Received: 25 June 2019 | Accepted: 12 December 2020

DOI: 10.1111/rssb.12443

ORIGINAL ARTICLE

Prediction and outlier detection in classification problems

Leying Guan[1] | Robert Tibshirani[2]

[1]Yale University, New Haven, CT, USA

[2]Stanford University, Stanford, CA, USA

Correspondence
Leying Guan, Department of Biostatistics, Yale University, CT, 06510, USA
Email: leying.guan@yale.edu

Funding information
National Institutes of Health, Grant/Award Number: 5R01 EB001988-16; National Science Foundation, Grant/Award Number: 19 DMS1208164

Abstract

We consider the multi-class classification problem when the training data and the out-of-sample test data may have different distributions and propose a method called BCOPS (balanced and conformal optimized prediction sets). BCOPS constructs a prediction set $C(x)$ as a subset of class labels, possibly empty. It tries to optimize the out-of-sample performance, aiming to include the correct class and to detect outliers x as often as possible. BCOPS returns no prediction (corresponding to $C(x)$ equal to the empty set) if it infers x to be an outlier. The proposed method combines supervised learning algorithms with conformal prediction to minimize a misclassification loss averaged over the out-of-sample distribution. The constructed prediction sets have a finite sample coverage guarantee without distributional assumptions. We also propose a method to estimate the outlier detection rate of a given procedure. We prove asymptotic consistency and optimality of our proposals under suitable assumptions and illustrate our methods on real data examples.

KEYWORDS

BCOPS, conformal inference, distributional change, label shift, set-valued prediction

1 | INTRODUCTION

We consider the multi-class classification problem where the training data and the test data may be mismatched. That is, the training data and the test data may have different distributions. We assume access to the labelled training data and unlabelled test data. Let $\{(x_i, y_i), i = 1, \ldots, n\}$ be the training data set, with continuous features $x_i \in \mathbb{R}^p$ and response $y_i \in \{1, \ldots, K\}$ for K classes.

In classification problems, one usually aims to produce a good classifier using the training data that predicts the class k at each x. As a contrast, we construct a *prediction set* $C(x)$ at each x by solving an optimization problem minimizing the out-of-sample loss directly. The prediction set $C(x)$ might contain multiple labels or be empty. When $K = 2$, for example, $C(x) \in \{\{1\},\{2\},\{1,2\},\emptyset\}$. If $C(x)$ contains multiple labels, it would indicate that x could be any of the listed classes. If $C(x) = \emptyset$, it would indicate that x is likely to be far from the training data, and we could not assign it to any class and consider it an outlier.

There are many powerful supervised learning algorithms that try to estimate $P(y = k|x)$, the conditional probability of y given x, for $k = 1, \ldots, K$. When the test data and the training data have the same distribution, we can often have reasonably good performance and relatively accurate evaluation of the out-of-sample performance using sample slitting. However, the posterior probability $P(y = k|x)$ may not reveal the fact that the training and test data are mismatched. In particular, when erroneously applied to mismatched data, the standard approaches may yield predictions for x far from the training samples, where it is usually better not to make a prediction at all. Figure 1 shows a two-dimensional illustrative example. In this example, we have a training data set with two classes and train a logistic regression model. The average misclassification loss based on sample splitting of the training data is extremely low. The test data come from a very different distribution. We plot the training data in the upper left panel (black points = class 1, blue points = class 2), and plot the test data in the upper right panel using red points. The black and blue dashed curves in these two plots are the boundaries for $P(y = 1|x) = 0.05$ and $P(y = 1|x) = 0.95$ from the logistic regression model. Based on the predictions from the logistic regression, we are confident that most of the red points are from class 1. However, in this case, since the test samples are relatively far from the training data, most likely, we consider them to be outliers and do not want to make predictions.

As an alternative, the density-level set (Cadre, 2006; Hartigan, 1975; Lei et al., 2013; Rigollet & Vert, 2009; Sadinle et al., 2019) considers $f_y(x)$, the density of x given y. For each x, it constructs a prediction set $C(x) = \{k: x \in A_k\}$where $A_k = \{x|f_k(x) \geq f_{k,\alpha}\}$ and $f_{k,\alpha}$ is the lower α percentile of $f_k(x)$ under the distribution of class k. In Figure 1, the second-row plots show the result of the density-level set with $\alpha = 0.05$. Again, the middle left plot contains the training samples, and the middle right plot contains the test samples. The black and blue dashed ellipses are the boundaries for the decision regions A_1 and A_2 from the oracle density-level sets (with given densities). We call the prediction with $C(x) = \emptyset$ as the abstention. In this example, the oracle density-level sets have successfully abstained from predictions while assigning correct labels for most training samples. The density-level set is suggested as a way of making predictions with abstention in Hechtlinger et al. (2018).

However, the density-level set has its drawbacks. It does not try to utilize information comparing different classes, potentially leading to a large deterioration in performance. Figure 2 shows another example where the oracle density-level set has less than ideal performance. In this example, we have two classes with $x \in \mathbb{R}^{10}$, and the two classes are well separated in the first dimension and follow the standard normal distribution in other dimensions. In Figure 2, we show only the first two dimensions. In the left plot of Figure 2, we have coloured the samples based on their actual class. The black points represent class 1, and the blue points represent class 2. In the middle plot of Figure 2, we have coloured the data based on their oracle density-level set results: x is coloured green if $C(x) = \{1, 2\}$, black if $C(x) = \{1\}$, blue if $C(x) = \{2\}$ and red if $C(x) = \emptyset$. Even

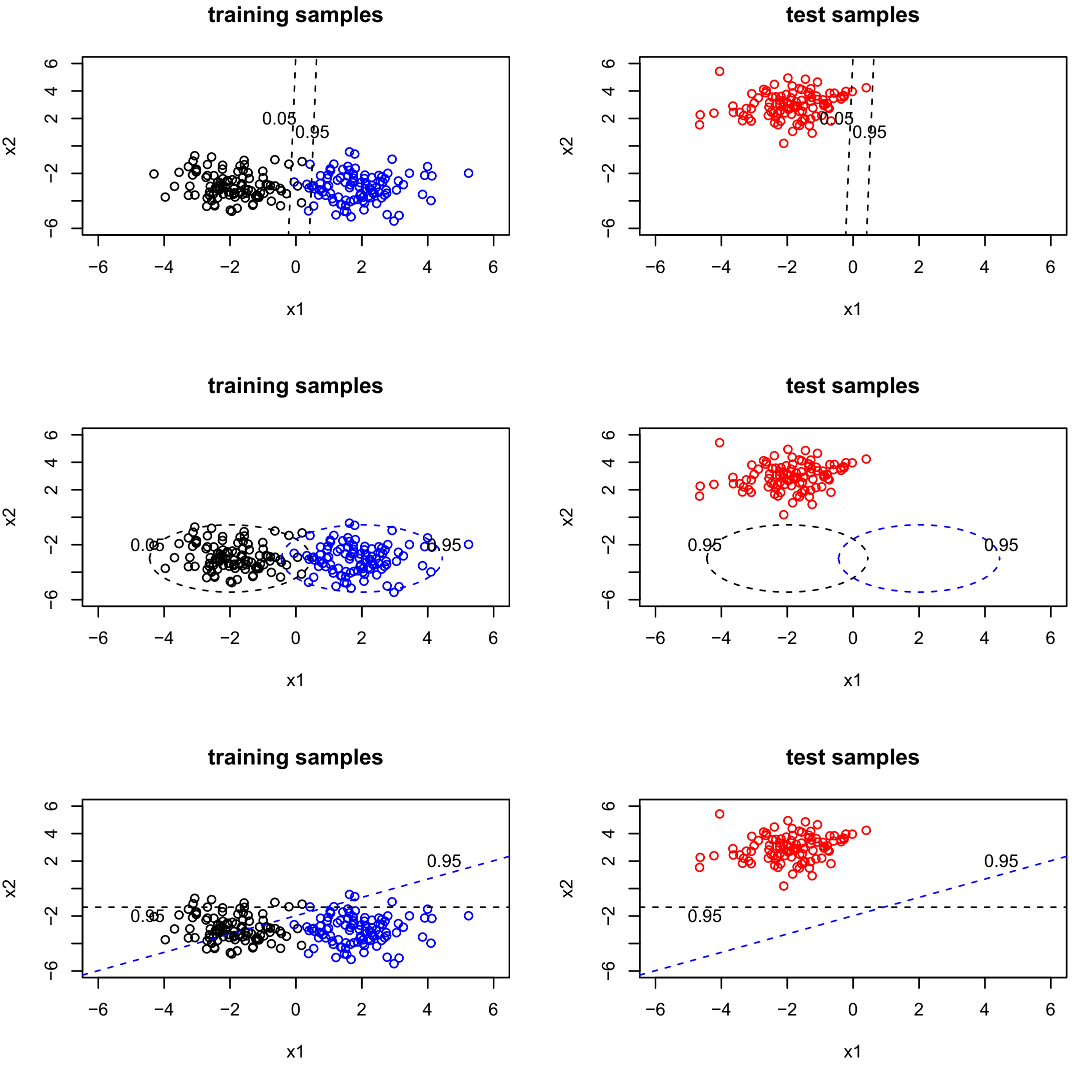

FIGURE 1 Illustrative example I. We use black/blue points to represent class 1/2 in the training samples and red points to represent the test samples. In the first row, the black and blue dashed lines are the boundaries for $P(y = 1|x) = 0.05$ and $P(y = 1|x) = 0.95$ based on the logistic model. In the second row of Figure 1, the dashed curves' interior represents the density-level sets achieving 95% coverage for class 1 and class 2. In the third row, the dashed lines represent decision boundaries for BCOPS achieving 95% coverage for class 1 and class 2 [Colour figure can be viewed at wileyonlinelibrary.com

though classes 1 and 2 can be well separated in the first dimension, we still have $C(x) = \{1, 2\}$ for a large portion of data, especially for samples from class 2.

In this paper, we propose a method called BCOPS (balanced and conformal optimized prediction set) to construct prediction set $C(x) \subseteq \{1, 2, \ldots, K\}$ for each x. The goals of BCOPS are to make good predictions for samples that are similar to the training data and refrain from making predictions otherwise. BCOPS usually has better performance in the test data than the density-level set because it combines information from different classes and unlabelled test samples when constructing $C(x)$. As an illustration, in the first illustrative example, we show the decision boundaries of class 1/2 using BCOPS with the black/blue dashed lines. We see that BCOPS refrains from making predictions for the outliers. In the

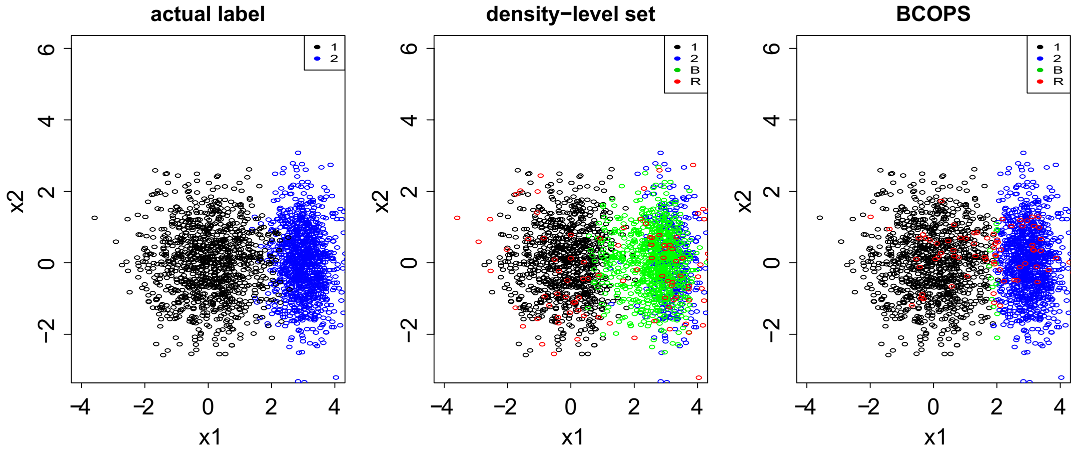

FIGURE 2 Illustrative example II. We have two classes with $x \in \mathbb{R}^{10}$, and the two classes are well separated in the first dimension and follow the standard normal distribution in other dimensions. The leftmost plot of Figure 2 shows the data coloured with their actual class. The black points represent class 1, and the blue points represent class 2. The middle/ right plot of Figure 2 shows the data with colour corresponding to their density-level set/ BCOPS: x is coloured green if $C(x) = \{1,2\}$, black if $C(x) = \{1\}$, blue if $C(x) = \{2\}$ and red if $C(x) = \varnothing$ [Colour figure can be viewed at wileyonlinelibrary.com

second illustrative example, we show the classification result using BCOPS in the right plot of Figure 2, which suggests that BCOPS distinguishes class 1 from class 2 in the test data much better than the density-level set. We also describe a new regression-based method to evaluate outlier detection ability under some assumptions on how the test data may differ from the training data.

The paper is organized as follows. In Section 2, we first describe our model and related works, then we will introduce BCOPS. In Section 3, we will describe methods to evaluate the performance regarding outlier detection. Some asymptotic behaviours of our proposals are given in Section 4. Finally, we provide real data examples in Section 5.

2 | BCOPS: MODELS AND METHODS

2.1 | A mixture model

It is often assumed that the distribution of the training data and out-of-sample data are the same. Let $\pi_k \in (0, 1)$ be the proportion samples from class $k \in \{1, ..., K\}$ in the training data, with $\sum_{k=1}^{K} \pi_k = 1$. Let $f_k(x)$ be the density of x from class k, and $f(x)/f_{test}(x)$ be the marginal in/out-of-sample densities. Under this assumption, we know that

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x), \;\; f_{test}(x) = \sum_{k=1}^{K} \pi_k f_k(x). \tag{1}$$

In other words, $f_{test}(x)$ can be written as a mixture of $f_k(x)$, and the mixture proportions π_k remain unchanged.

In this paper, we allow for distributional changes. We assume that the out-of-sample data may have different mixture proportions $\tilde{\pi}_k \in [0, 1)$ for $k \in \{1, ..., K\}$, and a new class **R** (outlier class) that is unobserved in the training data. We let

$$f_{test}(x) = \sum_{k=1}^{K} \tilde{\pi}_k f_k(x) + \epsilon \cdot e(x), \tag{2}$$

where $e(x)$ is the density of x from the outlier class, $\varepsilon \in [0, 1)$ is its proportion and $\sum_{k=1}^{K} \tilde{\pi}_k + \epsilon = 1$.

Under this new model assumption, we want to find a prediction set $C(x)$ that aims to minimize the length of $C(x)$ averaged over a properly chosen importance measure $\mu(x)$, and with guaranteed coverage $(1 - \alpha)$ for each class. Let $|C(x)|$ be the size of $C(x)$. We consider the optimization problem $\mathcal{P}$ below:

$$\begin{aligned} &\min \int |C(x)| \mu(x) dx \\ &s.t.\ \ P_k(k \in C(x)) \geq 1 - \alpha,\ \forall k = 1, \ldots, K. \end{aligned} \tag{3}$$

$P_k(\mathcal{A})$ is the probability of event $\mathcal{A}$ under the distribution of class k and $\mu(x)$ is a weighting function which we will choose later to tradeoff classification accuracy and power of outlier detection. The constraint (3) says that we want to have $k \in C(x)$ for at least $(1 - \alpha)$ of samples that are actually from an observed class k (coverage requirement). If $C(x) = \emptyset$, x is considered an outlier at the given level α, and we will refrain from making a prediction. Here, we aim for class-wise coverage rates instead of an average coverage rate. When controlling for the average coverage rate, the achieved coverage can be sensitive to misspecifications of test data distribution. In contrast, we have a coverage guarantee in the worst case scenario when controlling for the class-wise coverage rates.

It is easy to check that problem $\mathcal{P}$ can be decomposed into K independent problems for different classes, referred to as problem $\mathcal{P}_k$:

$$\min \int \mathbb{1}_{x \in A_k} \mu(x) dx \tag{4}$$

$$s.t.\ \ P_k(x \in A_k) \geq 1 - \alpha. \tag{5}$$

Let A_k be the solution to problem $\mathcal{P}_k$, then the solution to problem $\mathcal{P}$ is $C(x) = \{k : x \in A_k\}$. The set A_k has an explicit form using the density functions.

For an event $\mathcal{A}$, let $P_F(\mathcal{A})$ be the probability of $\mathcal{A}$ under distribution F and let $Q(\alpha;g,F)$ be the lower α percentile of a real-valued function $g(x)$ under distribution F:

$$Q(\alpha; g, F) = \sup\{t:\ P_F(g(x) \leq t) \leq \alpha\}.$$

We use $Q(\alpha;\ g(x_1), \ldots, g(x_n))$, or $Q(\alpha;\ g(x_{1:n}))$ to denote the lower α percentile of $g(x)$ using the empirical distribution $\{x_1, \ldots, x_n\}$. Let F_k be the distribution of x from class k. The solution to problem $\mathcal{P}_k$ is given by Proposition 1.

Proposition 1 *The solution to problem $\mathcal{P}_k$ can be expressed as,*

$$A_k = \{x : v_k(x) \geq Q(\alpha; v_k, F_k)\},\ \ v_k(x) = \frac{f_k(x)}{\mu(x)}. \tag{6}$$

Proof Problem $\mathcal{P}_k$ can be viewed a hypothesis testing problem where the null hypothesis is H_0: $x \sim f_k(x)$, the alternative is H_1: $x \sim \mu(x)$, and we want to find the decision region A_k for the null hypothesis corresponding to the most powerful level-α test. By Neyman–Pearson Lemma (Neyman & Pearson, 1933), we have $A_k = \{x: v_k(x) \geq Q(\alpha; v_k, F_k)\}$.

The same arguments for deriving the optimal solutions are also used in Lei et al. (2013) and Sadinle et al. (2019), with a different optimization objective.

We call the set A_k in Equation (6) the oracle set for class k. The oracle prediction set $C(x)$ for problem $\mathcal{P}$ is then constructed using the oracle sets A_k. Since $A_k(x)$ depends only on the ordering of $v_k(x)$, we can also use any order-preserving transformation of $v_k(x)$ when constructing A_k. (An order-preserving transformation o: $\mathbb{R} \to \mathbb{R}$ satisfies that $v_1 < v_2 \Leftrightarrow o(v_1) < o(v_2)$ for $\forall\, v_1, v_2 \in \mathbb{R}$.)

2.2 | Strategic choice for the weighting function $\mu(x)$

How should we choose $\mu(x)$? For any given $\mu(x)$, while the coverage requirement is satisfied by definition, the choice of $\mu(x)$ influences how well we separate different observed classes from each other, and inliers from outliers. In practice, except for the coverage, people also want to minimize the misclassification loss averaged over the out-of-sample data, for example,

$$\text{Err} = E_{x,y\sim f_{test}} \sum_{k\neq y} \mathbb{1}_{k\in C(x)}. \tag{7}$$

Recall that f_{test}(x) is the marginal density of X. With some abuse of notations, here, we use $x,\ y \sim f_{test}$ to represent that $x,\ y$ are generated from the test data distribution. To be more specific, Y takes label k with probability $\tilde{\pi}_k$ and R with probability ε. Given the label Y, X is generated according to the density for each class: $X|Y = k$ has density $f_k(x)$ for $k = 1, \ldots, K$ and $X|Y = R$ has density $e(x)$. According to Proposition 2, $f_{test}(x)$ is a natural choice for $\mu(x)$ if we want to minimize the above misclassification loss.

Proposition 2 *The problem $\mathcal{P}$ with $\mu(x) = f_{test}(x)$ is equivalent to the problem $\tilde{\mathcal{P}}$ below that minimizes the misclassification loss,*

$$\min \text{Err}, \quad s.t.\ \ P_k(k \in C(x)) \geq 1-\alpha,\ \ \forall k = 1,\ \ldots,\ K.$$

Proof Problem $\tilde{\mathcal{P}}$ can be decomposed into the following K independent problems,

$$\min \int \mathbb{1}_{x\in A_k}(f_{test} - \tilde{\pi}_k f_k)(x)dx, \quad s.t. \quad P_k(x \in A_k) \geq 1-\alpha,$$

where $(f_{test} - \tilde{\pi}_k f_k)$ denotes the unnormalized density of test data excluding class k. Same as in the proof of Proposition 1, by Neyman–Pearson Lemma, the solution to $\tilde{\mathcal{P}}$ is $A_k = \{x: \tilde{v}_k(x) \geq Q(\alpha; \tilde{v}_k, F_k)\}$, where $\tilde{v}_k = \frac{f_k(x)}{f_{test}(x) - \tilde{\pi}_k f_k(x)}$. By Proposition 1, $\{x: v_k(x) \geq Q(\alpha; v_k, F_k)\}$ with $v_k(x) = \frac{f_k(x)}{f_{test}(x)}$ is the solution to $\mathcal{P}$ when $\mu(x) = f_{test}(x)$. Since $v_k(x)$ is an order-preserving transformation of $\tilde{v}_k$, we have that problem $\tilde{\mathcal{P}}$ and problem $\mathcal{P}$ with $\mu(x) = f_{test}(x)$ are equivalent.

BCOPS constructs $\hat{C}(x)$ to approximate the oracle solution $C(x)$ to $\mathcal{P}$ with $\mu(x) = f_{test}(x)$. This optimization criterion depends on the unlabelled test data distribution. In practice, the unlabelled test data set can be much larger than the data sets with labels. Hence, we may provide good estimation for

$f_{test}(x)$ and its corresponding density ratios $v_k(x)$ for complicated $v_k(x)$. Some previous work is closely related or equivalent to other choices of $\mu(x)$. For example, the density-level set described in the introduction can also be written equivalently to the solution of problem $\mathcal{P}$ with $\mu(x)\propto 1$ (Lei et al., 2013).

Our prediction set $\widehat{C}(x)$ is constructed by combining properly chosen learning algorithms with the conformal prediction idea to meet the coverage requirement without distributional assumptions (Vovk et al., 2005). For the remainder of this section, we first have a brief discussion of some related methods in Section 2.3, and review the idea of conformal prediction in Section 2.4. We give details of BCOPS in Section 2.5, and we show a simulated example in Section 2.6.

2.3 | Related work

The new model assumption described in Equation (2)

$$f_{test}(x) = \sum_{k=1}^{K} \tilde{\pi}_k f_k(x) + \epsilon \cdot e(x).$$

It allows for changes in mixture proportions, and treats outliers as part of the distributional change that cannot be explained. This assumption is different from the assumption that fixes $f(y|x)$ and allows $f(x)$ to change. We use this model because $P(y = k)$ is much easier to estimate than $f(x)$, and it explicitly describes what kind of data we would like to reject.

Without the additional term $\varepsilon\cdot e(x)$, the change in mixture proportions π_k is also called label shift/target shift (Lipton et al., 2018; Zhang et al., 2013). In Zhang et al. (2013), the authors also allow for a location-scale transformation in $x|y$. When only the label shift happens, a better prediction model can be constructed through sample reweighting using the labelled training data and unlabelled out-of-sample data.

The term $\varepsilon\cdot e(x)$ corresponds to the proportion and distribution of outliers. We do not want to make a prediction if a sample comes from the outlier distribution. There is an enormous literature on the outlier detection problem, and interested readers can find a thorough review of traditional outlier detection approaches in Hodge and Austin (2004); Chandola et al. (2009). Here, we go back to the density-level set. Both BCOPS and the density-level set are based on the idea of prediction sets/tolerance regions/minimum volume sets(Chatterjee & Patra, 1980; Li & Liu, 2008; Wald, 1943; Wilks, 1941). For each observation x, we assign it a prediction set $C(x)$ instead of a single label to minimize a certain objective, usually the length or volume of $C(x)$, while having some coverage requirements. As we pointed out before, the density-level set is the optimal solution when $\mu(x)\propto 1$(Hechtlinger et al., 2018; Lei et al., 2013; Sadinle et al., 2019):

$$\min \int |C(x)| dx$$
$$s.t.\ \ P_k(k \in C(x)) \geq 1-\alpha,\ \forall k=1,\ \ldots,\ K.$$

Although the density-level prediction set achieves optimality with $\mu(x)$ being the Lebesgue measure, it is not obvious that $\mu(x)\propto 1$ is a good choice. In Section 1, we observe that the density-level set has lost the contrasting information between different classes, and choosing $\mu(x)\propto 1$ is a reason for why it happens. The work of Herbei and Wegkamp (2006); Bartlett and Wegkamp (2008) is closely related to the case where $\mu(x) = f(x)$, the in-sample density. When $\mu(x) = f(x)$, we encounter the same problem as in the usual classification methods that learn $P(y|x)$ and could assign confident predictions to test samples that are far from the training data. In contrast, BCOPS chooses

$\mu(x)$ to utilize as much information as possible to minimize $\mathrm{Err} = E_{x,y\sim f_{test}} \sum_{k\neq y} \mathbb{1}_{k\in C(x)}$, leading to good predictions for inliers and abstentions for the outliers.

In recent independent work, Barber et al. (2019) also used information from the unlabelled out-of-sample data under a different model and goal.

2.4 | Conformal prediction

BCOPS constructs $\widehat{C}(x)$ using the method of conformal prediction. We give a brief recap of the conformal prediction here for completeness.

Let $X_1, \ldots, X_n$ be n random observations. Conformal inference provides a way to construct a level $(1-\alpha)$ prediction interval for a new observation X_{n+1} assuming only data exchangeability. The key step is to construct a real-valued conformal score function of any observation value x. This score function may depend on $\{X_1, \ldots, X_n, X_{n+1}\}$ as long as this dependence is permutation invariant. To emphasize this potential dependence, we write the conformal score function as $\sigma(\{X_1, \ldots, X_n, X_{n+1}\}, x)$, where the first argument is an unordered set of our $n+1$ random observations.

Let $\sigma_i = \sigma(\{X_1, \ldots, X_n, X_{n+1}\}, X_i)$, for $i = 1, \ldots, n+1$, be the score function evaluated at each of the observations. Let $s_{n+1} = \frac{1}{n+1}\sum_{j=1}^{n+1} \mathbb{1}_{\sigma_{n+1}\geq\sigma_j}$ and $s_{n+1}(x) = \frac{1}{n+1}\sum_{j=1}^{n+1} \mathbb{1}_{\sigma_{n+1}\geq\sigma_j}|_{X_{n+1}=x}$ be s_{n+1} evaluated at $X_{n+1} = x$. The coverage result (Vovk et al., 2005) of the conformal prediction states that, when $X_1, \ldots, X_n, X_{n+1}$ are drawn exchangeably (e.g. if $X_1, \ldots, X_n, X_{n+1} \overset{i.i.d}{\sim} \mathcal{P}$ for an arbitrary distribution $\mathcal{P}$), then, $\sigma_1, \cdots, \sigma_{n+1}$ are exchangeable, and we have $P(X_{n+1}\in A)\geq 1-\alpha$ where $A = \{x: s_{n+1}(x) \geq \frac{\lfloor (n+1)\alpha \rfloor}{n+1}\}$.

The sample splitting conformal score $\sigma(\{X_1, \ldots, X_n\}, x) = \sigma(x)$ is a most familiar conformal score, where the conformal score function is independent of the new observation X_{n+1} and observations $\{X_1, \ldots, X_n\}$ used in s_{n+1} (but can depend on other training data). We call a procedure the sample-splitting conformal construction if it relies on this independence.

Another simple example where the conformal score function relies on the permutation invariance is given below:

$$\sigma(\{X_1, \ldots, X_n, X_{n+1}\}, x) = -(x - \frac{\sum_{i=1}^{n+1} X_i}{n+1})^2.$$

Since $\sigma(\{X_1, \ldots, X_n, X_{n+1}\}, x)$ is permutation invariant on $\{X_1, \ldots, X_n, X_{n+1}\}$, we will have the desired coverage with this score function. We call a procedure the data augmentation conformal construction if it relies on the permutation invariance but not the independence between observations and the conformal score function.

In BCOPS, we estimate $v_k(x)$ used in Equation (6) through either a sample-splitting conformal construction or a data-augmented conformal construction to have the coverage validity without distributional assumptions.

2.5 | BCOPS

With the observations from the out-of-sample data, we can consider the problem $\mathcal{P}_k$ directly with $\mu(x) = f_{test}(x)$:

$$\min \int \mathbb{1}_{x\in A_k} f_{test}(x)dx$$
$$s.t.\ \ P_k(x\in A_k)\geq 1-\alpha.$$

It has the solution

$$A_k = \{x\colon v_k(x) \geq Q(\alpha; v_k, F_k)\}, \quad v_k(x) = \frac{f_k(x)}{f_k(x) + f_{test}(x)}.$$

Here, we have applied an order-preserving transformation to the density ratio $\frac{f_k(x)}{f_{test}(x)}$ to get $v_k(x)$. A test point x will have an empty prediction set and be deemed an outlier if each class density $f_k(x)$ is low compared to the marginal density $f_{test}(x)$.

Remark 1 A_k does not favour regions where class k and other classes both have high densities. As a result, we may output $C(x) = \varnothing$ at places where multiple classes have high densities. However, in practice, we prefer a small α because a large α means a poor coverage guarantee. When α is small, A_k will not exclude a large portion of samples with high densities in class k by definition.

If we knew $f_k(x)$, $f_{test}(x)$ and hence $v_k(x)$, we can have the oracle A_k and $C(x)$. They are, of course, unknown. One could use the density estimation to approximate them, but this would suffer in high dimensions. Instead, our proposed BCOPS constructs a set $\widehat{A}_k$ to approximate the above A_k using conformal prediction. The conformal score function is learned via a supervised binary classifier $\mathcal{L}$. When the density ratio $v_k(x)$ has a low-dimensional structure, the learned density ratio from the binary classifier is often much better than the estimated density ratio from comparing two density estimations. Since we have used conformal construction when constructing $\widehat{A}_k$, the constructed prediction set $\widehat{C}(x) = \{k\colon x \in \widehat{A}_k\}$ will have finite sample coverage validity. Algorithm 1 gives details of its implementation, which consists of three main steps (1) data splitting of both the training and test data, (2) learning conformal score functions using onefold of the data and (3) performing conformal inference using the other fold of the data given the learned score functions.

Algorithm 1 BCOPS

```
function BCOPS($D^{tr}$, $D^{te}$, $\alpha$, $\mathcal{L}$)
Input  : Coverage level $\alpha$, a binary classifier $\mathcal{L}$, labeled training data $D^{tr}$, unlabeled
         test data $D^{te}$.
Output: For each $x \in D^{te}$, the prediction set $\widehat{C}(x)$.

  • Randomly split the training and test data into two roughly equal folds $\{D_1^{tr}, D_2^{tr}\}$
    and $\{D_1^{te}, D_2^{te}\}$. Let $D_{k,1}^{tr}$, $D_{k,2}^{tr}$ contain samples from class $k$ in $D_1^{tr}$, $D_2^{tr}$ respectively

  • For each $k$, apply $\mathcal{L}$ to $\{D_{k,1}^{tr}, D_1^{te}\}$ to separate $D_{k,1}^{tr}$ from $D_1^{te}$ and learn a prediction
    function $\hat{v}_{k,1}(x)$ for $v_k(x) = \frac{f_k(x)}{f_k(x)+f_{test}(x)}$. Do the same thing with $\{D_{k,2}^{tr}, D_2^{te}\}$, and
    denote the learned prediction function by $\hat{v}_{k,2}(x)$.

  • For $x \in D^{te}$, let $t$ be $\in \{1,2\}$ such that $x \in D_t^{te}$, and $t' = \{1,2\} \setminus t$. We construct

        $s_k(x) = \frac{1}{|D_{k,t}^{tr}| + 1} \sum_{z \in D_{k,t}^{tr} \cup \{x\}} \mathbb{1}_{\hat{v}_{k,t'}(x) \geq \hat{v}_{k,t'}(z)}$,

    and $\widehat{A}_k = \{x : s_k(x) \geq \frac{\lfloor (|D_{k,t}^{tr}|+1)\alpha \rfloor}{|D_{k,t}^{tr}|+1}\}$, $\widehat{C}(x) = \{k : x \in \widehat{A}_k\}$.
```

Remark 2 Algorithm 1 uses the sample-splitting conformal construction. We can also use the data augmentation conformal prediction. For a new observation x and class k, we can consider the augmented data $D_k = \{x_{k,1}, \ldots, x_{k,n_k}, x\}$, where $x_{k,i}$ for $i = 1, \ldots, n_k$ are samples from class k in the training data set. Then, we build a classifier separating D_k from $D^{te} \setminus \{x\}$, the test data excluding x. For each new observation, we let the trained prediction model be $\hat{v}_k(.\,|x)$ and let

$$s_k(x) = \frac{1}{n_k+1} \sum_{z \in D_k^{tr}} \mathbb{1}_{\hat{v}_k(x|x) \geq \hat{v}_k(z|x)}, \quad \hat{A}_k = \left\{ x: s_k(x) \geq \frac{\lfloor (n_k+1)\alpha \rfloor}{n_k+1} \right\}, \quad \hat{C}(x) = \{k | x \in \hat{A}_k\}.$$

By exchangeability, we can also have a finite sample coverage guarantee using this approach (data augmentation conformal construction). However, we use the sample-splitting conformal construction in this paper to avoid a huge computational burden at the cost of being less efficient.

Remark 3 The users can choose their preferred classifier to learn the density ratios here. The coverage guarantee will always be satisfied, but the classification error will depend on the classifier for any fixed level α. A classifier that outputs only class labels is usually a poor choice. The better the classifier learns the density ratios, the less the classification error tends to be. For example, in our real data examples, we used random forest and logistic regression as our classifiers. While both of them have reasonably good performance, the random forest performs relatively better. We also observe that, maybe as expected, classifiers that achieve low in-sample classification errors tend to achieve low out-of-sample classification errors using BCOPS.

By exchangeability, we know that the above procedure has finite sample validity (Cadre et al., 2009; Lei, 2014; Lei & Wasserman, 2014; Lei et al., 2013; Vovk et al., 2009):

Proposition 3 *$\hat{C}(x)$ is finite sample valid:*

$$P_k(k \in \hat{C}(x)) \geq 1 - \alpha, \ \forall k = 1, \ldots, K.$$

Remark 4 The coverage guarantee is in expectation. For every realization of training and test sets, the achieved actual coverage can be higher or lower than $1-\alpha$. However, the (expected) coverage defined in Proposition 3 is achieved for all classes if we repeat the same procedure infinitely many times.

Algorithm 1 produces a prediction set $\hat{C}(x)$ as good as the oracle prediction set $C(x)$ (we will refer to it as the asymptotic optimality) for minimizing the objective if $\hat{v}_{k,1}$, $\hat{v}_{k,2}$ are good estimations of $v_k(x)$, and $v_k(x)$ is well-behaved. A more rigorous statement can be found in Section 4.

2.6 | A simulated example

In this section, we provide a simple simulated example to illustrate the differences between three different methods: (1) BCOPS, (2) density-level set where $\mu(x) \propto 1$ in $\mathcal{P}$, and (3) in-sample ratio set where $\mu(x) = f(x)$ in $\mathcal{P}$. All three methods have followed the sample-splitting conformal construction with the level $\alpha = 0.05$. For both BCOPS and the in-sample ratio set, we have used the random forest classifier to learn $v_k(x)$ ($v_k(x) = \frac{f_k(x)}{f_k(x)+f_{test}(x)}$ for BCOPS and $v_k(x) = \frac{f_k(x)}{f(x)}$ the in-sample ratio set).

We let $x \in \mathbb{R}^{10}$ and generate 1000 training samples, half from class 1 and the other half from class 2. The feature x is generated as

$$x_1 \sim \begin{cases} N(0,1) & \text{if } y=1 \\ N(3,0.5) & \text{if } y=2 \end{cases}, \quad x_j \sim N(0,1),\ j=2,\ \ldots,\ 10.$$

We have 1500 test samples, one-third from class 1, one-third from class 2 and the other one-third of them follow the distribution (outliers, class **R**):

$$x_2 \sim N(3,1),\ \ x_j \sim N(0,1),\ j \neq 2.$$

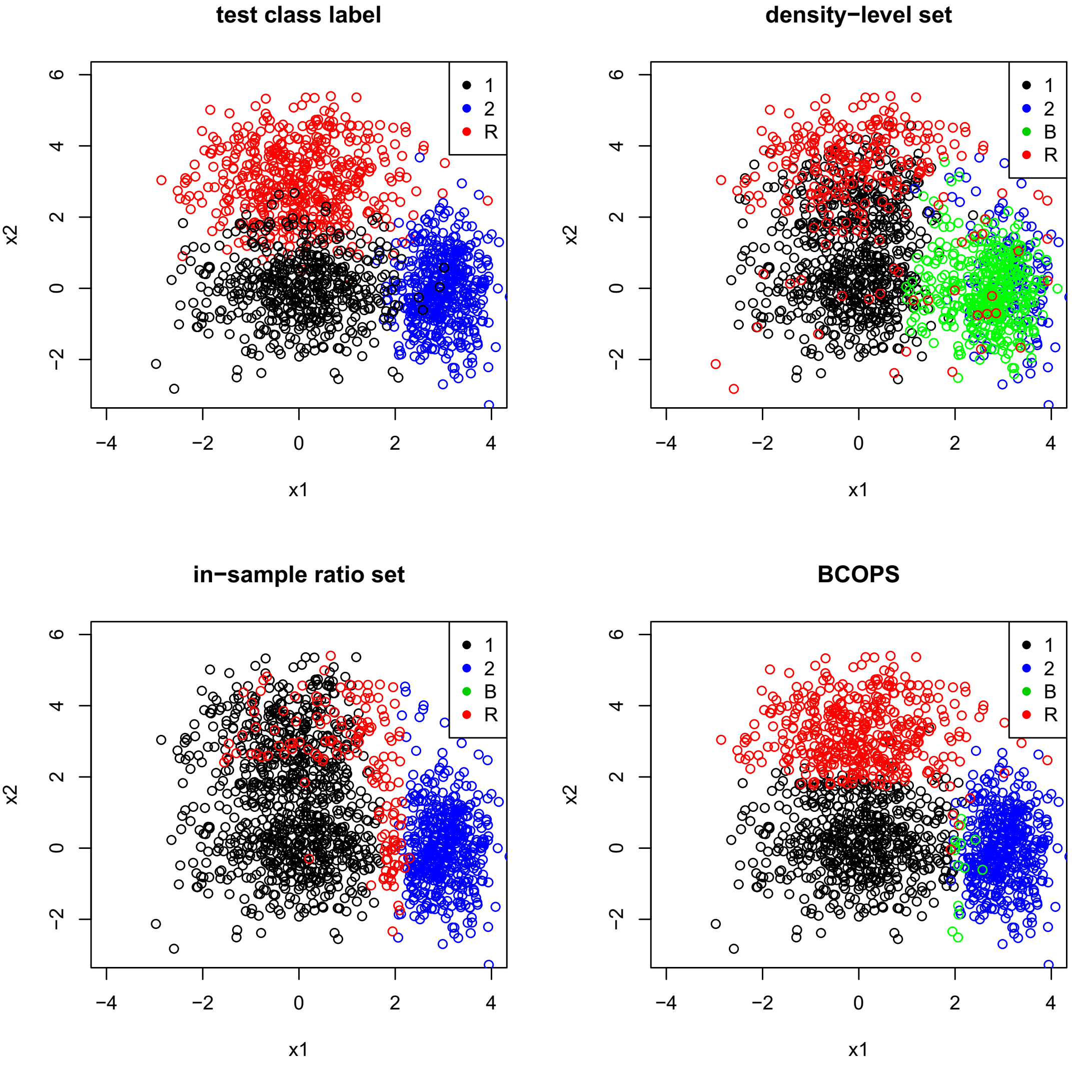

FIGURE 3 A simulated example. The upper left plot shows the class label for each sample in the test data set. The upper right, lower left, and lower right panels correspond to the prediction results using the density-level set, in-sample ratio set and BCOPS. The upper left plot colours the data based on its correct label, and is coloured black/blue/red if its label is class 1/class 2/outliers. For the remaining three plots, a sample is coloured black if $C(x) = \{1\}$, blue if $C(x) = \{2\}$, green if $C(x) = \{1,2\}$ and red if $C(x) = \varnothing$ [Colour figure can be viewed at wileyonlinelibrary.com

TABLE 1 An illustrative example. The second column **R** is the abstention rate for outliers; the third column is the prediction accuracy; the fourth and fifth columns are coverages for samples from class 1 and class 2

	R	Accuracy	Coverage I	Coverage II
Density level	0.46	0.57	0.96	0.97
In-sample ratio	0.20	0.94	0.94	0.95
BCOPS	0.84	0.95	0.96	0.97

In this example, we let the learning algorithm $\mathcal{L}$ be the random forest classifier. Figure 3 plots the first two dimensions of the test samples and shows the regions with 95% coverage for BCOPS, density-level set and in-sample ratio set. The upper left plot colours the data based on its correct label, and is coloured black/blue/red if its label is class 1/class 2/outliers. For the remaining three plots, a sample is coloured black if $C(x) = \{1\}$, blue if $C(x) = \{2\}$, green if $C(x) = \{1,2\}$ and red if $C(x) = \emptyset$. Table 1 shows results of the abstention rate in outliers (the higher, the better), the prediction accuracy for data from class 1 and 2 (a prediction is called correct if $C(x) = \{y\}$ for a sample (x, y)), coverages for class 1 and class 2.

We can see that the BCOPS achieves a much higher abstention rate in outliers and much higher accuracy in the observed classes compared with the density-level set. While the in-sample ratio set has similar accuracy as the BCOPS, it has the lowest abstention rate in this example.

We also observe that a small α might lead to making predictions on outliers that are far from the training data (especially for the density-level set and the in-sample ratio set). While the power for outlier detection varies for different approaches and problems, we want to learn about the outlier abstention rate no matter what method we are using. In Section 3, we provide methods for this purpose.

3 | OUTLIER ABSTENTION RATE AND FALSE LABELLING RATE

In the previous section, we proposed BCOPS for prediction set construction at a given level α. In this section, we describe a regression-based method to estimate the test set mixture proportions $\tilde{\pi}_k$ for $k = 1, \ldots, K$. We estimate the outlier abstention rate and FLR (false labelling rate) with this method. The outlier abstention rate is the expected proportion of outliers with an empty prediction set. The FLR is the expected ratio between the number of outliers with a non-empty prediction and the total number of observations with a non-empty prediction. For a fixed prediction set function $C(x)$, its outlier abstention rate γ and FLR are defined as

$$\gamma: = P_{\mathbf{R}}(C(x)=\emptyset),$$
$$\text{FLR}: = E\left(\frac{|\{x\in D^{te}: y(x)=\mathbf{R}, C(x)\neq\emptyset\}|}{|\{x\in D^{te}: C(x)\neq\emptyset\}|\vee 1}\right).$$

Here, we let $y(x)$ be the label of x for $x \in D^{te}$. The expectation is taken over the test data distribution. The outlier abstention rate is the power of $C(x)$ in terms of the outlier detection while FLR controls the per cent of outliers among samples with a non-empty prediction.

Information about the outlier abstention rate and FLR can be valuable when picking α. For example, while we want to have both small α and large γ, γ is negatively related to α. As a result, we might want to choose α based on the tradeoff curve of α and γ. There are different ways that we may want to utilize γ or FLR:

1. Set $\alpha \geq \alpha^*$ to control the FLR or γ. For example, let $\alpha^* = inf\{\alpha: \mathrm{FLR}(\alpha) \leq 10\%\}$ where $\mathrm{FLR}(\alpha)$ is the FLR at the given α.
2. Minimize a user-specified cost in terms of mis-coverage and FLR/ γ. For example, we may let $cost = \alpha + \mathrm{FLR}$, and we can choose an α to minimize the empirical estimation of the cost.
3. Set α without considering FLR or γ. However, in this case, we can still assign a score for each point to measure how likely it may be an outlier. For each point x, let $\alpha(x)$ be the largest α such that $C(x) = \varnothing$. We let its outlier score be $\gamma(x){:} = \gamma_{\alpha(x)}$, where $\gamma_{\alpha(x)}$ is the abstention rate at the required coverage is $(1 - \alpha(x))$. The interpretation of the outlier score is simple. If we want to refrain from making predictions for γ proportion of outliers, we do not predict a label for samples with $\gamma(x) \geq \gamma$ even if $C(x)$ itself is non-empty.

3.1 | Estimation of γ and FLR

When the proportion of outliers ε is greater than zero in the test samples, γ can also be expressed as

$$\gamma = \frac{E[\text{Number of outliers with } C(x) = \emptyset]}{\text{Total number of outliers}} = \frac{E[N_\emptyset] - \sum_{k=1}^{N} N\tilde{\pi}_k\gamma_k}{N(1 - \sum_{k=1}^{K} \tilde{\pi}_k)},$$

where N is the total number of test samples, $N_\emptyset$ is the total number of samples with abstention ($C(x) = \varnothing$), and $\gamma_k{:} = P_k(C(x) = \emptyset)$ is the abstention rate for class k. The FLR can be expressed as

$$\mathrm{FLR} = E[\frac{\overbrace{(N - N_\emptyset)}^{\text{All non-empty}} - \sum_{k=1}^{K} \overbrace{N\tilde{\pi}_k(1-\gamma_k)}^{\text{Class } k \text{ non-empty}}}{(N - N_\emptyset) \vee 1}].$$

When $\hat{\gamma}_k$ and $\hat{\pi}_k$, the estimates of γ_k and $\tilde{\pi}_k$, are available, we can construct empirical estimates $\hat{\gamma}$ and $\widehat{\mathrm{FLR}}$ of γ and FLR:

$$\hat{\gamma} = \frac{(N_\emptyset - \sum_{k=1}^{K} N\hat{\pi}_k\hat{\gamma}_k) \vee 0}{N(1 - \sum_{k=1}^{K} \hat{\pi}_k) \vee 1}, \quad \widehat{\mathrm{FLR}} = \frac{(N - N_\emptyset - \sum_k N\hat{\pi}_k(1 - \hat{\gamma}_k)) \vee 0}{(N - N_\emptyset) \vee 1}.$$

3.1.1 | Estimation of γ_k

We estimate γ_k using the empirical distribution for class k from the training data. More specifically, for BCOPS, we let

$$\hat{\gamma}_k = \frac{|\{x \in D_k^{tr}: \hat{C}(x) = \emptyset\}|}{|D_k^{tr}|},$$

where $\hat{C}(x)$ follows the same construction as in the BCOPS Algorithm: For $x \in D_{k,t}^{tr}$, we construct $\hat{C}(x)$ for x using training and test samples from fold $t' \in \{1, 2\} \setminus t$ as described in the BCOPS Algorithm.

3.1.2 | Estimation of $\tilde{\boldsymbol{\pi}}_k$

Let S_l be regions such that $P_{\mathbf{R}}(S_l) = 0$ for $l = 1, ..., K$. Then, by our model assumption, we know

$$P_{test}(S_l) = \sum_{k=1}^{K} \tilde{\pi}_k P_k(S_l).$$

Let P be the response vector with $\mathrm{P}_l = P_{test}(S_l)$ for $l = 1, ..., K$. Let $\mathbf{P}$ be a $K\times K$ design matrix with $\mathbf{P}_{l,k} = P_k(S_l)$ for k, $l = 1, ..., K$, and $\boldsymbol{\Sigma} = \mathbf{P}^T\mathbf{P}$. As long as $\boldsymbol{\Sigma}$ is invertible, $\tilde{\pi}$ is the solution to the regression problem that regresses P on $\mathbf{P}$. We now give a simple proposal to construct such S_l.

For a fixed function η: $\mathbb{R}^p \rightarrow \mathbb{R}^K$, let $g_{l,k}(.)$ be the density of η_l in class k. If the outliers happen with probability 0 at regions of η_l where class l has relatively high density, we can let S_l be the region with relatively high $g_{l,l}(.)$:

1. Let $\circ$ be the composition operator and $S_l = \{z{:}\, g_{l,l}(z) \geq Q(\zeta; g_{l,l}\circ\eta_l, F_l)\}$, $P_l = P_{test}(\eta_l(x) \in S_l)$, $\mathbf{P}_{l,k} = P_k(\eta_l(x) \in S_l)$ for a user-specific constant $\zeta \in (0,1)$ specifying the separation between inliers and outliers.
2. We would like to find the solution π to the linear system $J(\eta)$: $\mathrm{P} = \mathbf{P}\pi$, the oracle problem based on the function η.

We recommend taking $\eta_l(x) = \log\frac{f_l(x)}{f_{test}(x)}$, the log-odd ratio separating class l from the test data, since it automatically tries to separate class l from other classes, including the outliers.

Neither η nor P, $\mathbf{P}$ given η will be observed. In practice, we use sample-splitting to estimate η in onefold of the data and estimate P, $\mathbf{P}$ empirically in the other fold conditional on the estimated η. See Algorithm 2, MixEstimate (mixture proportion estimation) for details, which consists of three steps: (1) Sample splitting, (2) learning functions $\eta_l(x)$ using onefold of the data and (3) defining regions S_l and estimating the mixture proportions using the other fold of the data.

Algorithm 2 *MixEstimate*

```
function MixEstimate($D^{tr}$, $D^{te}$, $\zeta$, $\mathcal{L}$)
Input: Left-out proportion $\zeta$, a binary classifier $\mathcal{L}$, labeled training data $D^{tr}$, unlabeled test data $D^{te}$. By default, $\zeta = 0.2$.
Output: Estimated mixture proportion $\{\hat{\pi}_k,\ k = 1, \ldots, K\}$
(1) Randomly split the training and test data into roughly equal parts $\{D^{tr}_1, D^{tr}_2\}$ and $\{D^{te}_1, D^{te}_2\}$. For $t = 1, 2$, let $D^{tr}_{k,t}$ contain samples from class $k$ in $D^{tr}_t$ , and apply $\mathcal{L}$ to $\{D^{tr}_{k,t}, D^{te}_t\}$ to separate samples from class $k$ and the test set, we get $\hat{\eta}^t_l(x)$ as the estimate to $\eta_l(x) = \log\frac{f_l(x)}{f_{test}(x)}$.
(2) For fold $t = 1, 2$: let $t' = \{1, 2\} \setminus t$:
    • Let $\hat{F}^{t'}_l$ and $\hat{g}^{t'}_{l,l}(.)$ be empirical distribution of class $l$ and the gaussian kernel density estimation of the density of $\hat{\eta}^t_l(x)$ using $D^{tr}_{l,t'}$.
    • The empirical problem $\hat{J}(\hat{\eta}^t)$ is constructed with empirical probabilities of each class in fold $t'$ falling into regions $\widehat{S}_l = \{t : \hat{g}^{t'}_{l,l}(t) \geq Q(\zeta; \hat{g}^{t'}_{l,l} \circ \hat{\eta}^t, \hat{F}^{t'}_l)\}$. Let $\hat{\pi}^t_k$ for $k = 1, \ldots, K$ be the solutions to $\hat{J}(\hat{\eta}^t)$.
Output the average mixture proportion estimate and let $\hat{\pi}_k = \frac{\hat{\pi}_{k,1}+\hat{\pi}_{k,2}}{2}$, $\forall k = 1, \ldots, K$.
```

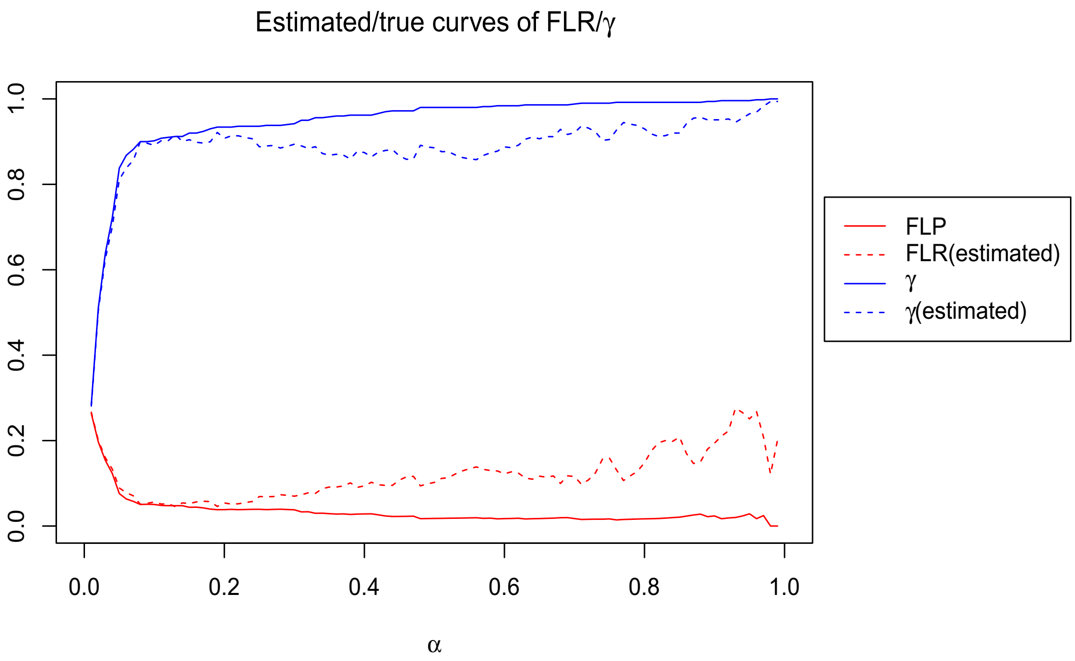

FIGURE 4 An illustrative example (continuation of Section 6). The red solid/dashed curves show the FLP and estimated FLR against different α. The blue solid/dashed curves show the actual outlier abstention rate γ and estimated γ against different α [Colour figure can be viewed at wileyonlinelibrary.com]

Remark 5

1. Our proposal is an extension to the BBSE method in Lipton et al. (2018) under the presence of outliers.
2. In practice, we can add a constraint on the optimization variables π_k and require that $\sum_{k=1}^{K} \pi_k \leq 1$ and $\pi_k \geq 0$. This constraint guarantees that both $\tilde{\pi}_k$ and the outlier proportion ϵ are non-negative.
3. The value of ζ indicates the desired separation between outliers and inlier in the given application. For example, we can interpret it as allowing outliers to exhibit the median behaviour of the inliers when ζ=0.5. Hence, we usually have $\zeta < 0.5$.

We show in Section 4 that the estimates from MixEstimate will converge to $\tilde{\pi}_k$ under proper assumptions. As a continuation of the example shown in Section 2.6. Figure 4 shows curves of estimated FLR, estimated outlier abstention rate $\hat{\gamma}$, as well as the FLP (false labelling proportion), which is the sample version of FLR using current test data, and the sample version of the γ against different α.

4 | ASYMPTOTIC PROPERTIES Of BCOPS AND MIXESTIMATE

Let n_k be the sample size of class k in the training data and n be the size of the training data. Let N be the size of the test data. In this section, we consider the asymptotic regime where $n \to \infty$, and assume that $lim_{n \to \infty} \frac{N}{n} \geq c$ for a constant $c > 0$ and $\frac{n_k}{n} \to c_k$ for a constant $c_k \in (0, 1), k = 1, \ldots, \quad K$. In this asymptotic regime, we show that

1. The prediction set $\widehat{C}(x)$ constructed using BCOPS achieves the same loss as the oracle prediction set $C(x)$ asymptotically if the estimation of $v_k(x) = \frac{f_k(x)}{f_k(x)+f_{test}(x)}$ is close to it, and under some conditions on the densities of x and distribution of $v_k(x)$ for $k = 1, \ldots, K$.
2. The mixture proportion estimations converge to the actual out-of-sample mixture proportions $\tilde{\pi}_k$ if the outliers are rare when the observed classes have high densities, and under some conditions on the densities of $\eta_l(x)$, functions used to construct S_l.

Proofs of Theorems 1 and 2 in this section are given in Appendix A.

4.1 | Asymptotic optimality of the BCOPS

Let $\hat{v}_k(x)$ be the estimate of $v_k(x)$, representing either $\hat{v}_{k,1}(x)$ or $\hat{v}_{k,2}(x)$ in the BCOPS Algorithm.

Assumption 1 *Densities $f_1(x), \ldots, f_k(x), e(x)$ are upper bounded by a constant. There exist constants $0 < c_1 \leq c_2$ and $\delta_0, \gamma > 0$, such that for $k = 1, \ldots, K$, we have*

$$c_1|\delta|^{\gamma} \leq |P_k(\{x|v_k(x) \leq Q(\alpha; v_k, F_k) + \delta\}) - \alpha| \leq c_2|\delta|^{\gamma}, \forall -\delta_0 \leq \delta \leq \delta_0.$$

Remark 6 We require that the underlying function $v_k(x)$ is neither too steep nor too flat around the boundary of the optimal decision region A_k. This requirement makes sure that the estimated boundary is not too sensitive to small errors in estimating $v_k(x)$, and the final loss is not too sensitive to small changes in the decision region.

Assumption 2 *The estimated function $\hat{v}_k(x)$ converges to the true model $v_k(x)$: there exist constants $B, \beta_1, \beta_2 > 0$ and a set A_n of x depending on n, such that*

$$P(\sup_{x\in A_n} |\hat{v}_k(x) - v_k(x)| < B(\frac{\log n}{n})^{\frac{\beta_1}{2}}) \to 1,\ P_{test}(x \in A_n) \geq 1 - B(\frac{\log n}{n})^{\frac{\beta_2}{2}},\ \textit{as } n \to \infty.$$

Remark 7 For such an assumption to hold in a high-dimensional setting, we usually need some parametric model assumptions and use a suitable parametric classifier $\mathcal{L}$ in BCOPS. For example, we could require the approximate correctness of the logistic model with nicely behaved features and sparse signals and let $\mathcal{L}$ be the logistic regression with lasso penalty (Van de Geer, 2008).

Theorem 1 *Let* C(x) *be the oracle BCOPS prediction set and* α > 0 *be any fixed level. Under Assumptions 1-2 and for a large enough constant* B, *we have*

$$P(\int (|\widehat{C}(x)| - |C(x)|) f_{test}(x)dx \geq B(\frac{\log n}{n})^{\frac{\min(\gamma\beta_1, \beta_2, 1)}{2}}) \to 0, \text{ as } n \to \infty.$$

4.2 | Asymptotic consistency of the mixture proportion estimates

In this section, let $\hat{\eta}$ represent $\hat{\eta}^t$ for $t = 1,2$ in the label shift estimation Algorithm (Algorithm 2). As defined in Section 3.1.2: ζ is a user-specific positive constant, $\eta: \mathbb{R}^p \to \mathbb{R}^K$ is a fixed function, and $g_{l,k}(t)$ is the density of η_l in class k, $S_l = \{t: g_{l,l}(t) \geq Q(\zeta; g_{l,l} \circ \eta_l, F_l)\}$ and P, $\mathbf{P}$ are the oracle response and design matrix based on η, $\mathbf{\Sigma} = \mathbf{P}^T\mathbf{P}$. We also let $\widehat{J}(\eta)$ be the problem with empirical P and $\mathbf{P}$ from fold $t = 1$ or 2, and h_n be the bandwidth of the Gaussian kernel density estimation in Algorithm 2. The bandwidth h_n satisfies $h_n \to 0$ and $\frac{\log n}{n h_n} \to 0$.

Assumption 3 *For $l = 1, \ldots, K$, the density $g_{l,k}(t)$ is bounded for $k = 1, \ldots, K, \mathbf{R}$ and $g_{l,l}(t)$ is Hölder continuous (e.g. there exist constants $0 < \gamma \leq 1$ and B, such that $|g_{l,l}(z) - g_{l,l}(z')| \leq B|z - z'|^\gamma$ for $\forall\, z, z' \in \mathbb{R}$).*

Assumption 4 *There exist constants $\gamma,\ c_1,\ c_2 > 0$ and $\delta_0 > 0$, such that for $\forall\, l = 1, \ldots, K$:*

$$c_1|\delta|^\gamma \leq |P_l(g_{l,l}(t) \leq Q(\zeta; g_{l,l} \circ \eta_l, F_l) + \delta) - \zeta| \leq c_2|\delta|^\gamma, \forall - \delta_0 \leq \delta \leq \delta_0.$$

Remark 8 Similar to Assumption 1, Assumption 4 requires that $g_{l,l}(t)$ is neither too steep nor too flat around the boundary of S_l.

Assumption 5 *$\mathbf{\Sigma}$ is invertible with the smallest eigenvalue $\sigma_{\min} \geq c$ for a constant $c > 0$.*

Assumption 6 *The outliers can be perfectly separated from the observed classes:*

$$P_{\mathbf{R}}(\eta_l(x) \in S_l) = 0,\ \forall l = 1, \ldots, K.$$

Remark 9

1. For Assumption 5, the invertibility of $\mathbf{\Sigma}$ guarantees that the oracle problem $\mathrm{P} = \mathbf{P}\pi$ has a unique solution. This requirement holds in the typical case where different classes can be reasonably well separated from each other. As a sufficient condition and an example, $\mathbf{\Sigma}$ is invertible if it is strictly diagonally dominant ($\sum_{k \neq l} P_k(\eta_l(x) \in S_l) < 1 - \zeta$ for all $l = 1, \ldots, K$). In practice, the invertibility can be checked empirically using the training data.
2. Assumption 6 says that the outliers can be separated from inliers. More specifically, outliers appear with negligible probabilities at where the density ratio between class κ and the test data is not small compared with observations from class k, for $k = 1, \ldots, K$. Since this density ratio is large at where density k has high density while other classes and outliers have low densities, Assumption 6 is partly consistent with the definition of outliers in many settings where people perceive samples that are far from inliers as abnormal.

Theorem 2 *Let $\{\widehat{\pi}_k, k = 1, \ldots, K\}$ be solutions to $\widehat{J}(\eta)$, and $\mathrm{P}_{\mathbf{R}} = (P_{\mathbf{R}}(\eta_1(x) \in S_1), \ldots, P_{\mathbf{R}}(\eta_K(x) \in S_K))^T$. Under Assumptions 3–5, we have*

$$\widehat{\pi}_k \xrightarrow{p} \tilde{\pi}_k + \epsilon \mathbf{\Sigma}^{-1}\mathbf{P}^T \mathrm{P}_{\mathbf{R}}, \quad \text{as } n \to \infty.$$

If Assumption 6 also holds, we have $\widehat{\pi}_k \xrightarrow{p} \tilde{\pi}_k$.

By the independence of the twofold, once we have learned $\widehat{\eta}$ in Algorithm 2, we can condition on it and treat it as fixed. Hence, we have Corollary 1 as a direct application of Theorem 2.

Corollary 1 *Let $\mathcal{A}_0$ be the event that Assumptions 3–6 hold for $\eta = \widehat{\eta}$. Let $\{\widehat{\pi}_k, k = 1, \ldots, K\}$ be the solution to $\widehat{J}(\widehat{\eta})$. If $P(\mathcal{A}_0) \to 1$ as $n \to \infty$, then, $\widehat{\pi}_k \xrightarrow{p} \tilde{\pi}_k$ as $n \to \infty$.*

5 | REAL DATA EXAMPLES

5.1 | MNIST handwritten digit example

We look at the MNIST handwritten digit data set (LeCun & Cortes, 2010). We let the training data contain digits 0–5, while the test data contain digits 0–9. We subsample 10,000 training data and 10,000 test data and compare the type I and type II errors using different methods. The type I error is defined as (1−coverage), and no type I error is defined for new digits unobserved in the training data. Rather than considering Err defined in Equation (7), we define the type II error as $\text{Err} = E_{x,y \sim f_{test}} \mathbb{1}_{|C(x) \setminus \{y\}| > 0}$ for observations from the test distribution (y is the true label of x). We use this loss since it is more similar to 0–1 classification loss and the type II error is well-defined with it: the prediction incurs zero loss if the prediction set contains and only contains the correct label and incurs a misclassification loss of value one otherwise. In this part, we consider three methods:

1. BCOPS: the BCOPS with the supervised learning algorithm $\mathcal{L}$ being random forest (rf) or logistic+lasso regression (glm).
2. DLS: the density-level set (with $\mu(x) = 1$ in problem $\mathcal{P}$) with the sample-splitting conformal construction.
3. IRS: the in-sample ratio set (with $\mu(x) = f(x)$ in problem $\mathcal{P}$) with the sample-splitting conformal construction and a supervised K-class classifier $\mathcal{L}$ to learn $\frac{f_k(x)}{f(x)}$. Here, we also let $\mathcal{L}$ be either random forest (rf) or multinomial+lasso regression (glm).

Figure 5 plots the nominal type I error for digits showing up in the training data (average). We can see that all methods can control their claimed type I error.

Figure 6 shows the plot of the type II error against type I error, separately for digits in and not in the training set, as α ranges from 0.01 to 0.99. We observe that

1. For the unobserved digits in the training data, we see that

   $$\text{DLS} < \text{IRS(glm)} < \text{IRS(rf)} < \text{BCOPS(glm)} < \text{BCOPS(rf)} \text{ (from worse to better)}$$

   BCOPS achieves the best performance by borrowing information from the unlabelled data. In this example, IRS also has better results compared with DLS for the unobserved digits. IRS depends only on the predictions from the given classifier(s). It does not prevent us from making predictions at locations with sparse observations if the classifiers themselves do not watch out for this. Although we can easily come up with situations where such methods fail entirely in terms of outlier detection, for example, the simulated example in Section 6, in this example, the dimension learned by IRS for in-sample classification is also informative for outlier detection.
2. For the observed digits in the training data, we see that

   $$\text{DLS} < \text{BCOPS(glm)} < \text{IRS(glm)} < \text{BCOPS(rf)} < \text{IRS(rf)}.$$

   DLS performs much worse than both BCOPS and IRS, and BCOPS performs slightly worse than IRS for a given learning algorithm $\mathcal{L}$.

Overall, in this example, BCOPS trades off some in-sample classification accuracy for higher power in outlier detection. In practice, we do not have access to the curves in Figure 6. While we can estimate the abstention rate for observed digits using the training data, we will not have

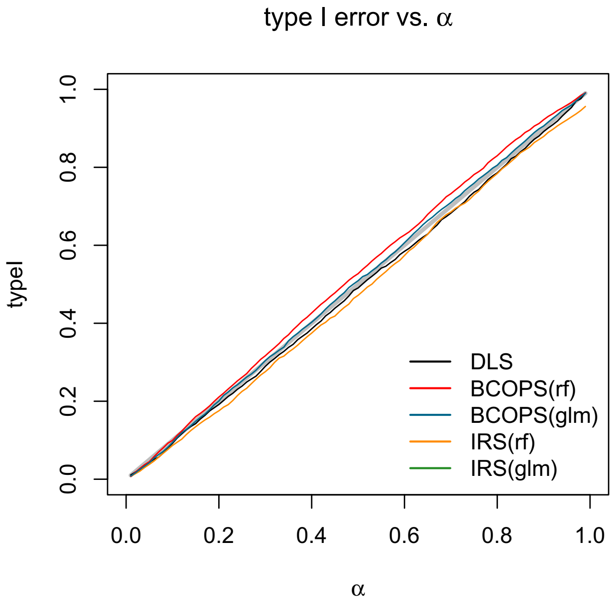

FIGURE 5 Actual type I error versus nominal type I error α for different methods [Colour figure can be viewed at wileyonlinelibrary.com]

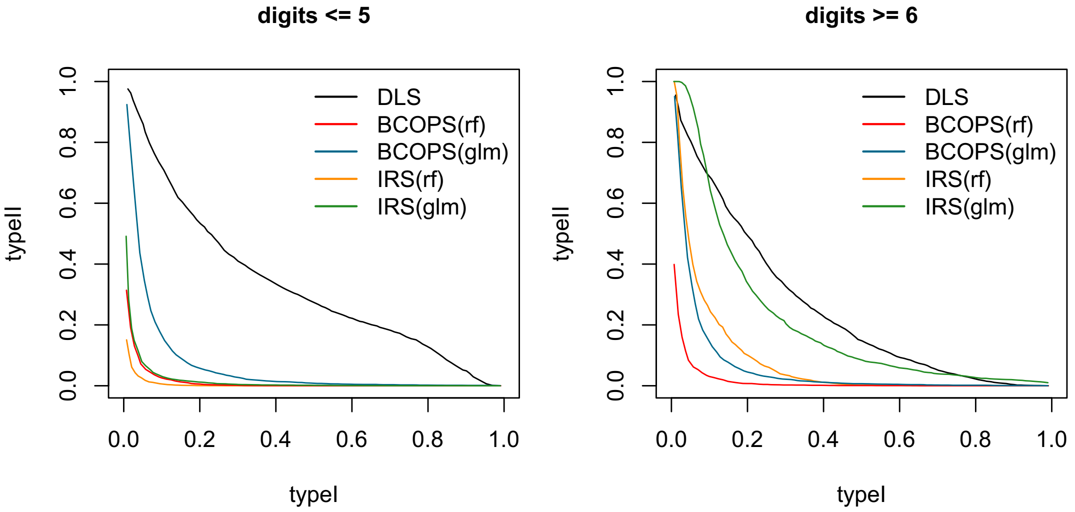

FIGURE 6 Comparisons of the Type II ~ Type I error curves using different conformal prediction sets. Results for observed digits (digits ≤ 5) and unobserved digits (digits ≥ 6) have been presented separately. BCOPS performs the best for the unobserved digits, and IRS is slightly better than BCOPS for the observed digits using the same learning algorithm. Both BCOPS and IRS are much better than DLS in this example [Colour figure can be viewed at wileyonlinelibrary.com]

much luck for the outliers. We can use methods proposed in Section 3 to estimate the FLR and the outlier abstention rate γ. Figure 7 compares the estimated FLR and γ with the actual sample versions of FLR and γ. We observe that the estimated FLR and γ match reasonably well with the actual FLP and γ (sample version) for both learning algorithm $\mathcal{L}$.

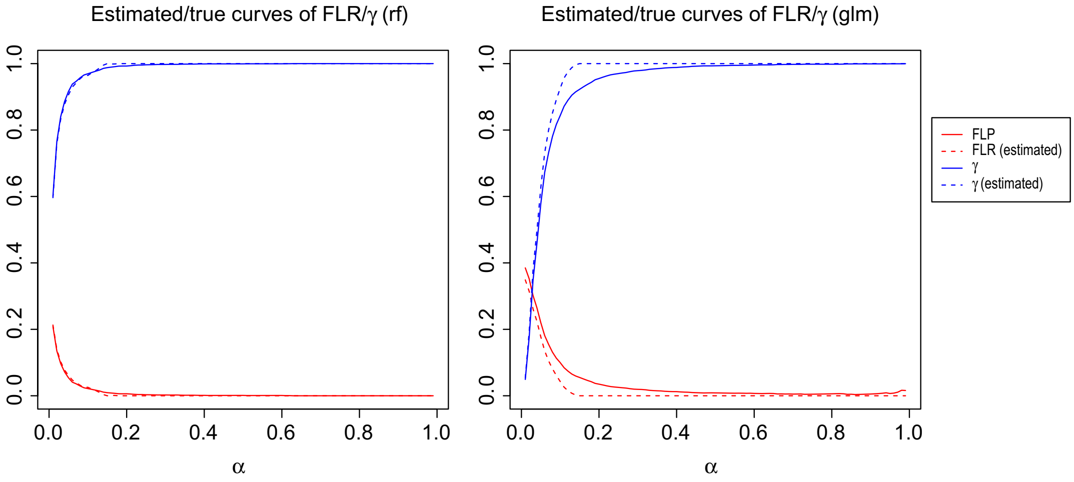

FIGURE 7 FLR and outlier abstention rate γ estimation. Red curves are actual FLP and estimated FLR, and the blue curves are the actual abstention rate γ realized on the current data set and the estimated γ [Colour figure can be viewed at wileyonlinelibrary.com]

5.2 | Network intrusion detection

The intrusion detector learning task is to build a predictive model capable of distinguishing between 'bad' connections, called intrusions or attacks, and 'good' normal connections (Stolfo et al., 2000). We use 10% of the data used in the 1999 KDD intrusion detection contest. We have four classes: normal, smurf, neptune and other intrusions. The normal, smurf, neptune samples are randomly assigned into the training and test samples while other intrusions appear only in the test data. We have 116 features in total and approximately 180,000 training samples and 180,000 test samples, and about 3.5% of the test data are other intrusions.

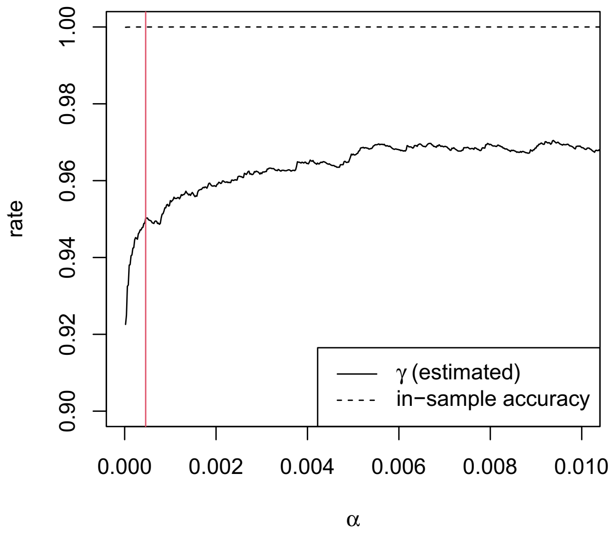

FIGURE 8 Estimated outlier abstention rate γ and in-sample accuracy against α. The vertical red line shows the suggested value for α if we let the estimated abstention rate be 95%. In this example, the in-sample accuracy remains almost 1 for extremely small α and is not very instructive for picking α [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Network intrusion results. The column names are the prediction sets from the BCOPS, and the row names are the true class labels. In each cell, the upper half is the number of samples falling into the category and the lower half is the prediction accuracy from RF for sample in this category. For example, the cell in the column ``normal+intrusion" and row "normal" describes the number of normal connections with BCOPS prediction set contains both normal and at least one intrusion label (upper half) and the prediction accuracy based on RF for these samples (lower half)

BCOPS Label Class	normal	intrusion	normal + intrusion	abstention
normal	48635 1.000	0 NA	0 NA	29 0.931
observed intrusions	0 NA	193858 1.000	0 NA	113 0.991
other intrusions	171 0.164	1 1.000	0 NA	8581 0.471

We let $\mathcal{L}$ be random forest, and compare the BCOPS with the RF prediction. Figure 8 shows the estimated abstention rate and estimated in-sample accuracy defined as (1-estimated type II errors for observed classes). In this example, BCOPS takes $\alpha = 0.05\%$, the largest α achieving 95% of the abstention rate for outliers.

Table 2 shows prediction accuracy using BCOPS and RF. We pool smurf and neptune together and call them the observed intrusions. We say that a prediction is correct from RF if it correctly assigns normal label to normal data or assign intrusion label to intrusions. From Table 2, we can see that the original RF classifier assigns correct label to 99.999% of samples from the observed classes, but claims more 50% of other intrusions to be the normal. The significant deterioration on unobserved intrusion types is also observed in participants' prediction during the contest http://cseweb.ucsd.edu/~elkan/clresults.html (Results of the KDD'99 Classifier Learning Contest).

As a comparison, the BCOPS achieves almost the same predictive power as the vanilla RF for the observed classes while refrains from making predictions for most of the samples from the novel attack types: the coverage for the normal and intrusion samples using BCOPS are 99.940% and 99.942%, and we assign correct labels to 99.941% of all samples from the observed classes while refrain from making predictions for 97.3% of the unobserved intrusion types.

6 | DISCUSSION

We propose a conformal inference method with a balanced objective for outlier detection and class label prediction. The proposed method has achieved good performance at both outlier detection and the class label prediction. Moreover, we propose a method for evaluating the outlier detection rate. Simple as it is, it has achieved good performance in the simulations and the real data examples in this paper. Throughout our discussion, we consider the setting with sufficient test samples. When the number of test samples is small, a minimization over the test distribution can be a poor choice. One alternative option is to minimize the loss over a mixture of the training and test distributions and let $\mu(x) = \kappa f_{test}(x) + (1-\kappa)f(x)$. For example, we can $\kappa = \frac{N}{n+N}$ where n and N are the sizes of the training set and the test set. Here, we also discuss some potential future work:

1. One extension is to consider the case where just a single new observation is available. In this case, although we have little information about $f_{test}(x)$, we can still try to design objective functions that lead to good predictions for samples close to the training data set while accepting our ignorance at locations where the training samples are rare. For example, we can use a truncated version of $f(x)$ and let $\mu(x) = f(x)\mathbb{1}_{f(x)\geq c} + c\mathbb{1}_{f(x)<c}$ for some properly chosen value c.
2. With just a single new observation available, another interesting question is to ask whether we can use this one sample to learn the direction separating the training data and the new sample. To prevent over-training, especially in high dimension, we can incorporate this direction searching step into the conformal prediction framework. If the direction learning step is too complicated, it will introduce too much variance, if there is no structure learning step, the decision rule can suffer in high dimension even if the problem itself may have simple underlying structure. Hence, it is again important to find a balance.
3. Another useful extension is to make BCOPS robust to small changes in the conditional distribution $f(x|y)$. The problem is not identifiable without proper constraints. If we allow only for the transformation $x_{test} \leftarrow ax + b$ from the training data distribution to the test data distribution for a reasonable simple a and b (Zhang et al., 2013), we may develop a modified BCOPS that is robust to small perturbation to features.

ACKNOWLEDGEMENTS

The authors thank Larry Wasserman, who suggested to us that conformal prediction might be a good approach for outlier detection. Robert Tibshirani was supported by NIH grant 5R01 EB001988-16 and NSF grant 19 DMS1208164.

ORCID

Leying Guan http://orcid.org/0000-0003-0609-1073
Robert Tibshirani http://orcid.org/0000-0003-0553-5090

REFERENCES

Barber, R.F., Candes, E.J., Ramdas, A. & Tibshirani, R.J. (2019) Conformal prediction under covariate shift. *arXiv preprint arXiv:1904.06019*.

Bartlett, P.L. & Wegkamp, M.H. (2008) Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9, 1823–1840.

Cadre, B. (2006) Kernel estimation of density level sets. *Journal of multivariate analysis*, 97, 999–1023.

Cadre, B., Pelletier, B. & Pudlo, P. (2009) Clustering by estimation of density level sets at a fixed probability.

Chandola, V., Banerjee, A. & Kumar, V. (2009) Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41, 15.

Chatterjee, S.K. & Patra, N.K. (1980) Asymptotically minimal multivariate tolerance sets. *Calcutta Statistical Association Bulletin*, 29, 73–94.

Hartigan, J.A. (1975) Clustering algorithms.

Hechtlinger, Y., Póczos, B. & Wasserman, L. (2018) Cautious deep learning. *arXiv preprint arXiv:1805.09460*.

Herbei, R. & Wegkamp, M.H. (2006) Classification with reject option. *Canadian Journal of Statistics*, 34, 709–721.

Hodge, V. & Austin, J. (2004) A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22, 85–126.

LeCun, Y. & Cortes, C. (2010) MNIST handwritten digit database. Available from http://yann.lecun.com/exdb/mnist/.

Lei, J. (2014) Classification with confidence. *Biometrika*, 101, 755–769.

Lei, J. & Wasserman, L. (2014) Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 71–96.

Lei, J., Robins, J. & Wasserman, L. (2013) Distribution-free prediction sets. *Journal of the American Statistical Association*, 108, 278–287.

Li, J. & Liu, R.Y. (2008) Multivariate spacings based on data depth: I. construction of nonparametric multivariate tolerance regions. *The Annals of Statistics*, 36, 1299–1323.

Lipton, Z.C., Wang, Y.-X. & Smola, A. (2018) Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916*.

Neyman, J. & Pearson, E.S. (1933) Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337.

Rigollet, P. & Vert, R. (2009) Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15, 1154–1178.

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

Stolfo, S.J., Fan, W., Lee, W., Prodromidis, A. & Chan, P.K. (2000) Cost-based modeling for fraud and intrusion detection: Results from the jam project. In: *Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00*, vol. 2, 130–144. IEEE.

Van de Geer, S.A. (2008) High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36, 614–645.

Vovk, V., Gammerman, A. & Shafer, G. (2005) *Algorithmic learning in a random world*. Berlin: Springer Science & Business Media.

Vovk, V., Nouretdinov, I. & Gammerman, A. (2009) On-line predictive linear regression. *The Annals of Statistics*, 37, 1566–1590.

Wald, A. (1943) An extension of wilks' method for setting tolerance limits. *The Annals of Mathematical Statistics*, 14, 45–55.

Wilks, S.S. (1941) Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12, 91–96.

Zhang, K., Schölkopf, B., Muandet, K. & Wang, Z. (2013) Domain adaptation under target and conditional shift. In: *International Conference on Machine Learning*, 819–827.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Guan, L. & Tibshirani, R. (2022) Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 84:524–546. https://doi.org/10.1111/rssb.12443