle columns), `InfoSCOP` on full-calibrated p-values has better power than the alternatives for prediction sets excluding a null class. Its advantage over `InfoSP` is primarily due to the fact that after pre-processing, almost all test examples are non-null, as illustrated in Figure 10 in the SM for a single data generation. The differences between the procedures are qualitatively the same, but even greater, when the overlap between components is larger, see Figure 12 in the SM. This is because when the overlap with the null class is large, after pre-processing only examples with better scores are considered, as illustrated in Figure 13 in the SM. For completeness, we also provide `InfoSP` on class-calibrated p-values in the iid setting, to demonstrate numerically the potentially large power advantage from using the full-calibrated p-values over the class-calibrated p-values.

For the class-conditional model (Figure 4 right column), `InfoSP` has lower power than classic conformal, but the power is reasonable. In Figure 6 in the SM we compare it to

two procedures that weigh the classes according to their estimated relative frequencies.

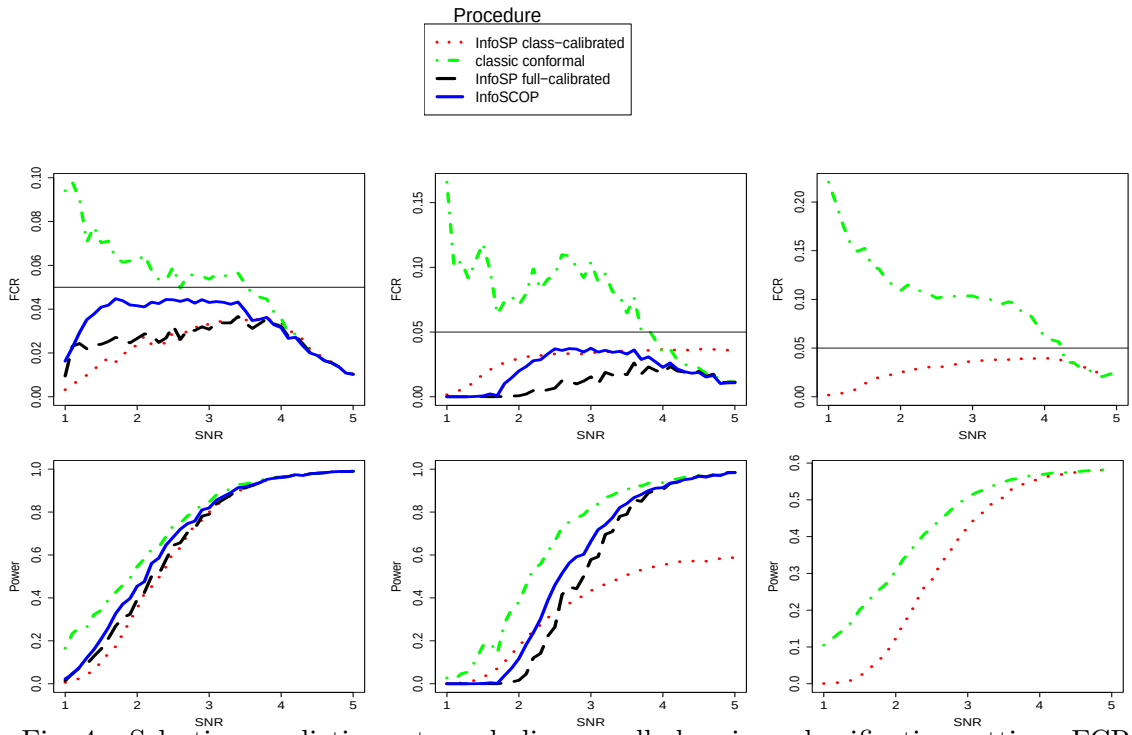


Fig. 4: Selecting prediction sets excluding a null class in a classification setting. FCR (top row), and resolution-adjusted power (bottom row) versus SNR. The iid setting in columns 1 and 2, with balanced classes and unbalanced classes, respectively. The class-conditional setting in column 3, with a large label shift: the class probabilities are equal in the calibration sample and 0.2,0.2, and 0.6 (the null class) in the test sample. The number of data generations was 2000, 1000 data points were used for training, and $n = m = 500$. See details of the data generation in § 5.2.

5.3. Selecting non-trivial prediction sets

Suppose the analyst is interested in reporting prediction sets that are not equal to $[K]$ (see first item of Examples 1.2 and 2.2). In that case, we argue that `InfoSP` has an FCR close to α in the iid model. Intuitively, this comes from the selection rule $\mathcal{S} = \mathrm{BH}(\overline{\mathbf{p}})$ which is such that $\mathbf{1}\{Y_{n+i} \notin \mathcal{C}_{n+i}^{\alpha|\mathcal{S}|/m}(\overline{\mathbf{p}}), i \in \mathcal{S}\} = \mathbf{1}\{Y_{n+i} \notin \mathcal{C}_{n+i}^{\alpha|\mathcal{S}|/m}(\overline{\mathbf{p}})\}$. It means that it is not possible to fail to cover at the adjusted level without being selected (because otherwise the prediction set is trivial). This is not the case for other selection rules, e.g., excluding a null class, where it is possible that the true class label is not covered at the adjusted level even if the example is not selected (thus implying that the adjusted level is conservative, since for FCR control we guard against non-coverage at the adjusted level for all examples). We formalize fully the argument for $K = 2$ in the following result.

PROPOSITION 5.1. *In the iid classification model with $K = 2$, consider the non-trivial informative subset collection $\mathcal{I} = \{C \subset [K] \ : \ |C| \leq 1\}$ and assume that the score functions satisfy Assumption 1 with $\sum_{k\in[K]} S_k(x) = 1$ and $S_k(x) \geq 0$. Let p_0 be the probability that $S_{Y_i}(X_i)$ is the maximum score $\max\{S_1(X_i), S_2(X_i)\}$. Then if $(n+1)\alpha/m$ is an integer, we have $\mathrm{FCR}(\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\overline{\mathbf{p}}), P_{X,Y}) = \alpha(1-(1-p_0)^{n+1})$.*

The proof is given in § E.4, which also shows that $\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\overline{\mathbf{p}})$ coincides with the procedure of Zhao and Su (2023) for $K = 2$. In typical applications (where classes are not very well separated) the value of $(1-(1-p_0)^{n+1})$ is close to one, which means that the FCR of our procedure should be close to α, at least for $K = 2$.

To complement our theoretical result, we provide numerical results in Figure 11. `InfoSP` has the best power, with `InfoSCOP` a close second, on full-calibrated p-values. `InfoSP` on class-calibrated p-values has much lower power in the unbalanced setting. As expected, the classic conformal procedure does not control the FCR and the level for `InfoSP` is about 0.05 for a range of SNR values. All other procedures control the FCR.

6. Informative prediction sets for 3 classes of animals

In this section, we illustrate the performance of our methods on real data. We use the image dataset CIFAR-10 (`https://www.cs.toronto.edu/~kriz/cifar.html`), which consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. We restrict the analysis to 3 classes: birds, cats and dogs.

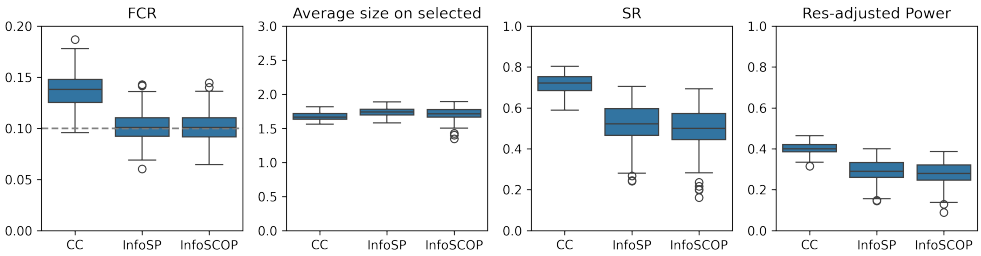
We consider 4 scenarios: in the iid setting, non-trivial classification (scenario a) and non-null classification (scenario b); in the class-conditional setting, non-trivial classification (scenario c) and non-null classification (scenario d). The null class is the bird class in scenarios b and d. By default, the classes are in equal proportions. We introduce a label shift between the calibration sample and test sample for the class-conditional settings, by modifying the classes proportions: of the calibration sample to 20% for the bird class versus 40% for the cat/dog class in scenario c; of the test sample to 50% for the bird class versus 25% for the cat/dog class in scenario d. In each experiment, the test size is $m = 1000$ and the calibration size is $n = 5000$.

In the iid settings (scenarios a and b), we evaluate the procedures `InfoSP` and `InfoSCOP` with full-calibrated p-values and in the class-conditional settings (scenarios c) and d)) we evaluate `InfoSP` with class-calibrated p-values, each being also compared with classical conformal prediction (denoted by `CC`). For all procedures the non-conformity score is $S_y(x) = 1 - \pi_y(x)$ where $\pi_y(x)$ is an estimator of the probability that the class of x is y and is learned using a convolutional neural network (CNN) with 2 convolutional layers, one pooling layer, and 3 fully-connected layers, trained for 20 epochs with a learning rate of 0.01 on a sample size of 5000. To assess the power of the methods, in addition to the resolution-adjusted power, we plot the selection rate (SR) of the procedures, defined as the average proportion of informative prediction sets returned, and the average size of the informative prediction sets.

In each setting, the FDR and each power metric for the methods are evaluated by using 100 runs and the results are reported in Figure 5 scenarios a and b and in Figure 14 scenarios c and d for $\alpha = 0.1$. The conclusions are qualitatively similar to the experiments of § 5: in all settings, classical conformal prediction yields an FCR that severely

exceeds the marginal level with an inflation of about 50%. By contrast, our procedures `InfoSP` and `InfoSCOP` control the FCR at the target nominal level, both without and with label shift. In terms of power, concerning the iid settings, pre-processing is not useful for non-trivial classification as expected and the performances of `InfoSP` and `InfoSCOP` are similar with an FCR close to the nominal level for both in that case. For the non-null classification task, `InfoSP` is conservative while `InfoSCOP` is more powerful and displays an FCR close to α. When there is label shift, in the case of non-trivial classification `InfoSP` displays an FCR close to α. In the non-null classification case, however, the label shift increases the difficulty of the task in the sense that for a fixed selection, under-covering the null class in the test sample results in more false informative sets. Hence, the power is low in that case. Finally, in all scenarios, our procedures `InfoSP` and `InfoSCOP` output a lower number of informative sets compared to `CC`, but this is necessary in order to control the FCR. The average size of the prediction sets that are informative is comparable.

a) Non-trivial classification



b) Non-null classification

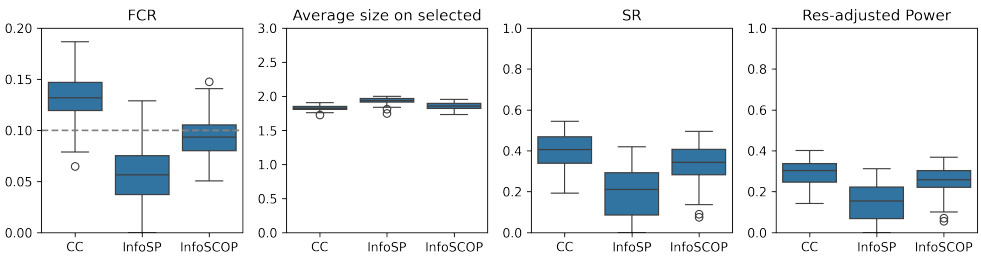


Fig. 5: FCR, average size of the selected, SR, and resolution-adjusted power for the methods, for $\alpha = 0.1$.

7. Conclusion and discussion

In this paper, we have introduced new methods for providing conformal prediction sets after selection with controlled FCR, that impose a user-specific constraint on the produced prediction sets, corresponding to a collection of so-called informative subsets $\mathcal{I}$. In contrast with previous literature in the field, the selection and prediction stages are intertwined, which results in a BH-type selection procedure on adjusted p-values (the

q_i's) that can further be explicitly derived in specific settings and that by definition produces prediction sets satisfying the desired constraint of belonging to $\mathcal{I}$.

Our methods are very general, and they are relevant to applications in classification and regression. We showed examples in § 4-§ 6 for informative subsets of interest. We used common scores for the examples, but many other scores from the literature can be used with our suggested procedures `InfoSP` and `InfoSCOP` as long as Assumption 1 is satisfied. In addition, transfer learning scores can also be handled by our theory (see Assumption 6 in SM). This is known to greatly improve the conformal inference when there is a domain shift between the learning sample and the calibration+test samples (Courty et al., 2017; Gazin et al., 2023) and developing specific application cases for the latter is of interest for future investigations.

For the iid model, `InfoSCOP` improves over `InfoSP` in all considered examples, except when selecting non-trivial prediction subsets in classification (Example 1.2 item 2), for which we establish that `InfoSP` almost exhausts the FCR level (see Proposition 5.1for $K = 2$). The procedure `InfoSCOP` splits the calibration sample in order to apply an efficient initial selection step on part of the calibration sample and the test sample. There are many ways to perform the initial selection. The choice is important because it defines the pre-processed p-values (22) that can be seen as p-values "conditionally on being selected". Different ways of choosing $S^{(0)}$ have been investigated: trying to rule out all the examples in null class (§ 5.2) or trying to mimic the BH($\mathbf{q}$) selection that will be applied at the second stage to reduce the selection effect (§ 4). Finding an optimal way of calibrating $S^{(0)}$ is an interesting avenue for future research.

For non-trivial prediction subsets in classification, `InfoSP` is optimal when $K = 2$ for oracle scores, i.e., $S_k(X_j) = \mathbb{P}(Y_j \neq k \mid X_j), k \in [K], j \in [n+m]$. Specifically, Zhao and Su (2023) showed that their classification procedure, which coincides with `InfoSP` for $K = 2$, is optimal for controlling the expected number of non-covering prediction sets divided by the expected number of selected examples, denoted by mFSR in their paper. An open question is whether `InfoSP` for $K > 2$ is optimal when the scores are oracle scores for the resolution adjusted power objective or a variant thereof. More generally, developing an optimality theory for selective informative prediction sets (for non-trivial prediction sets as well as for other notions informativeness) is of great interest.

We provided a class-conditional variant of `InfoSP`, with class-conditional guarantees. We proved that our strategy can be followed in the case where the classes of the calibration and test samples are arbitrary fixed, even when the class proportions in the calibration are very different than in the test. The main point is that working with the class-calibrated p-value collection allows to maintain the FCR control in this strong sense. In § D we suggest additionally weighted procedures, that incorporate the estimated class proportions. These procedures are not necessarily more powerful than `InfoSP`, and further research is needed in order to make recommendations about when to use the weighted procedures, and about weight adaptation to the specifics of the data.

Acknowledgments

The authors acknowledge grants ANR-21-CE23-0035 (ASCAI) and ANR-23-CE40-0018-01 (BACKUP) of the French National Research Agency ANR, the Emergence project

MARS of Sorbonne Université, and Israeli Science Foundation grant no. 2180/20.

References

Bao, Y., Huo, Y., Ren, H., and Zou, C. (2023). Selective conformal inference with false coverage-statement rate control.

Bao, Y., Huo, Y., Ren, H., and Zou, C. (2024). Selective conformal inference with false coverage-statement rate control. *Biometrika*, page asae010.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2023). Testing for outliers with conformal p-values. *Ann. Statist.*, 51(1):149–178.

Benjamini, Y. and Bogomolov, M. (2013). Selective Inference on Multiple Families of Hypotheses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):297–318.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81.

Blanchard, G. and Roquain, E. (2008). Two simple sufficient conditions for FDR control. *Electron. J. Stat.*, 2:963–992.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. In *Advances in neural information processing systems 30 (NIPS 2017)*, volume 30.

Ding, T., Angelopoulos, A. N., Bates, S., Jordan, M. I., and Tibshirani, R. J. (2023). Class-conditional conformal prediction with many classes. *arXiv preprint arXiv:2306.09335*.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456):1151–1160.

Ferreira, J. A. and Zwinderman, A. H. (2006). On the Benjamini-Hochberg method. *Ann. Statist.*, 34(4):1827–1849.

Gao, Z., Hu, W., and Zhao, Q. (2023). A constructive approach to selective risk control. *arXiv preprint arXiv:2401.16651*.

Gazin, U., Blanchard, G., and Roquain, E. (2023). Transductive conformal inference with adaptive scores. *arXiv preprint arXiv:2310.18108*.

Guo, W. and Romano, J. (2015). On stepwise control of directional errors under independence and some dependence. *Journal of Statistical Planning and Inference*, 163:21–33.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496.

Jin, Y. and Candes, E. J. (2023). Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41.

Jin, Y. and Ren, Z. (2023). Confidence on the focal: Conformal prediction with selection-conditional coverage. *arXiv preprint arXiv:2403.03868*.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *J. Amer. Stat. Assoc.*, 113(523):1094–1111.

Lei, J., Rinaldo, A., and Wasserman, L. (2014). A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2024). Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157–183.

Marandon, A., Rebafka, T., Roquain, E., and Sokolovska, N. (2022). False clustering rate control in mixture models. *arXiv preprint arXiv:2203.02597*.

Mary-Huard, T., Perduca, V., Martin-Magniette, M.-L., and Blanchard, G. (2022). Error rate control for classification rules in multiclass mixture models. *The international journal of biostatistics*, 18(2):381–396.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2023). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-13.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *13th European Conference on Machine Learning (ECML 2002)*, pages 345–356. Springer.

Ramdas, A. K., Barber, R. F., Wainwright, M. J., and Jordan, M. I. (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*, 47(5):2790–2821.

Rava, B., Sun, W., James, G. M., and Tong, X. (2021). A burden shared is a burden halved: A fairness-adjusted approach to classification. *arXiv preprint arXiv:2110.05720*.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.*, 100(469):94–108.

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

Roquain, E. and Villers, F. (2011). Exact calculations for false discovery proportion with application to least favorable configurations. *Ann. Statist.*, 39(1):584–612.

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

Sarkar, S. K. (2008). On methods controlling the false discovery rate. *Sankhya, Ser. A*, 70:135–168.

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

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):479–498.

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

Weinstein, A. and Ramdas, A. (2020). Online control of the false coverage rate and false sign rate. *Proceedings of the 37th International Conference on Machine Learning, PMLR*, 119:10193–10202.

Weinstein, A. and Yekutieli, D. (2020). Selective sign-determining multiple confidence intervals with FCR control. *Statistica Sinica*, 30:531–555.

Zhao, G. and Su, Z. (2023). Controlling FSR in Selective Classification. *arXiv preprint arXiv:2311.03811*.

A. Connections to existing works

For the class-conditional model, we can view $(Y_{n+i})_{i\in[m]}$ as fixed. Thus, informative prediction sets can be viewed as informative confidence sets for parameters. This has been considered in a particular setting by Weinstein and Yekutieli (2020). They considered building confidence intervals only for the selected parameters that will be sign determining. They showed that if the test statistics are independent and the confidence intervals satisfy some monotonicity properties, then the FCR can be controlled. Their theoretical framework is different than the one we consider, but their approach of selecting only sign-determining confidence intervals is very similar to ours, of selecting only informative prediction sets when informativeness is defined by sign determination. Moreover, this approach has been considered in Weinstein and Ramdas (2020) with the broader scope of only reporting confidence intervals if they are "localizing" appropriately the true parameter in the sense that the confidence interval is entirely contained in one element of a pre-specified partition of the $\mathcal{Y}$ space. They investigate this task in the online setting where the sequence of unknown parameters is fixed, and at each time step an independent observation is observed for the corresponding parameter, which is substantially different than our batch setting where the conformal p-values are dependent and the outcome may be random. We show in Remark 2.3 that our informativeness theory covers their localizing notion in our setting. Next, we discuss inspiring works connected to ours that assume the iid model.

Zhao and Su (2023) suggested procedures for average error control in multi-class classification, so $\mathcal{C}_{n+i}$ are singletons. Since their procedure only reports a single class for each selected example, ambiguous examples will not be selected. However, in most classification tasks, there are examples whose true class is difficult to determine, yet it is possible to narrow down the possible set of classes (Sadinle et al., 2019). We suggest procedures that produce $\mathcal{C}_{n+i}$ that are not necessarily singletons for $K > 2$. For $K = 2$, their procedure coincides with an instance in our framework, see details in § 5.3. However, for $K > 2$, while our suggestion as well as their suggestion provides level α FCR control, we select more examples, and although the prediction sets may be at a coarser resolution than singletons, they are still informative since they narrow down the possible set of classes. We note that for $K > 2$, we can recover their procedure if we define as informative only prediction sets of size one, since then their procedure coincides with `InfoSP` for the iid model.

Jin and Candes (2023) addressed the problem of discovering outcomes with values above a threshold. So $\mathcal{C}_{n+i}$ is of the form (c_i, ∞) for predefined $(c_i)_{i\in[m]}$. They cast the problem as that of testing the family of null hypotheses $\{Y_{n+i} \leq c_i, i \in [m]\}$. In § G.2, it is demonstrated that by defining $Y_{n+i} \leq c_i$ as uninformative, we can complement the discoveries of Jin and Candes (2023) with one-sided prediction intervals, while providing the same false discovery rate guarantee on the selected. Moreover, we show how to obtain two-sided prediction intervals for the informative examples in § 4.1.

Bao et al. (2024) considered the regression framework. Their first result (Proposition 1 therein) is to prove that for selection rules that do not depend on the calibration sample, classic conformal prediction intervals at level $\alpha|\mathcal{S}|/m$ for the $\mathcal{S}$ selected examples (i.e., the correction factor $|\mathcal{S}|/m$ for selection suggested in Benjamini and Yekutieli, 2005), provide level α FCR control. For our purpose of informative selection, this result is not useful because informative selection involves all conformal p-values and therefore involves the calibration sample in a specific way. We need a different set of assumptions that are detailed in our novel Theorem C.1.

Bao et al. (2024) further argue that the resulting prediction intervals, $\left(\mathcal{C}_{n+i}^{\alpha|\mathcal{S}|/m}\right)_{i\in\mathcal{S}}$ are too wide. They suggest a novel approach that performs selection on both the calibration set and test set, and then constructs α level conformal prediction intervals for the selected test candidates using the conditional empirical distribution obtained by the post-calibration set. For exchangeable selection rules, they show that the FCR is controlled at level α. Their selection process cannot guarantee that all the constructed prediction intervals are of interest to the analyst. For example, for predicting the affinity of drug-target pairs, the analyst may not be interested in pairs with affinity below, say, y_0 (the case in item 1 of Example 1.1). Using the novel approach of Bao et al. (2024), prediction intervals will be constructed following the selection of calibration and test examples for which the predicted affinity from the machine learning algorithm is above a selection threshold. They require that the selection procedure be a thresholding procedure of the scores $S_{Y_j}(X_j)$ with a threshold τ that is either independent of the calibration sample (their Proposition 1), or exchangeable with respect to both calibration and test samples (their Theorem 1). Some of these prediction intervals may include y_0, and thus be useless for the analyst (an illustration is given in § G.3 in SM, see Figure 9 therein). However, it is not possible to additionally select only the examples with pre-

diction intervals above y_0, since after performing this additional selection, the prediction intervals of Bao et al. (2024) on the selected no longer have an α level FCR guarantee. Our procedures for regression thus complement the work of Bao et al. (2024) when the focus is that all the prediction intervals eventually constructed are informative.

In a very recent work¶, Jin and Ren (2023) generalize the work of Bao et al. (2024), by considering more general selection rules for finite-sample exact coverage conditional on the unit being selected. Their conditional guarantee is achieved by a careful swapping argument which identifies for each selection rule, the appropriate subset of the calibration examples for each example from the test sample. Their conditional error guarantee implies FCR control under some conditions. They provide various examples where selection takes place and prediction sets on the selected are of interest. Even though their selection may be based on some notion of informativeness, like in Bao et al. (2024) the final prediction sets that are constructed can violate this notion. For example, after selecting by a multiple testing procedure on the family of null hypotheses $\{Y_{n+i} \leq c_i, i \in [m]\}$ in the setting of Jin and Candes (2023), their prediction sets may include the c_i's for some of the discoveries. We suggest procedures where selection and construction of prediction sets are inseparable, since we require that each selected prediction set be informative (along with the requirement that FCR $\leq \alpha$).

B. Application to directional FDR control

Consider for this section that we have $(X_i, Z_i)_{i\in[n+m]}$ with real-valued outcomes $Z_i \in \mathbb{R}$, and we aim at excluding $Z_{n+i} \in [a, b]$, as well as at deciding whether $Z_{n+i} < a$ or $Z_{n+i} > b$ (without producing prediction intervals for the Z_{n+i}), for two benchmark values $a \leq b$. More formally, we want to build a selection $\mathcal{S} \subset [m]$ and a (point-wise) decision $\widehat{Y} \in \{1, 3\}$ from the observed samples such that

$$\mathrm{FDR}_{\mathrm{dir}}(\mathcal{S}, \widehat{Y}) := \sup_{P_{X|Y},Y} \mathbb{E}_{X\sim P_{X|Y}} \left[\frac{\sum_{i\in\mathcal{S}} \mathbf{1}\{Y_{n+i} \neq \widehat{Y}_{n+i}\}}{1 \vee |\mathcal{S}|} \right] \leq \alpha, \tag{24}$$

where $Y_j = \mathbf{1}\{Z_j < a\} + 2\mathbf{1}\{Z_j \in [a, b]\} + 3\mathbf{1}\{Z_j > b\}$, $j \in [n + m]$. Our theory yields the following result.

Corollary B.1. *Consider the class-conditional (classification) model on $(X_i, Y_i)_{i\in[n+m]}$ with $K = 3$ classes. Consider any score function satisfying Assumption 1 (in this classification model). Consider the procedure that selects $\mathcal{S} = \mathrm{BH}(\mathbf{q})$ with*

$$q_i = \max(\tilde{p}_i^{(2)}, \min(\tilde{p}_i^{(1)}, \tilde{p}_i^{(3)})),\ i \in [m],$$

where $\tilde{p}_i^{(y)}$ are the class-conditional p-values computed as in (8)*, and with the decision*

$$\widehat{Y}_{n+i} = \mathbf{1}\{S_1(X_{n+i}) \geq S_3(X_{n+i})\} + 3\mathbf{1}\{S_3(X_{n+i}) > S_1(X_{n+i})\}, i \in \mathcal{S}.$$

Then this procedure controls the directional FDR at level α in the sense of (24)*.*

¶This work appeared when we were in the final stage of writing, our work has been done independently.

PROOF. We consider $\mathcal{I} = \{C \subset [K] \,:\, y_0 \notin C, |C| \leq 1\}$ for $y_0 = 2$ (see Example 2.3) in the classification setting based on the sample (X_j, Y_j)'s ($K = 3$), and we note that the FCR coincides with the directional FDR in that case. Hence, the result comes directly from Theorem 3.1 (note that `InfoSP` is post-processed here, see § 5).

REMARK B.1. *For $b = a$ and continuous outcomes (with, say, $\tilde{p}_i^{(2)} = 0$), there are almost surely only two classes, $Z_i = 1$ corresponding to $Y_i < a$ and $Z_i = 3$ corresponding to $Y_i > a$. The procedure is thus* `InfoSP` *for non-trivial classification, with class-calibrated p-values for $K = 2$. This procedure coincides with the directional FDR procedure in Guo and Romano (2015) applied to conformal p-values for the parameters $Y_{n+1}, \ldots, Y_{n+m}$. The proof of validity in Guo and Romano (2015) assumes that the test statistics for the m hypotheses are independent. Interestingly, with the dependence induced by the conformal p-values, the same procedure is still valid, as formalized in Corollary B.1.*

REMARK B.2. *If one wants to obtain prediction intervals in addition to the directional FDR control, it turns out that in the setting of Corollary 4.1, not only the FDR control (ii) holds but also the directional FDR control*

$$\mathrm{FDR}_{dir}(\mathcal{R}) := \sup_{P_{X,Y}} \mathbb{E}_{(X,Y)\sim P_{X,Y}} \left[\frac{\sum_{i\in\mathcal{S}} \mathbf{1}\{D_i = 1, Y_{n+i} > a\} + \mathbf{1}\{D_i = 3, Y_{n+i} < b\}}{1 \vee |\mathcal{S}|} \right] \leq \alpha,$$

for the procedure $\mathcal{R} = (\mathcal{C}_{n+i})_{i\in\mathcal{S}}$ defined therein with the directional rule $D_i = \mathbf{1}\{\mathcal{C}_{n+i} \subset (-\infty, a)\} + 3\mathbf{1}\{\mathcal{C}_{n+i} \subset (b, +\infty)\}$ for $i \in \mathcal{S}$. It can be slightly less powerful than the directional FDR controlling procedure of Corollary B.1 (because the latter uses classification scores), but provides additional information.

C. FCR control for BY-type selection rules

In this section, we present a general approach for FCR control (§ C.1), which relies on the following:

- a general class of p-values, including both full-calibrated and class-calibrated conformal p-values (§ C.3);
- a general selection rule, including informative selection rules (§ C.4).

The assumption on the selection rule is inspired by Benjamini and Yekutieli (2005); Benjamini and Bogomolov (2013).

C.1. General statement

Let $m \geq 1$, $\mathcal{Y} \subset \mathbb{R}$ and consider a family $\mathbf{p} = (p_i^{(y)}, i \in [m], y \in \mathcal{Y})$ of random variables taking values in $[0, 1]$ and a vector $Y = (Y_{n+i}, i \in [m])$ taking values in $\mathcal{Y}$. We make use of the notation $\mathbf{p}_{-i} := (p_j^{(y)})_{j\neq i, y\in\mathcal{Y}}$ for all $i \in [m]$.

First, we introduce the following assumption on $\mathbf{p}$ and Y:

ASSUMPTION 4. *There exists a vector $W = (W_i, i \in [m])$ of multivariate random variables with*

(i) for all $i \in [m]$, the random vector $\mathbf{p}_{-i} = (p_j^{(y)})_{j \neq i, y \in \mathcal{Y}}$ can be almost surely written as $\Psi_i(p_i^{(Y_{n+i})}, W_i)$ where $u \in [0,1] \mapsto \Psi_i(u, W_i) \in \mathbb{R}^{[m-1]\times\mathcal{Y}}$ is a nondecreasing function (in a coordinate-wise sense for the image space).

(ii) for all $i \in [m]$, the following super-uniformity property holds

$$\mathbb{P}(p_i^{(Y_{n+i})} \leq t \mid W_i) \leq t, \;\; t \in [0,1]. \tag{25}$$

Second, for any selection rule $\mathcal{S} \subset [m]$, that is, any measurable function of $\mathbf{p}$ valued in the subsets of $[m]$, we introduce the quantity (similar to Benjamini and Yekutieli (2005); Bao et al. (2024))

$$s_i^{\min}(\mathbf{p}_{-i}) = \min_{y \in \mathcal{A}(\mathbf{p}_{-i})} |\mathcal{S}(z, \mathbf{p}_{-i})|, \quad i \in [m], \tag{26}$$

where $\mathcal{A}(\mathbf{p}_{-i}) = \{z \in [0,1]^{\mathcal{Y}} : i \in \mathcal{S}(z, \mathbf{p}_{-i})\}$ and by convention $s_i^{\min}(\mathbf{p}_{-i}) = 0$ if $\mathcal{A}(\mathbf{p}_{-i})$ is empty. We consider the following assumption.

ASSUMPTION 5. *For $i \in [m]$, $\mathcal{A}(\mathbf{p}_{-i})$ is almost surely not empty and $s_i^{\min}(\mathbf{p}_{-i}) \geq 1$ is a coordinate-wise nonincreasing function of $\mathbf{p}_{-i}$.*

While Assumption 5 is not satisfied for a selection rule that selects the k-smallest p-values, for some $k \geq 1$ (because $\mathcal{A}(\mathbf{p}_{-i})$ can be empty), it is satisfied for p-value thresholding rules which are of interest in our setting (see Section C.4).

THEOREM C.1. *Let us consider a p-value family $\mathbf{p}$ and a label/outcome vector Y satisfying Assumption 4, for a selection rule $\mathcal{S} \subset [m]$ with $s_i^{\min} = s_i^{\min}(\mathbf{p}_{-i})$ (26) satisfying Assumption 5. Then, the procedure $\mathcal{R}_\alpha = \left(\mathcal{C}_{n+i}^{\alpha s_i^{\min}/m}(\mathbf{p})\right)_{i \in \mathcal{S}}$, for which the prediction set is defined as in (4) with level $\alpha s_i^{\min}/m$, satisfies $\mathbb{E}(\mathrm{FCP}(\mathcal{R}_\alpha, Y)) \leq \alpha$, for which the expectation $\mathbb{E}$ is taken w.r.t. the same probability than the one of (25) in Assumption 4 (ii).*

The inequality $\mathbb{E}(\mathrm{FCP}(\mathcal{R}_\alpha, Y)) \leq \alpha$ in Theorem C.1 will be used both in the cases where Y_{n+i} is fixed (conditional model) or not (iid model). The proof is provided in § C.2. Note that the considered assumptions make the proof particularly simple. Assumptions 4 and 5 are studied in § C.3 and § C.4, respectively.

C.2. Proof of Theorem C.1

By definition (2), we have

$$\begin{aligned}
\mathrm{FCP}(\mathcal{R}_\alpha, Y) &= \frac{\sum_{i \in \mathcal{S}} \mathbf{1}\{Y_{n+i} \notin \mathcal{C}_{n+i}^{\alpha s_i^{\min}/m}(\mathbf{p})\}}{1 \vee |\mathcal{S}|} = \sum_{i \in [m]} \frac{\mathbf{1}\{i \in \mathcal{S}, p_i^{(Y_{n+i})} \leq \alpha s_i^{\min}/m\}}{1 \vee |\mathcal{S}|} \\
&\leq \sum_{i \in [m]} \frac{\mathbf{1}\{i \in \mathcal{S}, p_i^{(Y_{n+i})} \leq \alpha s_i^{\min}/m\}}{s_i^{\min}} \leq \sum_{i \in [m]} \frac{\mathbf{1}\{p_i^{(Y_{n+i})} \leq \alpha s_i^{\min}/m\}}{s_i^{\min}},
\end{aligned}$$

where the first inequality follows from definition (26) and the second inequality follows from ignoring the fact that $i \in \mathcal{S}$. This entails

$$\begin{aligned}\mathbb{E}\left(\mathrm{FCP}(\mathcal{R}_\alpha, Y)\right) &\leq \sum_{i\in[m]} \mathbb{E}\left(\frac{\mathbf{1}\{p_i^{(Y_{n+i})} \leq \alpha s_i^{\min}/m\}}{s_i^{\min}}\right) \\ &= \sum_{i\in[m]} \mathbb{E}\left(\mathbb{E}\left[\frac{\mathbf{1}\{p_i^{(Y_{n+i})} \leq \alpha s_i^{\min}(\mathbf{p}_{-i})/m\}}{s_i^{\min}(\mathbf{p}_{-i})} \,\middle|\, W_i\right]\right),\end{aligned}$$

by using the random vector $W = (W_i, i \in [m])$ defined in Assumption 4. Now, combining Assumption 4 (i) with Assumption 5, we have that

$$s_i^{\min}(\mathbf{p}_{-i}) = s_i^{\min}(\Psi_i(p_i^{(Y_{n+i})}, W_i))$$

is a nonincreasing function of $p_i^{(Y_{n+i})}$. By Assumption 4 (ii) and applying Lemma F.1 (conditionally on W_i), we obtain

$$\mathbb{E}\left[\frac{\mathbf{1}\{p_i^{(Y_{n+i})} \leq \alpha s_i^{\min}(\mathbf{p}_{-i})/m\}}{s_i^{\min}(\mathbf{p}_{-i})} \,\middle|\, W_i\right] \leq \frac{\alpha}{m},$$

Putting this back into the FCR bound implies the result.

C.3. Examining Assumption 4

We show here that the full-calibrated and class-calibrated p-value families satisfy Assumption 4 in the iid and conditional models, respectively. For this, it is interesting to relax Assumption 1 by assuming that the score functions can use the covariates of the calibration+test samples in an exchangeable way, as suggested in Marandon et al. (2024); Gazin et al. (2023). This is useful for instance when the learning sample and the calibration+test samples are not based on the same distribution, so that the scores may be improved by using transfer learning; we refer to Gazin et al. (2023) for more details on this.

Assumption 6. *For any $y \in \mathcal{Y}$; $S_y(\cdot)$ is of the form $S_y\big(\cdot\,;\, \mathcal{D}_{train}, (X_i)_{i\in[n+m]}\big)$ for an independent training data sample $\mathcal{D}_{train}$ and is invariant by permutation of the elements of $(X_i)_{i\in[n+m]}$. In addition, the scores have no ties and the score function is regular in the sense of Assumption 1.*

p-value family $\overline{\mathbf{p}}$ The p-value family $\overline{\mathbf{p}}$ given by (7) satisfies Assumption 4 in the iid model.

Proposition C.2. *Let us consider a model where the variables (X_i, Y_i), $i \in [n+m]$, are exchangeable conditionally on $\mathcal{D}_{train}$ and score functions satisfying Assumptions 6. Then the p-value family $\overline{\mathbf{p}}$ given by (7) satisfies Assumption 4.*

PROOF. Let us first establish Assumption 4 (i) by following an argument similar to the one of Bates et al. (2023); Marandon et al. (2024). By Assumption 6, we can work on an event where the elements of the sets $A_i = \{S_{Y_k}(X_k), k \in [n]\} \cup \{S_{Y_{n+i}}(X_{n+i})\}$ are all distinct. For any $i \in [m]$, we have by (7) that for all $j \in [m]\backslash\{i\}$ and $y \in \mathcal{Y}$,

$$\begin{aligned}\overline{p}_j^{(y)} &= \frac{1}{n+1}\Big(1 + \sum_{k=1}^{n} \mathbf{1}\{S_{Y_k}(X_k) \geq S_y(X_{n+j})\}\Big) \\ &= \frac{1}{n+1}\Big(\mathbf{1}\{S_{Y_{n+i}}(X_{n+i}) < S_y(X_{n+j})\} + \sum_{s \in A_i} \mathbf{1}\{s \geq S_y(X_{n+j})\}\Big).\end{aligned}$$

Denoting $A_i = \{a_{i,(1)}, \ldots, a_{i,(n+1)}\}$ with $a_{i,(1)} > \cdots > a_{i,(n+1)}$, and noting that $S_{Y_{n+i}}(X_{n+i}) = a_{i,(\ell)}$ with $\overline{p}_i^{(y)} = \ell/(n+1)$, we may write

$$\overline{\mathbf{p}}_{-i} := (\overline{p}_j^{(y)})_{j \in [m]\backslash\{i\}, y \in \mathcal{Y}} = \Psi(\overline{p}_i^{(Y_{n+i})}, \overline{W}_i),$$

by letting for $u \in (0, 1]$,

$$\begin{aligned}\overline{W}_i &:= (A_i, (S_y(X_{n+j}))_{j \in [m]\backslash\{i\}, y \in \mathcal{Y}}) \\ \Psi(u, \overline{W}_i) &:= \left(\frac{1}{n+1}\Big(\mathbf{1}\{a_{i,(\lceil u(n+1)\rceil)} < S_y(X_{n+j})\} + \sum_{s \in A_i} \mathbf{1}\{s \geq S_y(X_{n+j})\}\Big)\right)_{j \in [m]\backslash\{i\}, y \in \mathcal{Y}}.\end{aligned}$$

Clearly, each of the elements inside $\Psi(u, \overline{W}_i)$ is nondecreasing in u, which gives Assumption 4 (i) (note that Ψ does not depend on i in this context).

Next, we establish Assumption 4 (ii). Since the variables (X_i, Y_i), $i \in [n+m]$, are exchangeable and since the scores functions, the set A_i, and $(S_y(X_{n+j}))_{j \in [m]\backslash\{i\}, y \in \mathcal{Y}}$ are invariant by permutations of $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+i}, Y_{n+i})$ (by using Assumptions 6), we have that the random vector $(S_{Y_1}(X_1), \ldots, S_{Y_n}(X_n), S_{Y_{n+i}}(X_{n+i}))$ is exchangeable conditionally on $\overline{W}_i$. Since there are no ties in the vector, it follows that $(n+1)\overline{p}_i^{(Y_{n+i})}$ (i.e., the rank of $S_{Y_{n+i}}(X_{n+i})$ in A_i) is uniformly distributed in $[n+1]$ conditionally on $\overline{W}_i$. Thus Assumption 4 (ii) is satisfied (note that the conditional probability is well defined thanks to the regularity condition in Assumption 6).

p-value family $\widetilde{\mathbf{p}}$ The p-value family $\widetilde{\mathbf{p}}$ given by (8) satisfies Assumption 4 in the conditional model.

PROPOSITION C.3. *In the case $\mathcal{Y} = [K]$, let us consider score functions satisfying Assumptions 6 and a model for the variables (X_i, Y_i), $i \in [n+m]$, for which $(Y_i, i \in [n+m])$ is a deterministic vector and, for each $y \in \mathcal{Y}$, the variables $(X_i)_{i \in [n+m]: Y_i = y}$, are exchangeable. Then the p-value family $\widetilde{\mathbf{p}}$ given by* (8) *satisfies Assumption 4.*

The proof is similar to the proof of Proposition C.2. We provide it below for completeness.

PROOF. For any $i \in [m]$, we have by (8) that for all $j \in [m]\backslash\{i\}$ and $y \in \mathcal{Y}$,

$$\begin{aligned}\tilde{p}_j^{(y)} &= \frac{1}{|\mathcal{D}_{\text{cal}}^{(y)}|+1}\Big(1+\sum_{k\in\mathcal{D}_{\text{cal}}^{(y)}} \mathbf{1}\{S_y(X_k) \geq S_y(X_{n+j})\}\Big)\\ &= \frac{1}{n_i^{(y)}}\Big(\mathbf{1}\{S_y(X_{n+i}) < S_y(X_{n+j})\} + \sum_{s\in A_i^{(y)}} \mathbf{1}\{s \geq S_y(X_{n+j})\}\Big),\end{aligned}$$

by letting $n_i^{(y)} = |A_i^{(y)}|$ and $A_i^{(y)} = \{S_y(X_k), k \in \mathcal{D}_{\text{cal}}^{(y)}\} \cup \{S_y(X_{n+i})\}$ whose elements can be assumed to be all distinct by Assumptions 6 (strictly, this is only true for labels $y \in \mathcal{Y}$ that appears at least once in the fixed sample $(Y_i, i \in [n+m])$, but the labels y not appearing in $(Y_i, i \in [n+m])$ can be trivially handled because they correspond to p-values all equal to 1). Denoting $A_i^{(y)} = \{a_{i,(1)}^{(y)}, \ldots, a_{i,(n_i^{(y)})}^{(y)}\}$ with $a_{i,(1)}^{(y)} > \cdots > a_{i,(n_i^{(y)})}^{(y)}$, and noting that $S_y(X_{n+i}) = a_{i,(n_i^{(y)}\tilde{p}_i^{(y)})}^{(y)}$, we may write

$$\widetilde{\mathbf{p}}_{-i} := (\tilde{p}_j^{(y)})_{j\in[m]\backslash\{i\}, y\in\mathcal{Y}} = \Psi_i(\tilde{p}_i^{(Y_{n+i})}, \widetilde{W}_i),$$

by letting

$$\widetilde{W}_i := \big((A_i^{(Y_{n+i})}, (S_{Y_{n+i}}(X_{n+j}))_{j\in[m]\backslash\{i\}}); (X_j, Y_j)_{j\in[n+m]:Y_j\neq Y_{n+i}}\big) \tag{27}$$

and for $u \in (0,1]$, $\Psi_i(u, \widetilde{W}_i) := (\Psi_i^{(y)}(u, \widetilde{W}_i))_{y\in\mathcal{Y}}$ where

$$\Psi_i^{(y)}(u, \widetilde{W}_i) := \begin{cases} \Big(\frac{1}{n_i^{(y)}}\Big(\mathbf{1}\{a_{i,(\lceil un_i^{(y)}\rceil)} < S_y(X_{n+j})\} + \sum_{s\in A_i^{(y)}} \mathbf{1}\{s \geq S_y(X_{n+j})\}\Big)\Big)_{j\in[m]\backslash\{i\}} & \text{if } y = Y_{n+i};\\ (\tilde{p}_j^{(y)})_{j\in[m]\backslash\{i\}} & \text{if } y \neq Y_{n+i}.\end{cases}$$

Clearly, the elements inside $\Psi_i(u, \widetilde{W}_i)$ are nondecreasing in u which gives Assumption 4 (i).

Next, to establish Assumption 4 (ii), we use that by assumption the vector

$$(S_{Y_{n+i}}(X_{n+i}), S_{Y_{n+i}}(X_k), k \in \mathcal{D}_{\text{cal}}^{(Y_{n+i})})$$

is exchangeable conditionally on $\widetilde{W}_i$ (by permutation invariance of $\widetilde{W}_i$, which also comes from Assumptions 6). Since there are no ties in the vector, it follows that $n_i^{(Y_{n+i})}\tilde{p}_i^{(Y_{n+i})}$ (i.e., the rank of $S_{Y_{n+i}}(X_{n+i})$ in $A_i^{(Y_{n+i})}$) is uniformly distributed in $[n_i^{(Y_{n+i})}]$ conditionally on $\widetilde{W}_i$. Thus Assumption 4 (ii) is satisfied.

C.4. Examining Assumption 5

PROPOSITION C.4. *Assumption 5 holds for the informative selection rule* $\mathcal{S}(\mathbf{p}) = \mathrm{BH}(\mathbf{q})$ *with* $s_i^{\min}(\mathbf{p}_{-i}) = |\mathcal{S}(\mathbf{p})|$ *whenever* $i \in \mathcal{S}(\mathbf{p})$.

Note that the above result is also true for selection rule of the type $\mathcal{S}(\mathbf{p}) = \{i \in [m] : q_i \leq \tau\}$ for some fixed threshold τ.

PROOF. First, a classical property of BH procedure is the following leave-one-out property: for all $i \in [m]$, $i \in \mathrm{BH}(\mathbf{q})$ if and only if $\mathrm{BH}(\mathbf{q}) = \mathrm{BH}(\mathbf{q}^{0,i})$ where $\mathbf{q}^{0,i}$ is the vector $\mathbf{q}$ where the i-th coordinate has been replaced by 0, see for instance Ferreira and Zwinderman (2006); Sarkar (2008); Roquain and Villers (2011); Ramdas et al. (2019). This implies

$$s_i^{\min}(\mathbf{p}_{-i}) = \min_{z \in [0,1]^{\mathcal{Y}} : i \in \mathcal{S}(z, \mathbf{p}_{-i})} |\mathcal{S}(z, \mathbf{p}_{-i})| = |\mathrm{BH}(\mathbf{q}^{0,i})|.$$

Hence, the result is proved as soon as $\mathbf{q}$ is proved to be coordinate-wise nondecreasing in each p-value trajectory. This holds by Lemma C.5.

LEMMA C.5. *Suppose Assumption 2, then, almost surely,* $\mathbf{q}$ *defined by* (11) *is a non-decreasing function of each p-value* $p_i^{(y)}$, $i \in [m]$ *and* $y \in \mathcal{Y}$.

PROOF. Let $\mathbf{p}$ and $\mathbf{p}'$ be two p-value collections with $p_i^{(y)} \leq p_i'^{(y)}$, for all $i \in [m]$ and $y \in \mathcal{Y}$, with corresponding values $\mathbf{q}$ and $\mathbf{q}'$. Let $i \in [m]$. By definition, $\mathcal{C}_{n+i}^{q_i'}(\mathbf{p}') \in \mathcal{I}$ and $\mathcal{C}_{n+i}^{q_i'}(\mathbf{p}) \subset \mathcal{C}_{n+i}^{q_i'}(\mathbf{p}')$. By Assumption 2 (i) (iii), we have $\mathcal{C}_{n+i}^{q_i'}(\mathbf{p}) \in \mathcal{I}$, which in turn implies $q_i \leq q_i'$ by definition of q_i.

D. Procedures using weighted class-calibrated p-values

Procedure `InfoSP` does not take into account the proportion of labels in each class in the test sample. However, in the classification case, these proportions are estimable from the data, and the estimates can aid inference.

Let $\pi_k = \sum_{i=1}^m \mathbf{1}\{Y_{n+i} = k\}/m$ the true proportion of examples with label k in the test sample, $k \in [K]$, and consider the following possible estimates:

- Calibration-based estimator: $\hat{\pi}_k^{\mathrm{cal}} = (|\mathcal{D}_{\mathrm{cal}}^{(k)}|+1)/(n+1) = (1+\sum_{j\in[n]} \mathbf{1}\{Y_j = k\})/(n+1)$, $k \in [K]$;
- Storey-λ estimator: $\hat{\pi}_k^{\mathrm{Storey}} = \left(1 + \sum_{i=1}^m \mathbf{1}\{p_i^{(k)} > \lambda\}\right)/(m(1-\lambda))$, $k \in [K]$. It is similar to the classical estimator of true null hypotheses proportion in multiple testing (Storey, 2002).

Given the class-calibrated p-value family $\widetilde{\mathbf{p}}$ and one of the estimators $\hat{\pi}_k \in \{\hat{\pi}_k^{\mathrm{cal}}, \hat{\pi}_k^{\mathrm{Storey}}\}$, we define the corresponding adaptive (weighted) p-value collection $\widetilde{\mathbf{p}}_{\mathrm{adapt}} = (\tilde{p}_{i,\mathrm{adapt}}^{(k)}, k \in [K], i \in [m]\}$ by

$$\tilde{p}_{i,\mathrm{adapt}}^{(k)} = \frac{\hat{\pi}_k}{w_k} \tilde{p}_i^{(k)}, \;\; k \in [K], i \in [m], \tag{28}$$

where $(w_k, k \in [K])$ are deterministic nonnegative weights such that $\sum_{k\in[K]} w_k = 1$. The rationale behind (28) is that the term $\hat{\pi}_k$ balances the false coverage errors between classes by trying to decrease p-values related to labels which do not appear much in the test sample. The weights w_k are additional parameter that add flexibility, but they have to sum to one. If we use equal weights than the class-calibrated p-value is multiplied by $K \times \hat{\pi}_k$ which will be less than one only if $\hat{\pi}_k < 1/K$.

Applying `InfoSP` with these adaptive p-values gives rise to a new procedure $\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\widetilde{\mathbf{p}}_{\mathrm{adapt}})$ that we denote by `Adapt-InfoSP`.

PROPOSITION D.1. *Consider an informative subset collection $\mathcal{I}$ satisfying Assumption 2, score functions satisfying Assumption 1 and the p-value collection $\widetilde{\mathbf{p}}_{adapt\text{-}cal} = (\tilde{p}^{(k)}_{i,adapt\text{-}cal}, k \in [K], i \in [m]\}$ defined as in (28) with the calibration-based estimator $\hat{\pi}_k^{cal} = (|\mathcal{D}^{(k)}_{cal}|+1)/(n+1)$, $k \in [K]$. Then the corresponding* ***Adapt-InfoSP*** *procedure $\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\widetilde{\mathbf{p}}_{adapt\text{-}cal})$ satisfies the following:*

(i) in the class-conditional model,

$$\sup_{P_{X|Y},Y} \mathrm{FCR}(\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\widetilde{\mathbf{p}}_{adapt\text{-}cal}), P_{X|Y}, Y) \leq \alpha \sum_{y\in[K]} w_y \frac{n+1}{m} \frac{\sum_{i\in[m]} \mathbf{1}\{Y_{n+i} = y\}}{\sum_{j\in[n]} \mathbf{1}\{Y_j = y\} + 1}. \tag{29}$$

(ii) in the iid model, $\sup_{P_{X,Y}} \mathrm{FCR}(\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\widetilde{\mathbf{p}}_{adapt\text{-}cal}), P_{X,Y}) \leq \alpha$*, that is,* ***Adapt-InfoSP*** *satisfies the FCR control* (15).

Proposition D.1 is proved in § E.3. The bound (29) is only sharp when the labels are generated in the same way in the calibration and test sample (which implies the correct control in (ii)), so `Adapt-InfoSP` should not be used if the label proportions are expected to be (very) different between calibration and test samples.

We illustrate in Figure 6 the performance of the adaptive procedures for nonnull selection and non-trivial selection, respectively, in the set-up of unbalanced classes described in § 5.2. For the adaptive versions, $w_k = 1/K$ for all $k \in [K]$, and $\lambda = 1/2$. We consider two settings for the class-conditional model: the test sample has class probabilities 0.1,0.1, 0.8 (i.e., a small label shift), and class probabilities 0.4,0.4,0.2 (i.e., a large label shift).

The only procedure with a theoretical class-conditional FCR guarantee is `InfoSP` on class calibrated p-values. The adaptive procedure with $\hat{\pi}_k^{cal}, k \in [K]$ violates FCR control only when the label-shift is large for non-trivial selection. Interestingly, this procedure has excellent power when the label shift is small. The adaptive procedure with $\hat{\pi}_k^{Storey}, k \in [K]$ is a close second in this case, but when the label shift is large it is no better than `InfoSP` in the settings considered. The fact that the adaptive procedure with $\hat{\pi}_k^{cal}, k \in [K]$ tends to control the FCR (or inflate it only by a little), suggests (arguably) that for power purposes it may be reasonable to use it if the label shift is small.

In the simulations we carried out for the iid settings, `Adapt-InfoSP` on class-conditional p-values had worse power than `InfoSCOP` (omitted for brevity).

E. Proofs

E.1. Proof of Theorem 3.1

The proof is straightforward from the theory developed in § C. Namely, we apply Theorem C.1 by checking the two required assumptions: Assumption 4 holds for the two considered p-value collections (§ C.3); Assumption 5 holds for the considered BH($\mathbf{q}$) selection (§ C.4).

E.2. Proof of Theorem 3.3

The proof is a consequence of Lemma F.4 applied with the FCR criterion: condition (i) in Lemma F.4 is satisfied from Theorem 3.1 (which is true more broadly in the case of exchangeable samples, see Theorem C.1 and Proposition C.2); condition (ii) in Lemma F.4 follows from the assumed permutation preserving property of $\mathcal{S}^{(0)}$. Hence, the conclusion of Lemma F.4 applies which gives the FCR control of `InfoSCOP`.

Finally, the FDR control is a consequence of the FCR control by applying Lemma 2.1.

E.3. Proof of Proposition D.1

Let us first prove (i) by considering the class-conditional model. We follow the proof of Theorem C.1 (see § C.2) and we use that the p-value collection $\widetilde{\mathbf{p}}$ satisfies Assumption 4 (see Proposition C.3), where the probability in the super-uniform property (25) holds in the class-conditional model with $\widetilde{W}_i$, $i \in [m]$ given by (27). We also use that the informative selection rule $\mathcal{S}(\cdot)$ satisfies Assumption 5 (see Proposition C.4). Hence, following the same approach as in § C.2, and denoting $\widetilde{\mathbf{p}}_{\text{cal}} := \widetilde{\mathbf{p}}_{\text{adapt-cal}}$ and $\widetilde{\mathbf{p}}_{-i,\text{cal}} := (\tilde{p}^{(y)}_{j,\text{cal}})_{j\neq i, y\in\mathcal{Y}}$, we obtain

$$\begin{aligned}\mathrm{FCR}(\mathcal{R}^{\mathrm{InfoSP}}_{\alpha}(\widetilde{\mathbf{p}}_{\text{cal}}), P_{X|Y}, Y) &\leq \sum_{i\in[m]} \mathbb{E}_{P_{X|Y}}\left(\mathbb{E}\left[\frac{\mathbf{1}\{\tilde{p}^{(Y_{n+i})}_{i,\text{cal}} \leq \alpha s^{\min}_i(\widetilde{\mathbf{p}}_{-i,\text{cal}})/m\}}{s^{\min}_i(\widetilde{\mathbf{p}}_{-i,\text{cal}})} \,\middle|\, \widetilde{W}_i\right]\right)\\ &= \sum_{i\in[m]} \mathbb{E}_{P_{X|Y}}\left(\mathbb{E}\left[\frac{\mathbf{1}\{\tilde{p}^{(Y_{n+i})}_{i} \leq \alpha s^{\min}_i(\widetilde{\mathbf{p}}_{-i,\text{cal}}) w_{Y_{n+i}}/(m\hat{\pi}^{\text{cal}}_{Y_{n+i}})\}}{s^{\min}_i(\widetilde{\mathbf{p}}_{-i,\text{cal}})} \,\middle|\, \widetilde{W}_i\right]\right).\end{aligned}$$

Using now the super-uniform property (25), the fact that $\widetilde{\mathbf{p}}_{-i,\text{cal}} = (\frac{\hat{\pi}_k}{w_k}\tilde{p}^{(k)}_j)_{k\neq i, y\in\mathcal{Y}} = \Phi(Y, \widetilde{\mathbf{p}}_{-i})$, with $\Phi(Y,\cdot)$ coordinate-wise nondecreasing, and Assumption 4 (i) entail

$$s^{\min}_i(\widetilde{\mathbf{p}}_{-i,\text{cal}}) = s^{\min}_i(\widetilde{\mathbf{p}}_{-i}) = s^{\min}_i(\Phi(Y, \Psi_i(p^{(Y_{n+i})}_i, W_i))).$$

Hence, $s^{\min}_i(\widetilde{\mathbf{p}}_{-i,\text{cal}})$ can be written as a function $g(p^{(Y_{n+i})}_i)$, with $g : u \mapsto s^{\min}_i(\Phi(Y, \Psi_i(u, W_i)))$ nonincreasing and only depending on Y and W_i. Applying Lemma F.1 for $c = w_{Y_{n+i}}/(m\hat{\pi}^{\text{cal}}_{Y_{n+i}})$, we obtain

$$\mathbb{E}_{P_{X|Y}}\left[\frac{\mathbf{1}\{\tilde{p}^{(Y_{n+i})}_{i} \leq \alpha s^{\min}_i(\widetilde{\mathbf{p}}_{-i,\text{cal}}) w_{Y_{n+i}}/(m\hat{\pi}^{\text{cal}}_{Y_{n+i}})\}}{s^{\min}_i(\widetilde{\mathbf{p}}_{-i,\text{cal}})} \,\middle|\, \widetilde{W}_i\right] \leq \alpha \frac{w_{Y_{n+i}}}{m\hat{\pi}^{\text{cal}}_{Y_{n+i}}}.$$

As a consequence, we derive

$$\mathrm{FCR}(\mathcal{R}^{\mathrm{InfoSP}}_{\alpha}(\widetilde{\mathbf{p}}_{\text{cal}}), P_{X|Y}, Y) \leq \alpha \sum_{k\in[K]} w_k \left(\frac{\sum_{i\in[m]} \mathbf{1}\{Y_{n+i} = k\}}{m} \frac{1}{\hat{\pi}^{\text{cal}}_k}\right),$$

which proves (i). We deduce (ii) by a simple integration:

$$\begin{aligned}
\mathrm{FCR}(\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\widetilde{\mathbf{p}}_{\mathrm{cal}}), P_{X,Y}) &\leq \alpha \sum_{k\in[K]} w_k \mathbb{E}\left(\frac{\sum_{i\in[m]} \mathbf{1}\{Y_{n+i}=k\}}{1+\sum_{j\in[m]}\mathbf{1}\{Y_j=k\}}\frac{n+1}{m}\right) \\
&= \alpha \sum_{k\in[K]} w_k (n+1)\mathbb{P}(Y_1=k)\mathbb{E}\left(\frac{1}{1+\sum_{j\in[m]}\mathbf{1}\{Y_j=k\}}\right) \\
&\leq \alpha \sum_{k\in[K]} w_k (n+1)\mathbb{P}(Y_1=k)\frac{1}{(n+1)\mathbb{P}(Y_1=k)} = \alpha \sum_{k\in[K]} w_k = \alpha,
\end{aligned}$$

by using Lemma F.2 for the last inequality.

E.4. Proof of Proposition 5.1

First define R_j the rank of $S_{Y_j}(X_j)$ in $\{S_k(X_j), k\in[K]\}$ (ordered in increasing order) for $j\in[n+m]$, and consider the slightly smaller conformal p-values

$$\breve{p}_i^{(k)} = \frac{1}{n+1}\left(1+\sum_{j=1}^{n}\mathbf{1}\{R_j>1\}\mathbf{1}\{S_{Y_j}(X_j)\geq S_k(X_{n+i})\}\right) \leq \overline{p}_i^{(k)},\ \ i\in[m], k\in[K], \tag{30}$$

which means that the calibration is only made with examples having a label not minimizing the score function. The rationale behind using this p-value rather than $\overline{p}_i^{(k)}$ is that, due to the post-processing, the elements X_{n+i} of the test sample cannot produce an error provided that $R_{n+i}=1$ so that we can restrict the test sample to those with $R_{n+i}>1$ when computing the FCR.

Assume $K=2$. We first prove that $\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\overline{\mathbf{p}}) = \mathcal{R}_\alpha^{\mathrm{InfoSP}}(\breve{\mathbf{p}})$. It is enough to prove that the adjusted p-values q_i obtained from $\overline{\mathbf{p}}$ and $\breve{\mathbf{p}}$ are the same for this non-trivial selection (since for the selected i, the procedure always chooses $\arg\min_{k\in[K]} S_k(X_{n+i})$ due to the post-processing). Letting $S_{\min}(x)=\min_{k\in[K]} S_k(x)$, $S_{\max}(x)=\max_{k\in[K]} S_k(x)$, this comes from

$$\begin{aligned}
\min_{k\in[K]}\{\breve{p}_i^{(k)}\} &= \frac{1}{n+1}\left(1+\sum_{j=1}^{n}\mathbf{1}\{R_j>1\}\mathbf{1}\{S_{Y_j}(X_j)\geq S_{\max}(X_{n+i})\}\right) \\
&= \frac{1}{n+1}\left(1+\sum_{j=1}^{n}\mathbf{1}\{S_{Y_j}(X_j)\geq S_{\max}(X_{n+i})\}\right) = \min_{k\in[K]}\{\overline{p}_i^{(k)}\},
\end{aligned}$$

where the second equality holds because $S_{Y_j}(X_j)\geq S_{\max}(X_{n+i})$ is impossible for $R_j=1$, that is, when $S_{Y_j}(X_j)$ is a minimum ($S_{\min}(X_j)<1/2<S_{\max}(X_j)$ almost surely by the assumptions on the score function).

Now prove Proposition 5.1, by showing the equality for $\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\breve{\mathbf{p}})$ and by carefully modifying the proof of Theorem C.1 (§ C.2) in order to get only equalities, by using that we consider the case of the non-trivial selection, that is, $\mathcal{S}=\mathrm{BH}(\mathbf{q})$, with $q_i=$

$\min_{k\in[K]} \breve{p}_i^{(k)}$. By definition of the FCP (2), we have (remember also that the prediction set always includes $\arg\min_{k\in[K]} S_k(X_{n+i})$ due to the post-processing)

$$\begin{aligned}\mathrm{FCP}(\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\breve{\mathbf{p}}), Y) &= \frac{\sum_{i\in\mathcal{S}} \mathbf{1}\{Y_{n+i} \notin \mathcal{C}_{n+i}^{\alpha|\mathcal{S}|/m}(\breve{\mathbf{p}})\}}{1 \vee |\mathcal{S}|} \\ &= \sum_{i\in[m]} \mathbf{1}\{R_{n+i} > 1\} \frac{\mathbf{1}\{i \in \mathcal{S}, \breve{p}_i^{(Y_{n+i})} \le \alpha|\mathcal{S}|/m\}}{1 \vee |\mathcal{S}|},\end{aligned}$$

because no error can occur when $R_{n+i} = 1$. Now, since $\mathcal{S} = \mathrm{BH}(\mathbf{q})$ and $q_i \le \breve{p}_i^{(Y_{n+i})}$, $\mathbf{1}\{i \in \mathcal{S}, \breve{p}_i^{(Y_{n+i})} \le \alpha|\mathcal{S}|/m\} = \mathbf{1}\{\breve{p}_i^{(Y_{n+i})} \le \alpha|\mathcal{S}|/m\}$. This entails

$$\mathrm{FCP}(\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\breve{\mathbf{p}}), Y) = \sum_{i\in[m]} \mathbf{1}\{R_{n+i} > 1\} \frac{\mathbf{1}\{\breve{p}_i^{(Y_{n+i})} \le \alpha|\mathcal{S}|/m\}}{1 \vee |\mathcal{S}|}.$$

Let $\xi = (\mathbf{1}\{R_j > 1\})_{j\in[n+m]}$ and now prove

$$\mathbb{E}[\mathrm{FCP}(\mathcal{R}_\alpha^{\mathrm{InfoSP}}(\breve{\mathbf{p}}), Y) \mid \xi] = \sum_{i\in[m]} \mathbf{1}\{R_{n+i} > 1\} \mathbb{E}\Big[\frac{\lfloor (n+1)\alpha K_i/m \rfloor/(n'+1)}{K_i} \,\Big|\, \xi\Big], \quad (31)$$

with $n' = \sum_{j\in[n]} \mathbf{1}\{R_j > 1\}$ for some random variables $K_i \ge 1$, $i \in [m]$. This implies the result because the last display is at most

$$\alpha \frac{(n+1)\sum_{i\in[m]} \mathbf{1}\{R_{n+i} > 1\}}{m(\sum_{j\in[n]} \mathbf{1}\{R_j > 1\} + 1)}$$

(with equality if $(n+1)\alpha/m$ is an integer). By Lemma F.2, the expectation of the latter is equal to $\alpha(1 - (1 - \mathbb{P}(R_1 > 1))^{n+1})$.

Let us now prove (31). For this, fix $i \in [m]$ with $R_{n+i} > 1$ and note that for all $j \in [m]$, $j \ne i$,

$$\begin{aligned} q_j &= \frac{1}{n+1}\left(1 + \sum_{k=1}^{n} \mathbf{1}\{R_k > 1\}\mathbf{1}\{S_{Y_k}(X_k) \ge S_{\max}(X_{n+j})\}\right) \\ &= \frac{1}{n+1}\Big(\mathbf{1}\{S_{Y_{n+i}}(X_{n+i}) < S_{\max}(X_{n+j})\} + \sum_{s\in A_i} \mathbf{1}\{s \ge S_{\max}(X_{n+j})\}\Big) \\ &= \frac{1}{n+1}\Big(\mathbf{1}\{S_{\max}(X_{n+i}) < S_{\max}(X_{n+j})\} + \sum_{s\in A_i} \mathbf{1}\{s \ge S_{\max}(X_{n+j})\}\Big) \\ &\ge \frac{1}{n+1} \sum_{s\in A_i} \mathbf{1}\{s \ge S_{\max}(X_{n+j})\} =: q'_j, \end{aligned}$$

by letting $A_i = \{S_{Y_k}(X_j), j \in [n] : R_k > 1\} \cup \{S_{Y_{n+i}}(X_{n+i})\}$ (all distinct almost surely). The third equality above is true because $K = 2$ and $R_{n+i} > 1$ and thus $S_{Y_{n+i}}(X_{n+i}) = S_{\max}(X_{n+i})$. We apply now Lemma F.3 because $\mathbf{q} = (q_j, 1 \le j \le m)$ and $\mathbf{q}' = (q'_j, 1 \le$

$j \leq m$) (defined as above and with $q'_i = 1/(n+1)$) satisfy (32). Indeed, if $q_j > q_i$ for $j \neq i$, then $S_{\max}(X_{n+j}) \leq S_{\max}(X_{n+i})$ and $q_j = q'_j$. Hence, we obtain, by letting $K_i = |\mathrm{BH}(\mathbf{q}')|$ (note that $\mathbf{q}'$ depends on i)

$$\{q_i \leq \alpha |\mathrm{BH}(\mathbf{q})|/m\} = \{q_i \leq \alpha K_i/m\} \subset \{|\mathrm{BH}(\mathbf{q})| = K_i\}.$$

In addition, $\mathbf{q}' = (q'_j, 1 \leq j \leq m)$ is a vector measurable w.r.t. the variable $W_i = (A_i, (S_{\max}(X_{n+j}))_{j \in [m] \setminus \{i\}})$. Now, we use that by exchangeability of the elements of A_i conditionally on ξ, W_i, and $R_{n+i} > 1$) (with no ties),

$$\mathbb{P}(\breve{p}_i^{(Y_{n+i})} \leq t \mid W_i, \xi, R_{n+i} > 1) = \frac{\lfloor (n+1)t \rfloor}{n'+1}.$$

Applying this with $t = \alpha K_i/m$ entails (31).

F. Auxiliary results

Lemma F.1 (Lemma 3.2 in Blanchard and Roquain (2008)). *Let $g : [0,1] \to (0,\infty)$ be a nonincreasing function and U be a random variable which is super-uniform, that is, $\forall u \in [0,1]$, $\mathbb{P}(U \leq u) \leq u$. Then, for any $c > 0$, we have*

$$\mathbb{E}\left[\frac{\mathbf{1}\{U \leq cg(U)\}}{g(U)}\right] \leq c.$$

Lemma F.2 (Lemma 1 of Benjamini et al. (2006)). *If T is a Binomial variable with parameter $N-1 \geq 0$ and $t \in (0,1]$, we have*

$$\mathbb{E}[1/(T+1)] = (1-(1-t)^N)/(Nt) \leq 1/(Nt).$$

Lemma F.3 (Lemma D.6 of Marandon et al. (2024)). *Write $\widehat{\ell} = \widehat{\ell}(\mathbf{q}) = |\mathrm{BH}(\mathbf{q})|$ for the number of rejections of $\mathrm{BH}(\mathbf{q})$ (21). Fix any $i \in \{1,\dots,m\}$ and consider two collections $\mathbf{q} = (q_j, 1 \leq j \leq m)$ and $\mathbf{q}' = (q'_j, 1 \leq j \leq m)$ which satisfy almost surely that*

$$\forall j \in \{1,\dots,m\}, \begin{cases} q'_j \leq q_j & \text{if } q_j \leq q_i; \\ q'_j = q_j & \text{if } q_j > q_i. \end{cases} \tag{32}$$

Then $\{q_i \leq \alpha \widehat{\ell}(\mathbf{q})/m\} = \{q_i \leq \alpha \widehat{\ell}(\mathbf{q}')/m\} \subset \{\widehat{\ell}(\mathbf{q}) = \widehat{\ell}(\mathbf{q}')\}$.

The calibration splitting trick can be seen as a way to obtain statistical guarantees in conformal inference when making a data-driven choice regarding the inference . In Marandon et al. (2024); Gazin et al. (2023), the choices are about adaptive score functions. The next lemma presents this trick in the case that the choice is about which examples to select for (potentially more efficient and powerful) further inference ‖.

Lemma F.4 (Calibration splitting trick for selective conformal prediction sets). *Assume that for any training sample $\mathcal{D}_{train}$, calibration sample $\mathcal{D}_{cal} = ((X_j, Y_j))_{j \in [n_{cal}]}$ and test sample $\mathcal{D}_{test} = (X_j, Y_j)_{j \in [n_{cal}+1, n_{cal}+n_{test}]}$, such that $(\mathcal{D}_{train}, \mathcal{D}_{cal}, \mathcal{D}_{test})$ has some distribution Q, the procedure (that is, a prediction set collection on a selection) and initial selection rule are as follows:*

‖This idea was sketched in version 3 of the arXiv version Bao et al. (2023) of the work Bao et al. (2024)

(i) the procedure $\mathcal{R} = (\mathcal{C}_{n_{cal}+i})_{i\in\mathcal{S}}$, $\mathcal{C}_{n_{cal}+i} \subset \mathcal{Y}$, $\mathcal{S} \subset [n_{test}]$, *with* $\mathcal{R} = \mathcal{R}(\mathcal{D}_{cal}, \mathcal{D}^X_{test}; \mathcal{D}_{train})$ *built upon* $\mathcal{D}_{train}, \mathcal{D}_{cal}$ *and* $\mathcal{D}^X_{test} := (X_j)_{j\in[n_{cal}+1,n_{cal}+n_{test}]}$ *such that it controls a criterion* $\mathbb{E}_Q(\mathcal{E}(\mathcal{R}, Y)) \leq \alpha$ *if the entries of* $\mathcal{D}_{cal} \cup \mathcal{D}_{test}$ *are exchangeable conditionally on* $\mathcal{D}_{train}$*;*

(ii) an initial selection rule $\mathcal{S}^{(0)} = \mathcal{S}^{(0)}(\mathcal{D}_{cal}, \mathcal{D}^X_{test}; \mathcal{D}_{train}) \subset [n_{cal} + 1, n_{cal} + n_{test}]$ *built upon* $\mathcal{D}_{train}, \mathcal{D}_{cal}$ *and* $\mathcal{D}^X_{test}$ *which is permutation preserving in the latter, that is, for any permutation* σ *of* $[n_{cal} + 1, n_{cal} + n_{test}]$*, we have* $\sigma(\mathcal{S}^{(0)}(\mathcal{D}_{cal}, \mathcal{D}^X_{test}; \mathcal{D}_{train})) = \mathcal{S}^{(0)}(\mathcal{D}_{cal}, \sigma(\mathcal{D}^X_{test}); \mathcal{D}_{train})$.

Let us consider samples $(\mathcal{D}_{train}, \mathcal{D}_{cal}, \mathcal{D}_{test})$ *as above such that the entries of* $\mathcal{D}_{cal} \cup \mathcal{D}_{test}$ *are exchangeable conditionally on* $\mathcal{D}_{train}$ *and split the calibration set* $\mathcal{D}_{cal}$ *into two samples* $\mathcal{D}^{(1)}_{cal}$ *and* $\mathcal{D}^{(2)}_{cal}$ *of respective sizes* $n^{(1)}_{cal}$ *and* $n^{(2)}_{cal}$*. Let* $\mathcal{D}^{(2),X}_{cal} = (X_j)_{j\in[n^{(1)}_{cal}+1,n_{cal}]}$*. Then for* $\mathcal{S}^{(0)} = \mathcal{S}^{(0)}(\mathcal{D}^{(1)}_{cal}, \mathcal{D}^{(2),X}_{cal} \cup \mathcal{D}^X_{test}; \mathcal{D}_{train}) \subset [n^{(1)}_{cal} + 1, n_{cal} + n_{test}]$*, the procedure* $\mathcal{R}^{split} = \mathcal{R}((\mathcal{D}^{(2)}_{cal})_{\mathcal{S}^{(0)}}, (\mathcal{D}^X_{test})_{\mathcal{S}^{(0)}}; \mathcal{D}_{train})$ *is a procedure achieving* $\mathbb{E}_Q(\mathcal{E}(\mathcal{R}^{split}, Y) \mid \mathcal{S}^{(0)}) \leq \alpha$ *and thus also* $\mathbb{E}_Q(\mathcal{E}(\mathcal{R}^{split}, Y)) \leq \alpha$.

For instance, Lemma F.4 can be used for the criterion $\mathcal{E}(\mathcal{R}, Y) = \mathrm{FCP}(\mathcal{R}, Y)$ in which case it provides FCR control. For instance, since classical marginal prediction intervals always satisfy (i) with $S = [m]$ and the FCR criterion, Lemma F.4 shows that sample splitting can offer FCR control for any (permutation preserving) selection rule. This is an alternative method to the swapping approach of Jin and Ren (2023).

Proof. For convenience, let us write $\mathcal{D}_{\mathrm{cal}} \cup \mathcal{D}_{\mathrm{test}} = ((X_j, Y_j))_{j\in[n_{\mathrm{cal}}+n_{\mathrm{test}}]} = (Z_j)_{j\in[n_{\mathrm{cal}}+n_{\mathrm{test}}]}$. By (i), it is enough to prove that the entries of $(Z_j)_{j\in S}$ are exchangeable conditionally on $\mathcal{S}^{(0)} = S$, $\mathcal{D}^{(1)}_{\mathrm{cal}}$ and $\mathcal{D}_{\mathrm{train}}$, for any possible realisation S of $\mathcal{S}^{(0)}$. For any σ permutation of $[n^{(1)}_{\mathrm{cal}} + 1, n_{\mathrm{cal}} + n_{\mathrm{test}}]$ which only affects the indexes of S, we have

$$\begin{aligned}
&\mathcal{D}((Z_{\sigma(j)})_{j\in S, j\geq n^{(1)}_{\mathrm{cal}}+1} \mid \mathcal{S}^{(0)} = S, \mathcal{D}^{(1)}_{\mathrm{cal}}, \mathcal{D}_{\mathrm{train}}) \\
&= \mathcal{D}((Z_{\sigma(j)})_{j\in S, j\geq n^{(1)}_{\mathrm{cal}}+1} \mid \mathcal{S}(\mathcal{D}^{(1)}_{\mathrm{cal}}, \mathcal{D}^{(2),X}_{\mathrm{cal}} \cup \mathcal{D}^X_{\mathrm{test}}; \mathcal{D}_{\mathrm{train}}) = S, \mathcal{D}^{(1)}_{\mathrm{cal}}, \mathcal{D}_{\mathrm{train}}) \\
&= \mathcal{D}((Z_{\sigma(j)})_{j\in S, j\geq n^{(1)}_{\mathrm{cal}}+1} \mid \mathcal{S}(\mathcal{D}^{(1)}_{\mathrm{cal}}, (Z_{\sigma(j)})_{j\in[n^{(1)}_{\mathrm{cal}}+1, n_{\mathrm{cal}}+n_{\mathrm{test}}]}; \mathcal{D}_{\mathrm{train}}) = S, \mathcal{D}^{(1)}_{\mathrm{cal}}, \mathcal{D}_{\mathrm{train}}),
\end{aligned}$$

by using the permutation preserving property (ii) and because $\sigma(S) = S$ by definition of σ. Now using that the distribution of $(Z_{\sigma(j)})_{j\in[n^{(1)}_{\mathrm{cal}}+1,n_{\mathrm{cal}}+n_{\mathrm{test}}]}$ is the same as $(Z_j)_{j\in[n^{(1)}_{\mathrm{cal}}+1,n_{\mathrm{cal}}+n_{\mathrm{test}}]}$ conditionally on $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}^{(1)}_{\mathrm{cal}} = (Z_j)_{j\in[n^{(1)}_{\mathrm{cal}}]}$, the last display is equal to

$$\mathcal{D}((Z_j)_{j\in S, j\geq n^{(1)}_{\mathrm{cal}}+1} \mid \mathcal{S}^{(0)} = S, \mathcal{D}^{(1)}_{\mathrm{cal}}, \mathcal{D}_{\mathrm{train}}),$$

which provides the desired exchangeability for valid inference on $(\mathcal{D}^X_{\mathrm{test}})_{\mathcal{S}^{(0)}}$ using $(\mathcal{D}^{(2)}_{\mathrm{cal}})_{\mathcal{S}^{(0)}}$.

G. Additional illustrations in the regression case

G.1. Excluding a null value $y_0 \in \mathbb{R}$

We consider here $\mathcal{I} = \{I \text{ interval of } \mathbb{R} : y_0 \notin I\}$, which is useful in situation where the value y_0 corresponds to some "normal" value and the user wants to report only prediction

intervals for "abnormal" individuals, that is, when the outcome value deviates from this reference value. The score considered here is $S_y(x) = |y - \mu(x)|/\sigma(x)$, where μ, σ are predictors of the conditional mean and variance, respectively.

An illustration is provided in Figure 7 in the non-parametric regression Gaussian model for different situations: Rows 1,2 correspond to an homoscedastic model with perfect prediction of the variance and a mean predictor with various accuracy (less accurate in the middle for row 1, more accuracy in the middle for row 2). Rows 3,4 correspond to an heteroscedastic model with perfect prediction of the mean and a the variance predictor under-estimating the variance in the middle for row 3, and over-estimating the variance in the middle for row 4). The marginal prediction interval (no selection) is displayed in light blue while the prediction interval after selection is displayed in dark red. The quantities reported at the top of each panel are the FCR and adjusted power averaged over 1000 repetitions (while each panel only displays the last experiment as a typical realisation of the sample).

First, as expected, the naive classical conformal selection does not provide FCR control in all cases, while `InfoSP` and `InfoSCOP` do control the FCR in any case. Second, `InfoSCOP` always improves `InfoSP` in terms of power, and the range of improvement depends on where non-covering errors are likely to happen: when errors are more likely to arise on the selection (rows 1,3), the behavior of `InfoSP` and `InfoSCOP` are similar; when errors are less likely to arise on the selection (rows 2,4), `InfoSCOP` improves over `InfoSP` in a striking manner (and is even better than classical conformal). This is both because of the reduction of the selection effect and because the pre-processed p-values are calibrated with much lower residuals and thus are much more efficient than the original p-values.

G.2. Prediction intervals for Jin and Candes (2023) selection

We focus here on the case where the user wants to build prediction sets only for outcomes $Y_{n+i} > y_0$, which corresponds to excluding the set $\mathcal{Y}_0 = (-\infty, y_0]$ and is related to the selection proposed in Jin and Candes (2023).

We consider here the aim of finding one-sided prediction intervals on the selected. Let us assume that the score function is monotone in the following sense (Jin and Candes, 2023): for all $x \in \mathbb{R}^d$, for $y \leq y'$, $S_y(x) \geq S_{y'}(x)$. A classical example of monotonic score function is given by $S_y(x) = (\mu(x) - y)/\sigma(x)$. The following result summarizes our finding in this case.

Corollary G.1. *Consider the iid model in the regression case and assume that the score function is monotone (see above) and such that Assumption 2 (ii) (iii) and Assumption 1 hold. Then the following holds for* `InfoSP` *with informative collection* $\mathcal{I} = \{I \text{ interval of } \mathbb{R} \,:\, I \cap (-\infty, y_0] = \emptyset\}$ *and p-value collection* $\overline{\mathbf{p}}$ (7)*:*

(i) `InfoSP` *selects* $\mathcal{S} = \mathrm{BH}(\mathbf{q})$ *the rejected set of BH procedure at level* α *applied with the p-values*

$$q_i = \overline{p}_i^{(y_0)} = \frac{1}{n+1}\Big(1 + \sum_{j=1}^{n} \mathbf{1}\{S_{Y_j}(X_j) \geq S_{y_0}(X_{n+i})\}\Big),$$

which coincides with the rejection set of the procedure proposed in Jin and Candes (2023).

(ii) The selection $\mathcal{S} = \mathrm{BH}(\mathbf{q})$ *of* `InfoSP` *controls the FDR at level* α *in the following sense:*

$$\sup_{P_{XY}} \mathbb{E}_{(X,Y)\sim P_{XY}} \left[\frac{\sum_{i\in\mathcal{S}} \mathbf{1}\{Y_{n+i} \leq y_0\}}{1 \vee |\mathcal{S}|} \right] \leq \alpha.$$

(iii) The prediction intervals are of the form $\mathcal{C}_{n+i} = \{y > y_0 \ : \ S_y(X_{n+i}) \leq S_{(n_\alpha)}\}$, $i \in \mathcal{S}$, *where* $S_{(1)} \leq \cdots \leq S_{(n)}$ *are the ordered calibration scores* $S_{Y_j}(X_j)$, $1 \leq j \leq n$ *(with* $S_{(n+1)} = +\infty$*), and* $n_\alpha = \lceil (1-\alpha|\mathcal{S}(\mathbf{p})|/m)(n+1)\rceil$.

(iv) These prediction intervals control the FCR at level α *in the sense of* (15).

In other words, the above result provides complements the multiple testing procedure of Jin and Candes (2023), by providing in addition FCR controlling informative prediction sets on the selected outcomes.

Proof. The proof is direct with Example 2.1 and monotonicity (for (i)), Lemma 2.1 (for (ii)), Remark 2.1 (for (iii)), Theorem 3.1 (for (iv)).

Obviously, a similar result holds for `InfoSCOP`, for any initial selection step $\mathcal{S}^{(0)} \subset [r+1, n+m]$ that satisfies the permutation preserving Assumption 3. Corollary G.1 is illustrated on Figure 8 when $\mathcal{S}^{(0)}$ is taken here has $\mathrm{BH}(\mathbf{q})$ at level 2α with the score $S_y(x) = (\mu(x) - y)/\sigma(x)$. The comments are qualitatively similar to those of Figure 7: when covering errors are less likely on the selection, the improvement of `InfoSCOP` over `InfoSP` is substantial.

G.3. Comparison with existing selective prediction sets

Figure 9 is useful to visualize the difference between the selection proposed by Bao et al. (2024) and our informative selection. We display selective prediction intervals for two FCR controlling selections proposed in Bao et al. (2024): `SCOPa` uses a thresholding rule $\mathcal{S} = \{i \in [m] \ : \ \mu(X_{n+i}) \geq y_0\}$, while `SCOPb` selects the largest $\mu(X_{n+i})$'s. Each time, a suitable conditional conformal prediction set is built in Bao et al. (2024) and reported in Figure 9 in green and purple. As one can see, either the selection is too conservative, or the prediction interval could include the nominal b_0. This not the case of `InfoSCOP` which always exclude y_0 by essence.

H. Additional illustrations in the classification case

H.1. Three classes, each bivariate normal

For the iid model, as described in § 5.2, we demonstrate the initial selection step in `infoSCOP` for excluding a null class in one realization of the data generation in Figure 10.

For non-trivial classification, as described in § 5.3, we provide numerical results in Figure 11.

H.2. Three classes, each a mixture of bivariate normals

For the iid model, as in § 5.2, we consider the balanced and unbalanced cases of 3 classes. However, here we consider that each class comes from a mixture of two bivariate normals, where one component of the mixture is identical in all three classes, and given the class there is probability half of coming from that common mixture component. Thus the overlap between the classes is much greater than in the settings considered in § 5.2.

In Figure 12 we show that the qualitative conclusions with regard to the respective procedures are as in § 5.2. Interestingly, in the hardest setting at the bottom row we see that the power of `InfoSCOP` is even larger than classic conformal. This is because the initial selection step not only prevents paying too much for selection, but it can also improve the calibration set for computing the conformal p-values after selection. In this hard setting, there is a large improvement of the tail probability that matters for inference. We demonstrate this in one realization of the data generation in Figure 13, where we can see that the 0.95 quantile of the original calibration set is much larger than the 0.95 quantile of examples that remain in the calibration set after the initial selection step. This is because the data generation is such that examples in the central cloud tend to have larger nonconformity scores $S_{Y_i}(X_i)$ than the examples that are not in the central cloud, and most of the examples from the central cloud were eliminated from the calibration sample with the initial selection step. So the remaining nonconformity scores after the initial selection step tend to be smaller. Since the nonconformity scores of the examples from the test sample are unchanged, the improved calibration set after the initial selection step results in p-values that tend to be smaller when testing the null group, and in particular there are many more p-values that are below the 0.05 level.

H.3. Three classes of animals

For the class-conditional model with label shift described in § 6, Figure 14 provides the resulting FCR and power measures.

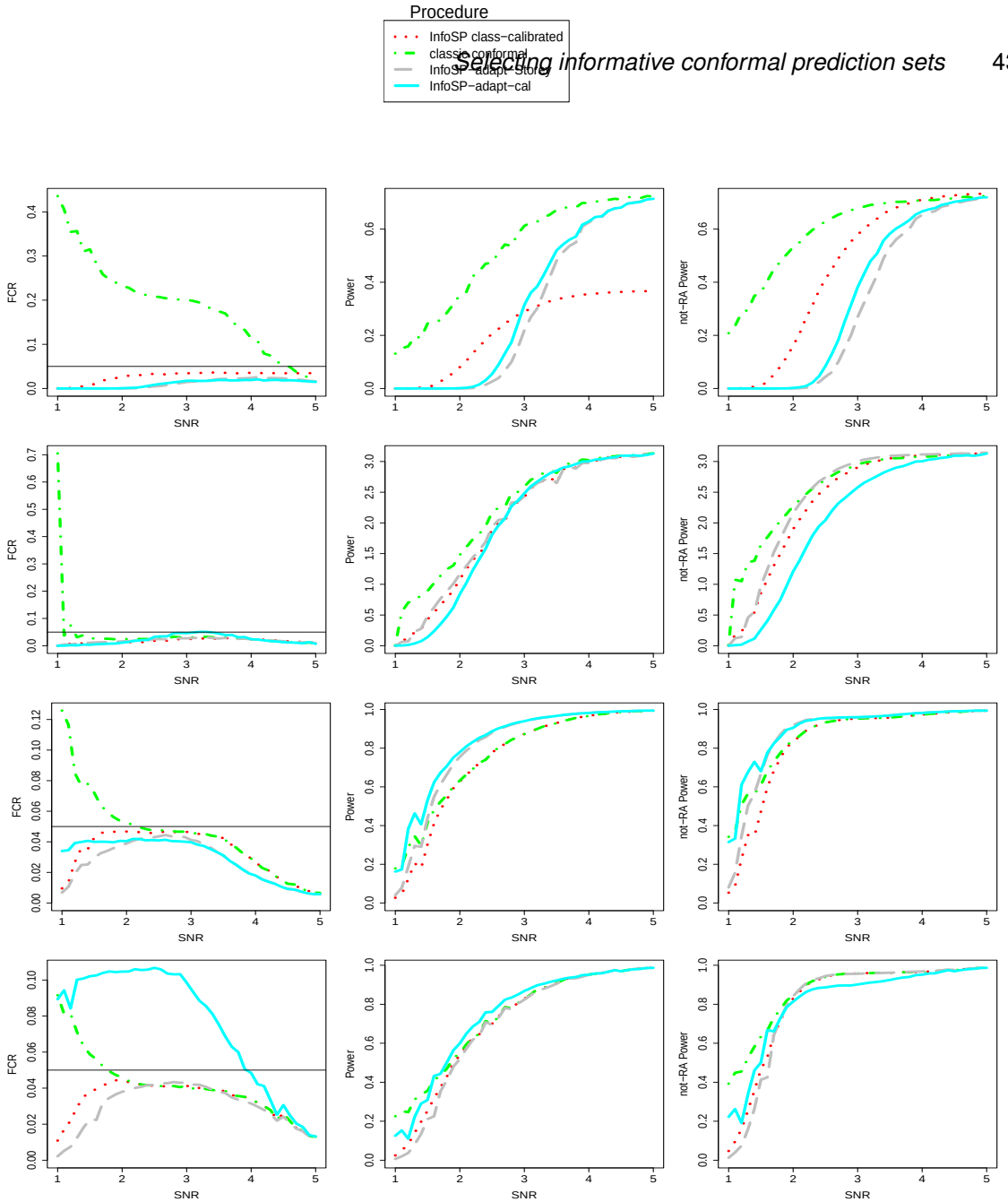


Fig. 6: Selecting informative prediction sets in the class-conditional setting. FCR (left column), resolution-adjusted power (middle column), and the expected fraction of covering prediction sets (right column) versus SNR in a classification setting where prediction sets excluding a null class are of interest (top two rows) and when prediction sets excluding a trivial class are of interest (bottom two rows). The class probabilities in the calibration sample are 0.15, 0.1, and 0.75; in the test sample, we have a small label shift (rows 1 and 3) and a large label shift (rows 2 and 4). The number of data generations was 2000, and $n = m = 500$.

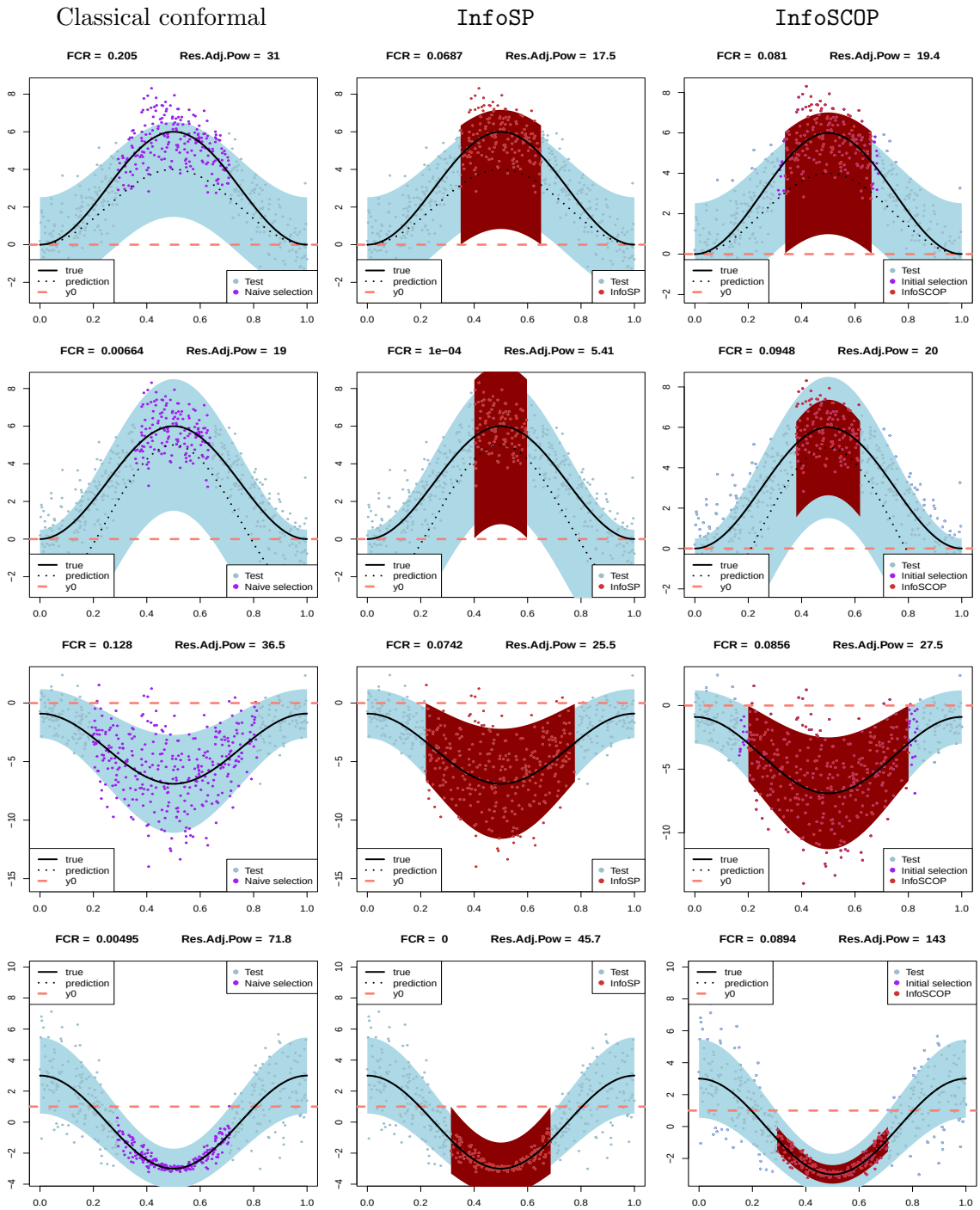


Fig. 7: Informative selection by excluding the null value y_0 in the (homoscedastic or heteroscedastic) regression case, see § G.1.

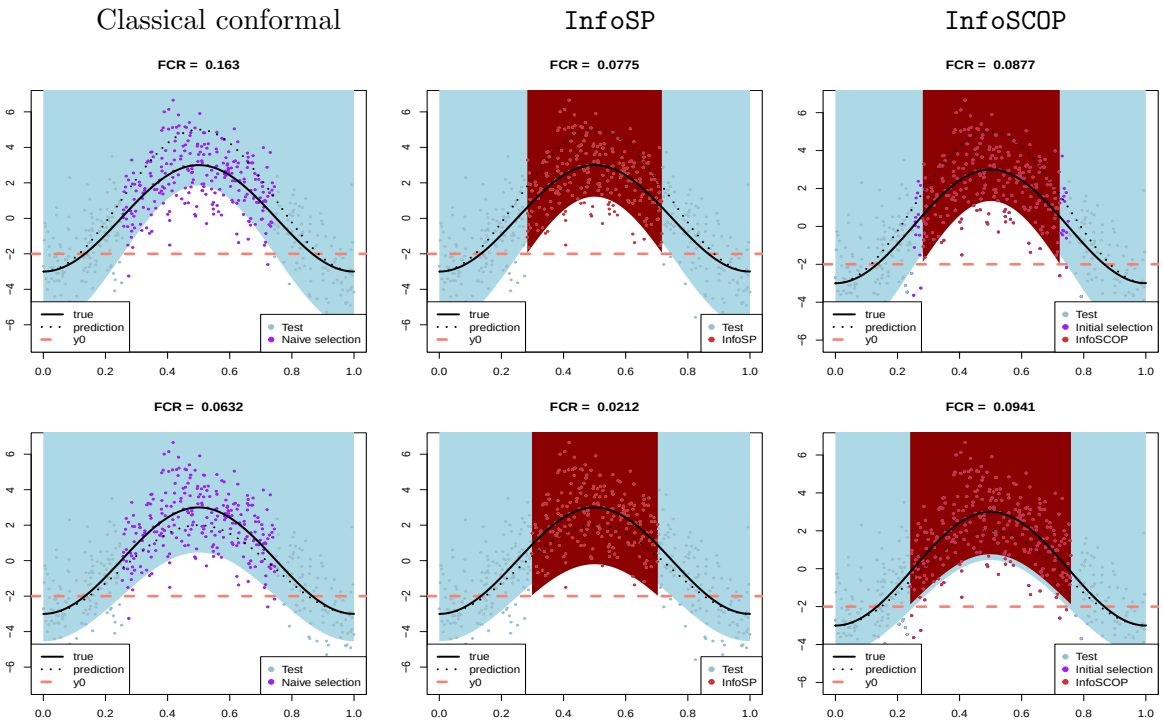


Fig. 8: Same as Figure 7 with Jin and Candes (2023) type selection and one-sided prediction intervals, see § G.2.

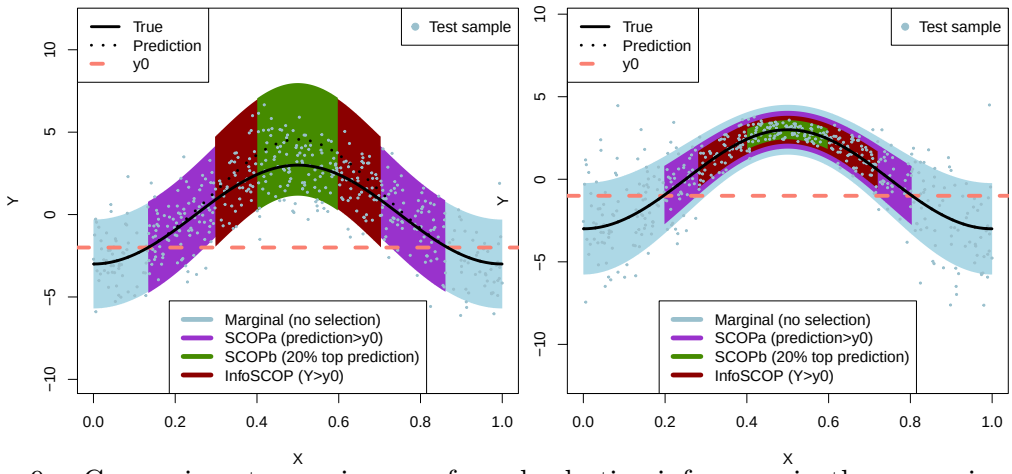


Fig. 9: Comparison to previous conformal selective inferences in the regression case: new informative prediction intervals (`InfoSCOP`) versus non-informative prediction intervals (SCOPab) (left: homoscedastic Gaussian regression with perfect variance prediction and errors in the mean prediction; right: heteroscedastic Gaussian regression with perfect mean prediction and errors in the variance prediction). Informative means here prediction intervals that does not contain y_0 (dashed line), see § G.3.

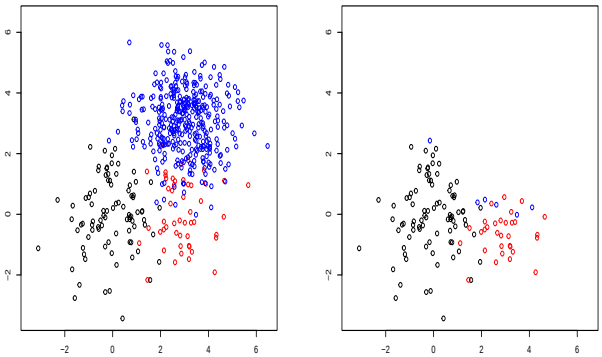


Fig. 10: The test sample in a single data generation for selecting informative prediction sets that exclude a null class. The setting is that of unbalanced classes, and the SNR is 3. The data points from the test sample of each of the three classes, where the null group is in blue: left panel for the entire test sample, right panel the remaining examples from the test set after the initial selection step.

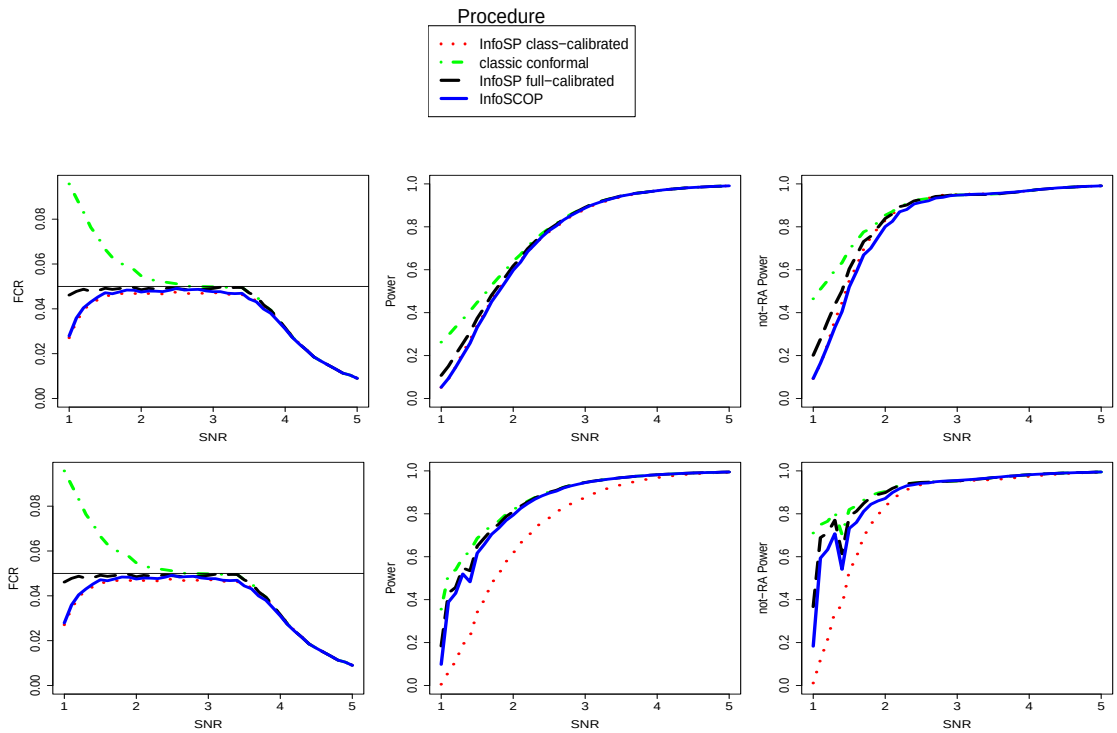


Fig. 11: Selecting non-trivial prediction sets in the classification iid model, in the case of balanced classes (top row) and unbalanced classes (bottom row), for the Gaussian mixture model with $K = 3$ classes described in § 5.2. FCR (first column), resolution-adjusted power (second column) and the expected fraction of covering prediction sets (third column) versus SNR. The number of data generations was 2000, 1000 data points were used for training, and $n = m = 500$.

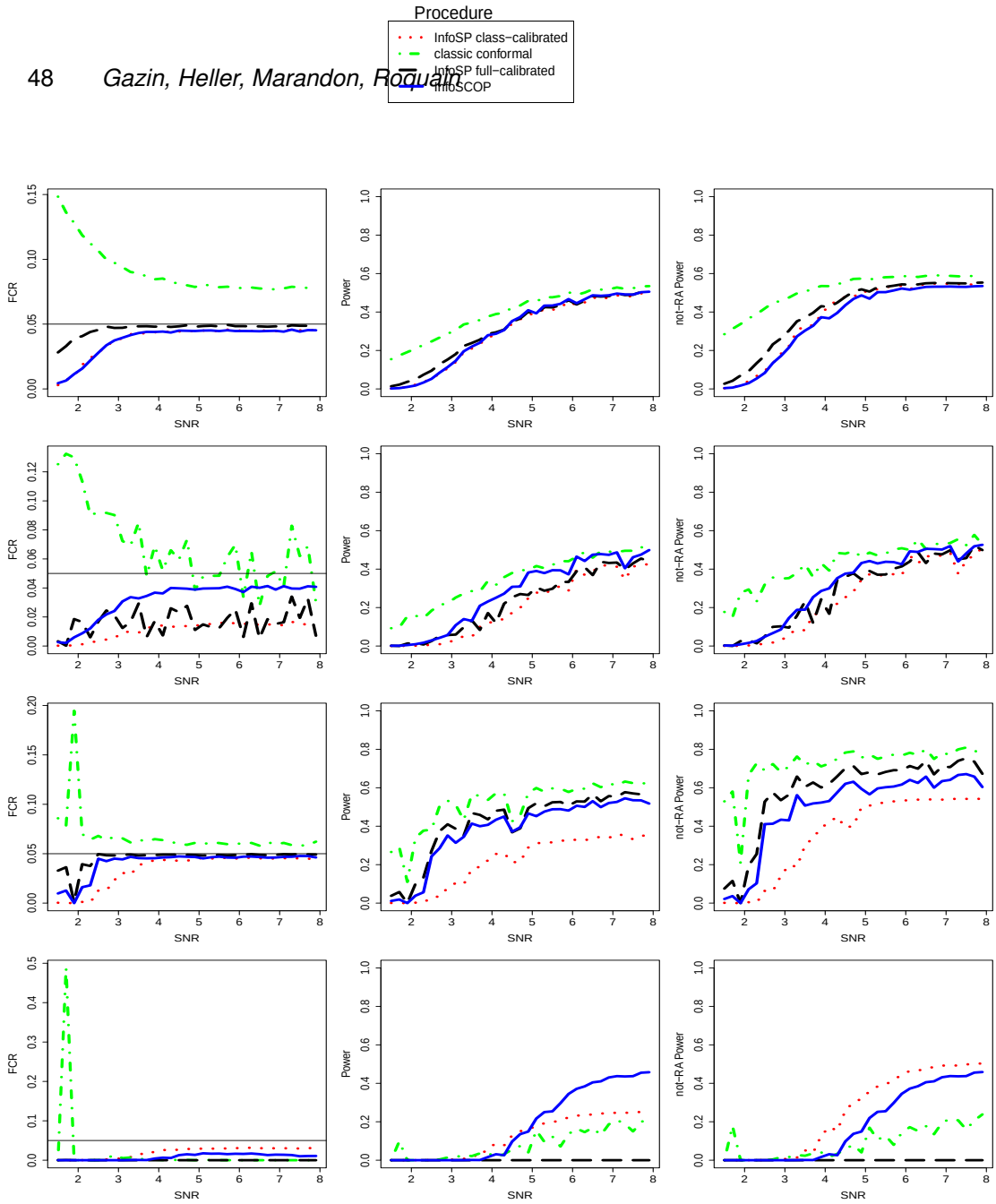


Fig. 12: Selecting informative prediction sets in the iid setting with a large overlap between three classes. FCR (left column), resolution-adjusted power (middle column) and the expected fraction of covering prediction sets (right column) versus SNR for: balanced classes for minimally informative prediction sets (first row) and for prediction sets excluding a null class (second row); unbalanced classes for minimally informative prediction sets (third row) and for prediction sets excluding the largest class (fourth row). Each class is a mixture of a common component and a unique component, see § H.2 for details. Based on 2000 data generations, 1000 data points were used for training, and $n = m = 500$.

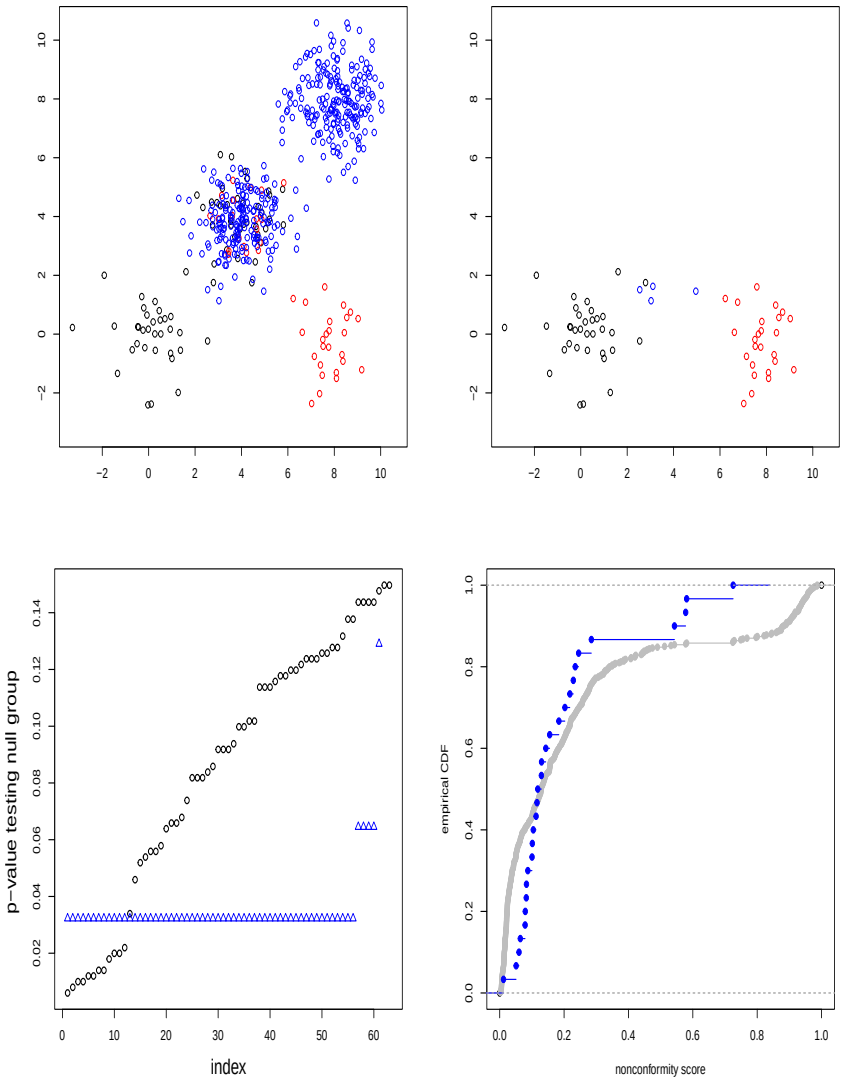
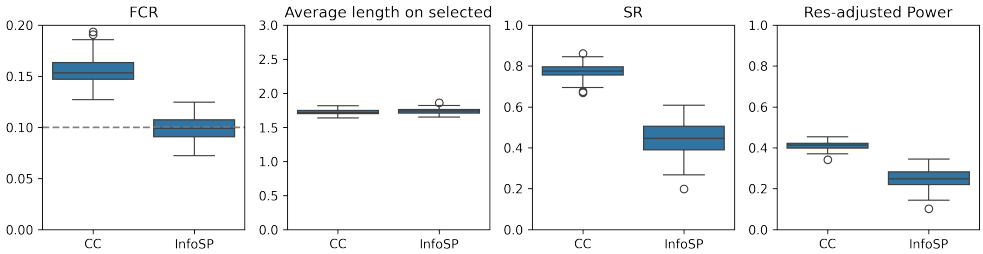


Fig. 13: The setting is that of unbalanced classes, with each class having probability half of being in the same mixture component, and the SNR is 8. In the first row, the data points from each of the three classes, where the null group is in blue: top left panel for the entire test sample, top right panel the remaining examples after the initial selection step. In the second row, left panel, their p-values using the entire calibration set (black circles) and using the examples from the calibration set remaining after the initial selection step (blue triangles). In the second row, right panel, the empirical CDF of the nonconformity scores for true classes in the entire calibration set (gray) and using the examples from the calibration set remaining after the initial selection step for the of the (blue).

c) Non-trivial classification with label shift



d) Non-null classification with label shift

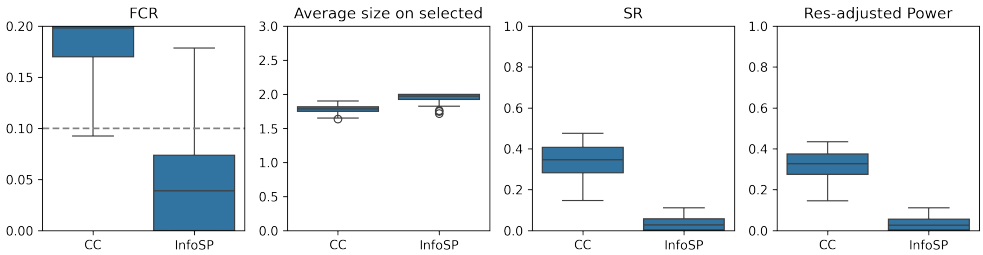


Fig. 14: For the class-conditional model, the FCR, average size of the selected, SR, and resolution-adjusted power for the methods, for $\alpha = 0.1$.