Complex & Intelligent Systems (2023) 9:5761–5778
https://doi.org/10.1007/s40747-023-01044-0

ORIGINAL ARTICLE

Probabilistic prediction with locally weighted jackknife predictive system

Di Wang[1,2] · Ping Wang[1,2] · Pingping Wang[3] · Cong Wang[1,2] · Zhen He[4] · Wei Zhang[4]

Received: 13 September 2022 / Accepted: 9 March 2023 / Published online: 5 April 2023

Abstract

Probabilistic predictions for regression problems are more popular than point predictions and interval predictions, since they contain more information for test labels. Conformal predictive system is a recently proposed non-parametric method to do reliable probabilistic predictions, which is computationally inefficient due to its learning process. To build faster conformal predictive system and make full use of training data, this paper proposes the predictive system based on locally weighted jackknife prediction approach. The theoretical property of our proposed method is proved with some regularity assumptions in the asymptotic setting, which extends our earlier theoretical researches from interval predictions to probabilistic predictions. In the experimental section, our method is implemented based on our theoretical analysis and its comparison with other predictive systems is conducted using 20 public data sets. The continuous ranked probability scores of the predictive distributions and the performance of the derived prediction intervals are compared. The better performance of our proposed method is confirmed with Wilcoxon tests. The experimental results demonstrate that the predictive system we proposed is not only empirically valid, but also provides more information than the other comparison predictive systems.

Keywords Probabilistic prediction · Predictive system · Jackknife prediction · Asymptotic analysis · Conformal prediction

✉ Wei Zhang
zhangweitju@tju.edu.cn

Di Wang
wangdi2015@tju.edu.cn

Ping Wang
wangps@tju.edu.cn

Pingping Wang
wangpingping@sdutcm.edu.cn

Cong Wang
wangc@tju.edu.cn

Zhen He
zhhe@tju.edu.cn

1 School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, People's Republic of China

2 Joint Laboratory of Intelligent Identification and Nowcasting Service for Convective System, CMA Public Meteorological Service Center, Beijing 100081, People's Republic of China

3 Qingdao Academy of Chinese Medical Science, Shandong University of Traditional Chinese Medicine, Qingdao 266112, Shandong, People's Republic of China

4 College of Management and Economics, Tianjin University, Tianjin 300072, People's Republic of China

Introduction

Machine learning techniques been widely applied to many areas because of their expressive power with the help of gradient descent [13, 15, 19, 23, 24] and metaheuristics [1–3, 16, 17] for efficient parameter searching. Based on machine learning, one can build predictive models considering the uncertainty of outputs. This paper concentrates on the predictive system providing predictive distribution for the test label, which is desirable for all walks of life [11]. For regression problems, the predictive distribution contains the full information of the uncertainty, as it can provide the probability of any event relevant to the test label and be transformed to prediction point or prediction interval by use of the corresponding first-order moment or quantiles. For many applications, especially the high-risk ones, the predictive distributions are required to be valid, which implies that the distributions or their derived prediction intervals have statistical compatibility with realizations, i.e., they ought to tell the truth [33].

Nowadays, many algorithms in the context of statistics or machine learning have been proposed to output predictive distributions for test labels. However, most of them

such as Bayesian regression and Gaussian process regression are highly dependent on their prior distribution assumptions, which can be far away from being valid if the prior assumptions are not correct [6, 18]. Recently, some frequentist approaches of probabilistic prediction algorithms have been proposed with compatibility with realizations in mind [25, 27]. While these approaches concern more about frequentist probability, they are limited in applications due to their original parametric forms. This issue has been tackled by a collection of promising works about conformal predictive systems (CPSs) [31, 33, 34], which build predictive system using learning framework of conformal prediction [6, 32] and extend the above frequentist approaches to a general nonparametric setting of being valid even in the small-sample cases.

The purpose of conformal prediction is to output valid prediction sets for test labels. One of the key characteristics of conformal prediction is that its p values calculated using conformity scores follows the uniform distribution on [0, 1] with the assumption of the samples being independent and identically distributed. This excellent property enables us to transform the unknown uncertainty from data to one of our most familiar distributions. CPSs utilize the p values of conformal prediction and transform them to the predictive distributions, which makes CPSs have the small-sample property of validity [31].

The pioneer work [31] first proposed CPSs with the classical least square procedure as the underlying algorithms and the asymptotic efficiency was proved with some strong assumptions. After that, [31] was done to answer some general questions about the existence and construction of consistent CPSs. In addition to the general theoretical studies above, two kinds of works concentrating on the applicability of CPSs were done. The first kind is to propose more flexible CPSs, whose representatives are [33] and [35]. The former extends using the classical least square procedure as underlying algorithm to using a more powerful algorithm named kernel ridge regression, and the latter proposed conformal calibration whose underlying algorithms are existing predictive systems. The second kind is to speed up the learning process of CPSs, as CPSs inherit the computational issue from conformal prediction [14, 20, 30]. To address this, there are two ways to try. One way is to modify the learning process of the original CPSs, such as split conformal predictive systems (SCPSs) and cross-conformal predictive systems (CCPSs) [34]. SCPSs are also valid even in small-sample cases, but they may lose predictive efficiency, as they split the data into two parts, one of which is used to train the underlying algorithm and the other of which is used to calculate conformity scores. Although CCPSs do not have the theoretical guarantee of validity, they improve the prediction performance by making full use of the data. Another way is to use a fast and well-performed underlying algorithm to compute the conformity scores, which was our previous work for building a fast probabilistic prediction algorithm [37]. In that work, based on the Leave-One-Out CCPS and extreme learning machine [12], we proposed a fast CPS named LOO–CCPS–RELM and analysed its asymptotic property of validity. LOO–CCPS–RELM takes advantage of jackknife prediction of residuals and their closed-form formula to make the whole learning process fast, which is competent in real-time applications.

This work extends our previous work about LOO–CCPS–RELM in two aspects. First, we design a more general learning framework in the spirit of LOO–CCPS–RELM to make probabilistic prediction, whose underlying algorithms can be any uniformly stable algorithm. Second, contrast with LOO–CCPS–RELM designed and proved to be asymptotically valid only for homoscedastic cases, the learning framework in this paper considers the heteroscedastic cases and a more general theoretical guarantee of the asymptotical validity is proved. The heteroscedastic cases are addressed by the idea of locally weighted jackknife prediction, whose theoretical analysis for prediction intervals has been conducted in our earlier work [38]. This paper extends the related concepts and analytical techniques to the probabilistic prediction. Since the predictive system we proposed is based on the idea of locally weighted jackknife prediction, it is named as locally weighted jackknife predictive system (LW-JPS) in this paper.

In summary, to build valid and computationally efficient predictive system, we develop locally weighted jackknife prediction approach with asymptotic guarantee of validity with the contributions as follows:

- A general predictive system based on the idea of locally weighted jackknife prediction is proposed for probabilistic prediction, which is easy-to-code and can learn fast if the underlying algorithms have the closed-form formula for leave-one-out residuals.
- The asymptotical validity of our predictive system is proved with some regularity assumptions, which extends the analysis of LOO–CCPS–RELM by considering a more general setting and the heteroscedastic cases.
- The experiments with 20 public data sets are conducted, which empirically proves the effective and efficiency of the proposed predictive system.

The rest of this paper is organized as follows. "Conformal predictive systems and locally weighted jackknife predictive system" reviews conformal predictive systems and defines the proposed LW-JPS. "Asymptotic analysis of locally weighted jackknife predictive system" proves the asymptotic validity of LW-JPS with some regularity assumptions and conditions. In "Experiments", the experiments

are designed to test the validity and efficiency of LW-JPS empirically and the conclusions of this paper are drawn in "Conclusion".

Conformal predictive systems and locally weighted jackknife predictive system

Throughout this paper, $\boldsymbol{X} \subseteq \boldsymbol{R}^n$ denotes the object space and $\boldsymbol{Y} \subseteq \boldsymbol{R}$ the label space. The observation space is denoted by $\boldsymbol{Z} = \boldsymbol{X} \times \boldsymbol{Y}$ and each observation $z = (\boldsymbol{x}, y) \in \boldsymbol{X} \times \boldsymbol{Y}$ comprises its object $\boldsymbol{x}$ and corresponding label y. $\boldsymbol{Z}l = \{\boldsymbol{Z}_i, i = 1, \cdots, l\}$ denotes a random training set whose realization is $z^l = \{z_i, i = 1, \cdots, l\}$. $\boldsymbol{Z}_0$ denotes a random test observation whose realization is z_0, where $\boldsymbol{Z}_0 = (\boldsymbol{X}_0, Y_0)$, $\boldsymbol{Z}_1 = (\boldsymbol{X}_1, Y_1), \cdots, \boldsymbol{Z}_l = (\boldsymbol{X}_l, Y_l)$ are independent and identically distributed and drawn from the distribution ρ on $\boldsymbol{Z} = \boldsymbol{X} \times \boldsymbol{Y}$. T denotes a random number uniformly distributed on [0, 1], which is independent of all observations and its realization is denoted by t.

For a fixed training set zl and a test input object $\boldsymbol{x}_0$, the goal of predictive systems is to construct a predictive distribution on $y \in \boldsymbol{R}$, which contains much of the information about y_0.

Predictive system and randomized predictive system

We first give the definition of predictive system which is first formally defined in [35].

Definition 1 *A measurable function* $Q : \boldsymbol{Z}^{l+1} \to [0, 1]$ *is a predictive system (PS) if it satisfies the following two conditions:*

A. For each realization z^l *and* $\boldsymbol{x}_0$*, the function* $Q\big(z^l, (\boldsymbol{x}_0, y)\big)$ *is increasing in* $y \in \boldsymbol{R}$.

B. For each realization z^l *and* $\boldsymbol{x}_0$,

$$\lim_{y \to -\infty} Q\Big(z^l, (\boldsymbol{x}_0, y)\Big) = 0$$

and

$$\lim_{y \to \infty} Q\Big(z^l, (\boldsymbol{x}_0, y)\Big) = 1.$$

Next, the notion of randomized predictive system is needed to introduce conformal predictive system.

Definition 2 *A measurable function* $Q : \boldsymbol{Z}^{l+1} \times [0, 1] \to [0, 1]$ *is a randomized predictive system (RPS) if it satisfies the following two conditions:*

A. For each realization z^l *and* $\boldsymbol{x}_0$*, the function* $Q\big(z^l, (\boldsymbol{x}_0, y), t\big)$ *is increasing in* $y \in \boldsymbol{R}$ *and* $t \in [0, 1]$.

B. For each realization z^l *and* $\boldsymbol{x}_0$,

$$\lim_{y \to -\infty} Q\Big(z^l, (\boldsymbol{x}_0, y), 0\Big) = 0$$

and

$$\lim_{y \to \infty} Q\Big(z^l, (\boldsymbol{x}_0, y), 1\Big) = 1.$$

In this paper, we use the shorthand notation $Q_{z^l, \boldsymbol{x}_0}(y) = Q\big(z^l, (\boldsymbol{x}_0, y)\big)$ to explicitly regard it as a function of y dependent on z^l and $\boldsymbol{x}_0$, and the shorthand notation $Q_{z^l, \boldsymbol{x}_0, t}(y) = Q\big(z^l, (\boldsymbol{x}_0, y), t\big)$ to explicitly regard it as a function of y dependent on z^l, $\boldsymbol{x}_0$ and t. $Q_{z^l, \boldsymbol{x}_0}(y)$ is the predictive distribution of PS, which is a cumulative distribution function (CDF) of Y_0 given z^l and $\boldsymbol{x}_0$. Different from that, RPS introduces a random number t to build the predictive distribution $Q_{z^l, \boldsymbol{x}_0, t}(y)$. For fixed training set z^l and $\boldsymbol{x}_0$, the lower bound and upper bound of $Q_{z^l, \boldsymbol{x}_0, t}(y)$ are $Q_{z^l, \boldsymbol{x}_0, 0}(y)$ and $Q_{z^l, \boldsymbol{x}_0, 1}(y)$, respectively. The gap $Q_{z^l, \boldsymbol{x}_0, 1}(y) - Q_{z^l, \boldsymbol{x}_0, 0}(y)$ can converge to 0 quickly for the existing designed RPSs [31]. Thus, one can use a CDF between or approximating $Q_{z^l, \boldsymbol{x}_0, 0}(y)$ and $Q_{z^l, \boldsymbol{x}_0, 1}(y)$ to remove the impact of t and build the predictive distribution of Y_0.

A predictive system $Q_{z^l, \boldsymbol{x}_0}(y)$ is valid, if the following holds:

$$P\Big\{Q_{\boldsymbol{Z}^l, \boldsymbol{X}_0}(Y_0) \le \eta\Big\} = \eta, \tag{1}$$

where $Q_{\boldsymbol{Z}^l, \boldsymbol{X}_0}(y)$ is a random function of y whose realization is $Q_{z^l, \boldsymbol{x}_0}(y)$. In addition, $Q_{z^l, \boldsymbol{x}_0}(y)$ is asymptotically valid if formula (1) holds asymptotically. Let $\widehat{q}^{(\eta/2)}_{z^l, \boldsymbol{x}_0}$ and $\widehat{q}^{(1-\eta/2)}_{z^l, \boldsymbol{x}_0}$ be the $\eta/2$ and $1 - \eta/2$ quantiles of $Q_{z^l, \boldsymbol{x}_0}(y)$. Then, the property of validity defined by formula (1) ensures that

$$P\Big\{Y_0 \in C^{(1-\eta)}_{\boldsymbol{Z}^l, \boldsymbol{X}_0}\Big\} = 1 - \eta, \tag{2}$$

where $C^{(1-\eta)}_{\boldsymbol{Z}^l, \boldsymbol{X}_0} = \Big[\widehat{q}^{(\eta/2)}_{z^l, \boldsymbol{x}_0}, \widehat{q}^{(1-\eta/2)}_{z^l, \boldsymbol{x}_0}\Big]$ is the prediction interval derived from $Q_{z^l, \boldsymbol{x}_0}(y)$, whose expected coverage rate is $1 - \eta$.

The predictive systems developed in the literature needs strong assumptions to be valid in small-sample cases [25, 27]. Therefore, to obtain validity in small-sample cases, randomized predictive system introduces the extra random number t, whose purpose is to define a similar property of validity for RPS as follows,

$$P\Big\{Q_{\boldsymbol{Z}^l, \boldsymbol{X}_0, T}(Y_0) \le \eta\Big\} = \eta \tag{3}$$

If $Q_{z^l, \boldsymbol{x}_0, t}(y)$ is the p value of conformal prediction, the corresponding RPS is called conformal predictive system, which has the property of validity in small-sample cases defined by formula (3) and the equation-like formula (2) holds by introducing T.

Next, we review SCPSs and CCPSs to demonstrate how to construct the function $Q_{z^l, \boldsymbol{x}_0, t}(y)$.

Split conformal predictive system

To build a CPS, the conformity scores of observations calculated by a conformity measure $A(S, z)$ are needed, where S is a data set and z is an observation. The conformity measure evaluates the degree of agreement between S and z. In the context of SCPSs, $A(S, z)$ should be a balance isotonic function [34]. In general, with a regression algorithm u, $A(S, z)$ can be designed as

$$A(S, z) = y - \widehat{\mu}_S(\boldsymbol{x}), \tag{4}$$

or

$$A(S, z) = \frac{y - \widehat{\mu}_S(\boldsymbol{x})}{\sqrt{\widehat{v}_S(\boldsymbol{x})}}, \tag{5}$$

where $\widehat{\mu}_S$ and $\widehat{v}_S$ are estimated mean function and conditional variance function learned from S, respectively.

The learning process of SCPSs splits the training set $\boldsymbol{z}^l$ into two parts, which are the proper training set $\boldsymbol{z}_1^m = \{(\boldsymbol{x}_j, y_j), j = 1, 2, \cdots, m\}$ and the calibration set $\boldsymbol{z}_m^l = \{(\boldsymbol{x}_j, y_j), j = m+1, \cdots, l\}$. For each possible label $y \in \boldsymbol{R}$, $l - m + 1$ conformity scores can be computed as follows:

$$\alpha_i = A\left(\boldsymbol{z}_1^m, (\boldsymbol{x}_i, y_i)\right),$$

$$\alpha_0^y = A\left(\boldsymbol{z}_1^m, (\boldsymbol{x}_0, y)\right),$$

where $\boldsymbol{x}_0$ is a test input, y is the corresponding label of $\boldsymbol{x}_0$ and $i = m+1, m+2, \cdots, l$. Based on the above calculation, $Q_{\boldsymbol{z}^l, \boldsymbol{x}_0, t}(y)$ can be obtained as formula (5) in [37]. The theory in [34] shows that the above $Q_{\boldsymbol{z}^l, \boldsymbol{x}_0, t}(y)$ is a valid RPS.

Different $A(S, z)$ leads to different $Q_{\boldsymbol{z}^l, \boldsymbol{x}_0, t}(y)$. Suppose that formula (5) is the conformity measure and define C_i as

$$C_i = \widehat{\mu}_{\boldsymbol{z}_1^m}(\boldsymbol{x}_0) + \frac{y_{m+i} - \widehat{\mu}_{\boldsymbol{z}_1^m}(\boldsymbol{x}_{m+i})}{\sqrt{\widehat{v}_{\boldsymbol{z}_1^m}(\boldsymbol{x}_{m+i})}} \times \sqrt{\widehat{v}_{\boldsymbol{z}_1^m}(\boldsymbol{x}_0)}.$$

Sort C_i to obtain $C_{(1)} \leq \cdots \leq C_{(l-m)}$ and let $C_{(0)} = -\infty$ and $C_{(l-m+1)} = \infty$. Then, the corresponding $Q_{\boldsymbol{z}^l, \boldsymbol{x}_0, t}(y)$ is calculated as formula (7) in [37], which can be further modified to become a formal CDF as formula (8) in [37], i.e., the empirical CDF of $\{C_{(i)}, i = 1, \cdots, l - m\}$.

The split process of SCPSs may not make full use of the training data, which is the reason of the development of CCPSs.

Cross-conformal predictive system

Based on the idea of cross validation, CCPSs first partition the training data into k folds. Let o_i denote the ordinals of training data in the i th fold and $z_{(o_i)}^l$ denote the training data without the i th fold. For each $i \in \{1, \cdots, k\}$, a CCPS with conformity measure $A(S, z)$ calculates the conformity scores with $z_{(o_i)}^l$ being the proper training set and $\{z_j | j \in o_i\}$ the calibration set. The corresponding conformity scores are

$$\alpha_{j,i} = A\left(z_{(o_i)}^l, z_j\right)$$

and

$$\alpha_{0,i}^y = A\left(z_{(o_i)}^l, (\boldsymbol{x}_0, y)\right).$$

Finally, the function $Q_{\boldsymbol{z}^l, \boldsymbol{x}_0, t}(y)$ of the CCPS is written as formula (9) in [37].

Suppose that formula (5) is the conformity measure and for $j \in o_i$, $C_{j,i}$ is written as

$$C_{j,i} = \widehat{\mu}_{\boldsymbol{z}_{(o_i)}^l}(\boldsymbol{x}_0) + \frac{y_i - \widehat{\mu}_{\boldsymbol{z}_{(o_i)}^l}(\boldsymbol{x}_i)}{\sqrt{\widehat{v}_{\boldsymbol{z}_{(o_i)}^l}(\boldsymbol{x}_i)}} \times \sqrt{\widehat{v}_{\boldsymbol{z}_{(o_i)}^l}(\boldsymbol{x}_0)}.$$

Sort all $C_{j,i}$ to obtain $C_{(1)} \leq \cdots \leq C_{(l)}$ and set $C_{(0)} = -\infty$ and $C_{(l+1)} = \infty$. Then, the $Q_{\boldsymbol{z}^l, \boldsymbol{x}_0, t}(y)$ of the above CCPS can be written as formula (10) in [37], which can be further modified to become a formal CDF as formula (11) in [37], i.e., the empirical CDF of $\{C_{(i)}, i = 1, \cdots, l\}$.

Leave-One-Out CCPS with formula (5) as conformity measure can be obtained by choosing $k = l$, whose predictive distribution is the empirical CDF of $\{C_i, i = 1, \cdots, l\}$, with C_i being written as.

$$C_i = \widehat{u}_{\boldsymbol{z}_{(i)}^l}(\boldsymbol{x}_0) + \frac{y_i - \widehat{u}_{\boldsymbol{z}_{(i)}^l}(\boldsymbol{x}_i)}{\sqrt{\widehat{v}_{\boldsymbol{z}_{(i)}^l}(\boldsymbol{x}_i)}} \times \sqrt{\widehat{v}_{\boldsymbol{z}_{(i)}^l}(\boldsymbol{x}_0)}.$$

We summarize the Leave-One-Out CCPS in Algorithm 1, since our proposed predictive system based on locally weighted jackknife prediction is highly related to it.

Algorithm 1 Leave-One-Out CCPS

Input:

Training set $\boldsymbol{z}^l$, test object $\boldsymbol{x}_0$, conformity score $A(S, \boldsymbol{z})$ as formula (5), regression algorithms μ and v.

Output:

Predictive distribution for Y_0.

1: For $i = 1,2,\cdots, l$, calculate C_i as follows,

$$C_i = \hat{u}_{\boldsymbol{z}_{(i)}^l}(\boldsymbol{x}_0) + \frac{y_i - \hat{u}_{\boldsymbol{z}_{(i)}^l}(\boldsymbol{x}_i)}{\sqrt{v_{\boldsymbol{z}_{(i)}^l}(\boldsymbol{x}_i)}} \times \sqrt{v_{\boldsymbol{z}_{(i)}^l}(\boldsymbol{x}_0)}.$$

2: **return** The empirical CDF of $\{C_i, i = 1, \cdots, l\}$.

Locally weighted jackknife predictive system

Jackknife prediction employs leave-one-out predictions for training data, which was proposed in the context of conformal prediction to build interval predictors [14, 36, 38]. Here, we extend it to build predictive systems inspired by Leave-One-Out CCPS. Locally weighted jackknife prediction is the jackknife prediction with the square root of $\widehat{v}_{\boldsymbol{z}^l_{(i)}}(\boldsymbol{x}_i)$ in Algorithm 1 as the local weight. In fact, Algorithm 1 can be modified to base on locally weighted jackknife prediction by changing $\widehat{u}_{\boldsymbol{z}^l_{(i)}}(\boldsymbol{x}_0)$ and $\widehat{v}_{\boldsymbol{z}^l_{(i)}}(\boldsymbol{x}_0)$ to $\widehat{u}_{\boldsymbol{z}^l}(\boldsymbol{x}_0)$ and $\widehat{v}_{\boldsymbol{z}^l}(\boldsymbol{x}_0)$, respectively, which reduces the times of the computation for regressors from l to 1. In addition, one also needs a way of calculating or approximating $\widehat{v}_{\boldsymbol{z}^l}$ and $\widehat{v}_{\boldsymbol{z}^l_{(i)}}$ efficiently to build the predictive system. In this paper, we employ the way of approximating them developed in our previous works [36, 38] about conformal prediction, which leads to our proposed predictive system based on locally weighted jackknife prediction in Algorithm 2.

Algorithm 2 Locally Weighted Jackknife Predictive System (LW-JPS)

Input:

Training data $\boldsymbol{z}^l$, test object $\boldsymbol{x}_0$, two regression algorithms u, v,

Output:

Predictive distribution for Y_0.

1: Fit $\boldsymbol{z}^l$ using u to obtain $\hat{u}_{\boldsymbol{z}^l}(\cdot)$.

2: Calculate $\hat{y}_i = \left(y_i - \hat{u}_{\boldsymbol{z}^l_{(i)}}(\boldsymbol{x}_i)\right)^2$, for $i = 1,2,\cdots,l$.

3: Fit the data set $\hat{\boldsymbol{z}}^l = \{(\boldsymbol{x}_i, \hat{y}_i), i = 1,2,\cdots,l\}$ using v to get $\hat{v}_{\boldsymbol{z}^l}(\cdot)$.

4: Calculate $\hat{v}_{\boldsymbol{z}^l_{(i)}}(\boldsymbol{x}_i)$, for $i = 1,2,\cdots,l$.

5: For $i = 1,2,\cdots,l$, calculate C_i as follows,

$$C_i = \hat{u}_{\boldsymbol{z}^l}(\boldsymbol{x}_0) + \frac{y_i - \hat{u}_{\boldsymbol{z}^l_{(i)}}(\boldsymbol{x}_i)}{\sqrt{\hat{v}_{\boldsymbol{z}^l_{(i)}}(\boldsymbol{x}_i)}} \times \sqrt{\hat{v}_{\boldsymbol{z}^l_{(i)}}(\boldsymbol{x}_0)}.$$

6: **return** The empirical CDF of $\{C_i, i = 1,\cdots,l\}$.

Algorithm 2 utilizes the jackknife prediction $\widehat{u}_{\boldsymbol{z}^l_{(i)}}$ and calculates the locally weighted leave-one-out residuals with the square root of $\widehat{v}_{\boldsymbol{z}^l_{(i)}}(\boldsymbol{x}_i)$ as the weight to build the predictive system.

Although Algorithm 2 needs to compute leave-one-out residuals, the learning process can be fast if the underlying algorithms u and v are linear smoothers [39], which have closed-form formula for computation.

Asymptotic analysis of locally weighted jackknife predictive system

This section provides the asymptotic analysis of LW-JPS. We first give the related definition, assumptions and conditions, and then prove the asymptotic validity of LW-JPS.

Definitions, assumptions and conditions

Throughout the paper, we assume that the labels are bounded by D, i.e., $sup_{y\in\boldsymbol{Y}}|y| \le D$. The regularity properties of the probability distribution ρ on $\boldsymbol{Z} = \boldsymbol{X} \times \boldsymbol{Y}$ will be assumed when it is needed as in [9]. All observations $(\boldsymbol{X}_i, Y_i)$ are i.i.d. samples. The generalization error of a function $f : \boldsymbol{X} \to \boldsymbol{Y}$ is measured by

$$\xi(f) = E\Big[(f(\boldsymbol{X}) - Y)^2\Big] = \int_{\boldsymbol{Z}} (f(\boldsymbol{x}) - y)^2 d\rho.$$

Denote the marginal probability distribution of ρ on $\boldsymbol{X}$ as $\rho_{\boldsymbol{X}}$, which is $\rho_{\boldsymbol{X}}(S) = \rho(S \times \boldsymbol{Y})$ for the measurable set $S \subseteq \boldsymbol{X}$. The conditional distribution of y given $\boldsymbol{x}$ is $\rho(y|\boldsymbol{x})$ and the regression function of ρ is

$$\mu_\rho(\boldsymbol{x}) = E[Y|\boldsymbol{X} = \boldsymbol{x}] = \int_{\boldsymbol{Y}} y d\rho(y|\boldsymbol{x}).$$

Therefore, based on Proposition 1.8 in [9], μ_ρ is the minimizer of $\xi(f)$ and for each $f : \boldsymbol{X} \to \boldsymbol{Y}$,

$$\xi(f) - \xi\big(\mu_\rho\big) = \int_{\boldsymbol{X}} \big(f(\boldsymbol{x}) - \mu_\rho(\boldsymbol{x})\big)^2 d\rho_{\boldsymbol{X}}.$$

It can be concluded that μ_ρ is bounded by D as $Y \le D$.

For the regression problem, we assume that the samples satisfy Assumption 1, where $\|f\|_\infty$ is the infinite norm of f on its domain, i.e., $\|f\|_\infty = \sup_{\boldsymbol{x}\in\boldsymbol{X}}|f(\boldsymbol{x})|$.

Assumption 1 *Each observation $(\boldsymbol{X}, Y)$ satisfies the following formula:*

$$Y = \mu_\rho(\boldsymbol{X}) + \sqrt{v_\rho(\boldsymbol{X})} \times \zeta,$$

where $v_\rho(\boldsymbol{X})$ is the conditional variance function and ζ is a random variable with zero mean and unit variance. ζ is independent of $\boldsymbol{X}$ and $0 < v_{min} \le \|v_\rho\|_\infty \le v_{max} < \infty$. In addition, $|\zeta| \le \zeta_{max}$ whose cumulative distribution function $F(b) = P\{\zeta \le b\}$ is continuous and strictly increasing on $\{b|F(b) \in (0, 1)\}$.

The formula in Assumption 1 is a standard assumption for regression problems with heteroscedastic setting, where the conditional variance of Y is dependent on $\boldsymbol{X}$ instead of a constant. Since Y is bounded, $|\zeta| \le \zeta_{max}$ is assumed.

An array of random variables X_l for $l \in N^+$ converges to a random variable X in probability is written as $X_l \to_p X$, whose definition can be found from the Definition 1 in [38].

To prove the asymptotic validity of LW-JPS, the following four conditions are needed for algorithms μ and v, which were also introduced in our earlier work about the theoretical analysis of locally weighted jackknife prediction [38]. In the conditions, r represents a general regression algorithm and $\boldsymbol{Z}^l$ is a general random data set for training r, whose samples are i.i.d.. $\widehat{r}_{\boldsymbol{Z}^l}$ is the learned regressor whose randomness is from $\boldsymbol{Z}^l$ and $\widehat{r}_{z^l}$ is the corresponding realization.

Condition 1. *The regression algorithm r is symmetric in the observations, such that for each l, each z^l and each permutation π of $\{1, \cdots, l\}$, there holds*

$$\widehat{r}_{z^l} = \widehat{r}_{\pi_l(z^l)},$$

where $\pi_l(z^l) = \{z_{\pi(j)},\ j = 1, \cdots, l\}$.

Condition 2. *The regressor $\widehat{r}_{\boldsymbol{Z}^l}$ uniformly converges in probability to the regression function μ_ρ of $\boldsymbol{Z}$, i.e.,*

$$\|\widehat{r}_{\boldsymbol{Z}^l} - \mu_\rho\|_\infty \to_p 0.$$

Condition 3. *The regression algorithm r is a uniformly stable algorithm [8], whose uniform stability with respect to the square loss is $\beta = \beta(l)$, i.e., for each l and each z^l,*

$$\sup_i \left(\sup_{(\mathbf{x}, y) \in \mathbf{X} \times Y} \left| \left(y - \hat{r}_{\boldsymbol{z}^l}(\mathbf{x})\right)^2 - \left(y - \hat{r}_{\boldsymbol{z}^l_{(i)}}(\mathbf{x})\right)^2 \right| \right) \le \beta(l),$$

where $\lim_{l \to \infty} \beta(l) = 0$.

Condition 4. *For two fixed data sets $\widehat{z}^l = \{(\boldsymbol{x}_j, \widehat{y}_j),\ j = 1, \cdots, l\}$ and $\widetilde{z}^l = \{(\boldsymbol{x}_j, \widetilde{y}_j),\ j = 1, \cdots, l\}$ with the same input objects, if for each l, the labels satisfy*

$$\sup_{i \in \{1, 2, \cdots, l\}} |\widehat{y}_i - \widetilde{y}_i| \le \eta,$$

there holds

$$\|\widehat{r}_{\widehat{z}^l} - \widehat{r}_{\widetilde{z}^l}\|_\infty \le \eta.$$

With the same mathematical skills in [38], we need the algorithm μ to satisfy Condition 1, 2, and 3 and v to satisfy Condition 1, 2 and 4 to prove the asymptotic validity of LW-JPS. The conditions are not too restrict for applications, which we have analyzed in section 3.3 of [38].

Asymptotic validity of LW-JPS

We introduce Lemma 1 to guarantee that $\widehat{v}_{\widehat{\boldsymbol{Z}}^l}$ is a consistent estimator for the conditional variance function in Algorithm 2, which has been proved in [38].

Lemma 1 *With Assumption 1 being hold, μ satisfying Condition 2 and 3, and v satisfying Condition 2 and 4, we have*

$$\|\widehat{v}_{\widehat{\boldsymbol{Z}}^l} - v_\rho\|_\infty \to_p 0.$$

We will prove in Theorem 1 that Algorithm 2 is asymptotically valid by showing that the corresponding predictive distribution $\widehat{Q}_{z^l, \boldsymbol{x}_0}(y)$ satisfies

$$P\left\{ \widehat{Q}_{\boldsymbol{Z}^l, \boldsymbol{X}_0}(Y_0) \le \alpha | \boldsymbol{Z}^l \right\} \to_p \alpha, \tag{6}$$

which is an asymptotic version of formula (1). To do so, we need to prove that

$$P\left\{ Y_0 \le \widehat{q}^{(\alpha)}_{\boldsymbol{Z}^l, \boldsymbol{X}_0} | \boldsymbol{Z}^l \right\} \to_p \alpha, \tag{7}$$

where $\widehat{q}^{(\alpha)}_{z^l, \boldsymbol{x}_0}$ is the α quantile of $\widehat{Q}_{z^l, \boldsymbol{x}_0}(y)$. Formula (7) is equivalent to

$$P\left\{ \Gamma_{\boldsymbol{Z}^l} \le \widehat{q}^{(\alpha)}_{\boldsymbol{Z}^l} | \boldsymbol{Z}^l \right\} \to_p \alpha, \tag{8}$$

where Γ_{z^l} is the normalized residual defined by

$$\Gamma_{z^l} = \frac{Y_0 - \widehat{\mu}_{z^l}(\boldsymbol{X}_0)}{\sqrt{\widehat{v}_{z^l}(\boldsymbol{X}_0)}},$$

and $\widehat{q}^{(\alpha)}_{z^l}$ is the α quantile of the normalized leave-one-out residuals $\left\{a_{l,i},\ i = 1, \cdots, l\right\}$ defined by

$$a_{l,i} = \frac{y_i - \widehat{\mu}_{\boldsymbol{z}^l_{(i)}}(\boldsymbol{x}_i)}{\sqrt{\widehat{v}_{\boldsymbol{z}^l_{(i)}}(\boldsymbol{x}_i)}}.$$

Denote the CDF of Γ_{z^l} by $F_{z^l}(b)$, i.e.,

$$F_{z^l}(b) = P\left\{ \Gamma_{z^l} \le b | z^l \right\},$$

and $q^{(\alpha)}$ is the α quantile of $F(b)$ in Assumption 1. Since Lemma 1 confirms that the estimator $\widehat{v}_{\widehat{z}^l}$ uniformly converges to v_ρ in probability and μ satisfying Condition 2, we can make the connection between Γ_{z^l} and the normalized noise term of Assumption 1, which is

$$\zeta_0 = \frac{Y_0 - \mu_\rho(\boldsymbol{X}_0)}{\sqrt{v_\rho(\boldsymbol{X}_0)}},$$

and prove in Lemma 2 that

$$\sup_{b\in R}\left|F_{\mathbf{Z}^l}(b)-F(b)\right|\to_p 0.$$

Also, $\widehat{q}^{(\alpha)}_{\mathbf{Z}^l}$ and $q^{(\alpha)}$ are highly related as we show in Lemma 2 that

$$\widehat{q}^{(\alpha)}_{\mathbf{Z}^l}\to_p q^{(\alpha)}.$$

Building on Lemma 2, formula (8) and formula (6) can be proved in turn and the conclusion of LW-JPS being asymptotically valid can be drawn in Theorem 1.

The analysis techniques in Lemma 2 was first introduced in [28] for linear regression problems with homoscedastic errors, which was further improved for nonlinear regression problems with heteroscedastic errors by our earlier work for locally weighted jackknife prediction [38]. The above two works both concerns building interval prediction other than probabilistic prediction, which makes the detailed expressions different from this work. In addition, our work about LOO–CCPS–RELM [37] only considers nonlinear regression problems with homoscedastic errors, and its proofs are specific to extreme learning machine. Therefore, we introduce and prove Lemma 2, which is essential to proving Theorem 1 strictly in this paper.

Lemma 2 *Fix* $\alpha\in(0,1)$. *If the conditions of Lemma 1 hold and both* μ*and valso satisfy Condition 1, then we have.*

$$\sup_{b\in R}\left|F_{\mathbf{Z}^l}(b)-F(b)\right|\to_p 0, \tag{9}$$

and

$$\widehat{q}^{(\alpha)}_{\mathbf{Z}^l}\to_p q^{(\alpha)}. \tag{10}$$

Proof Since $\widehat{\mu}_{z^l}$ satisfies that

$$\|\widehat{\mu}_{\mathbf{Z}^l}-\mu_\rho\|_\infty\to_p 0,$$

and $\widehat{v}_{\widehat{z}^l}$ satisfies that

$$\|\widehat{v}_{\widehat{\mathbf{Z}}^l}-v_\rho\|_\infty\to_p 0,$$

for each l we can define a nonempty set $B(l)$ as

$$B(l)=\left\{z^l\left|\max\left\{\|\widehat{\mu}_{z^l}-\mu_\rho\|_\infty,\ \|\widehat{v}_{\widehat{z}^l}-v_\rho\|_\infty\right\}\le g(l)\right.\right\},$$

where $g(l)$ is nonnegative and converges to 0 sufficiently slow. Then, we can construct an array of random variables Γ_{z^l} by taking an arbitrary z^l in $B(l)$. As $z^l\in B(l)$, $v_{min}>0$ and $g(l)$ converges to 0, there exists a l_1, such that for all $l>l_1$, there holds

$$\|\widehat{v}_{\widehat{z}^l}\|_\infty\ge v_{min}-g(l)\ge v_{min}-g(l_1)>0.$$

For all $l>l_1$, by the definitions, we have

$$\left|\Gamma_{z^l}-\zeta_0\right|\le\frac{\frac{g(l)}{\left|\sqrt{v_{min}}+\sqrt{v_{min}-g(l_1)}\right|}\times\zeta_{max}+g(l)}{\sqrt{v_{min}-g(l_1)}},$$

which guarantees that

$$\Gamma_{z^l}\to_p\zeta_0.$$

Since convergence in probability implies convergence in distribution and the CDF of ζ_0 is continuous, according to Proposition 1.16 of [26], we have

$$\lim_{l\to\infty}\sup_{b\in\boldsymbol{R}}\left|F_{z^l}(b)-F(b)\right|=0.$$

The arbitrarily chosen z^l from $B(l)$ leads to

$$\lim_{l\to\infty}\sup_{\mathbf{z}^l\in B(l)}\sup_{b\in\boldsymbol{R}}\left|F_{z^l}(b)-F(b)\right|=0,$$

which implies that formula (9) is correct [38].

Next, we prove formula (10). Since for every $\epsilon>0$, we have

$$P\left\{\left|\widehat{q}^{(\alpha)}_{\mathbf{Z}^l}-q^{(\alpha)}\right|>\epsilon\right\}=P\left\{\widehat{q}^{(\alpha)}_{\mathbf{Z}^l}>q^{(\alpha)}+\epsilon\right\}+P\left\{\widehat{q}^{(\alpha)}_{\mathbf{Z}^l}<q^{(\alpha)}-\epsilon\right\}.$$

Thus, we need to show that

$$P\left\{\widehat{q}^{(\alpha)}_{\mathbf{Z}^l}>q^{(\alpha)}+\epsilon\right\}$$

and

$$P\left\{\widehat{q}^{(\alpha)}_{\mathbf{Z}^l}<q^{(\alpha)}-\epsilon\right\}$$

converges to 0, respectively. Define $F_l(b)$ by

$$\begin{aligned}F_l(b)&=P\left\{\frac{Y_1-\hat{\mu}_{\mathbf{Z}^l_{(1)}}(\boldsymbol{X}_1)}{\sqrt{\hat{v}_{\hat{\mathbf{Z}}^l_{(1)}}(\boldsymbol{X}_1)}}\le b\right\}\\&=E\left[P\left\{\frac{Y_1-\hat{\mu}_{\mathbf{Z}^l_{(1)}}(\boldsymbol{X}_1)}{\sqrt{\hat{v}_{\hat{\mathbf{Z}}^l_{(1)}}(\boldsymbol{X}_1)}}\le b|\mathbf{Z}^l_{(1)}\right\}\right]\\&=E\left[F_{\mathbf{Z}^l_{(1)}}(b)\right]\end{aligned}$$

whose distance from $F(b)$ can be bounded by

$$\sup_{b\in\boldsymbol{R}}|F_l(b)-F(b)| \le E\left[\sup_{b\in\boldsymbol{R}}\left|F_{\boldsymbol{Z}^l_{(1)}}(b)-F(b)\right|\right].$$

From formula (9) and the definition of leave-one-out samples, the bounded random variable $\sup_{b\in\boldsymbol{R}}\left|F_{\boldsymbol{Z}^l_{(1)}}(b)-F(b)\right|$ converges to 0 in probability. This leads to

$$\lim_{l\to\infty}\sup_{b\in\boldsymbol{R}}|F_l(b)-F(b)| \le \lim_{l\to\infty} E\left[\sup_{b\in\boldsymbol{R}}\left|F_{\boldsymbol{Z}^l_{(1)}}(b)-F(b)\right|\right], =0$$

i.e.,

$$\lim_{l\to\infty}\sup_{b\in\boldsymbol{R}}|F_l(b)-F(b)|=0. \tag{11}$$

Let the CDF of the normalized leave-one-out residuals $\{a_{l,i},\, i=1,\cdots,l\}$ be denoted by $F_{al}(b)$. Therefore, $F_{A_l}(b)$ and $\widehat{q}^{(\alpha)}_{\boldsymbol{Z}^l}$ are the corresponding random function and random variable by introducing the randomness of $\boldsymbol{Z}^l$. Define $J_{l,i}=1_{\{A_{l,i}>q^{(\alpha)}+\epsilon\}}$, which is the indicator function of $\{A_{l,i}>q^{(\alpha)}+\epsilon\}$. As algorithms μ and v being exchangeable implies that $\{J_{l,j},\, j=1,\cdots,l\}$ are exchangeable, based on the property of quantile function [29], we have

$$\begin{aligned}
P\left\{\hat{q}^{(\alpha)}_{\boldsymbol{Z}^l}>q^{(\alpha)}+\in\right\} &= P\left\{\alpha>F_{A_l}\left(q^{(\alpha)}+\in\right)\right\}\\
&= P\left\{1-F_{A_l}\left(q^{(\alpha)}+\in\right)>1-\alpha\right\}\\
&= P\left\{\frac{1}{l}\sum_{i=1}^{l}\left(J_{l,i}-E\left[J_{l,1}\right]\right)>1-\alpha-E\left[J_{l,1}\right]\right\}\\
&= P\left\{\frac{1}{l}\sum_{i=1}^{l}\left(J_{l,i}-E\left[J_{l,i}\right]\right)>F_l\left(q^{(\alpha)}+\in\right)-\alpha\right\}.
\end{aligned}$$

Since formula (11) holds, it follows that

$$F\left(q^{(\alpha)}+\epsilon\right)>\alpha,$$

which implies that $F_l\left(q^{(\alpha)}+\epsilon\right)>0$ for sufficiently large l. Thus, it follows from Markov's inequality that for sufficiently large l, the probability,

$$P\left\{\frac{1}{l}\sum_{i=1}^{l}\left(J_{l,i}-E\left[J_{l,i}\right]\right)>F_l\left(q^{(\alpha)}+\epsilon\right)-\alpha\right\},$$

is bounded by

$$\frac{\frac{1}{l}var\left(J_{l,1}\right)+\frac{l(l-1)}{l^2}cov\left(J_{l,1},\,J_{l,2}\right)}{\left(F_l\left(q^{(\alpha)}+\epsilon\right)-\alpha\right)^2}, \tag{12}$$

where var and cov are the variance and covariance function, respectively. Therefore, to prove $P\left\{\widehat{q}^{(\alpha)}_{\boldsymbol{Z}^l}>q^{(\alpha)}+\epsilon\right\}$ approaches 0, we need to prove $cov\left(J_{l,1},\,J_{l,2}\right)$ converges to 0 as $l\to\infty$.

Let $\boldsymbol{Z}^l_{(1,2)}$ and $\widehat{\boldsymbol{Z}}^l_{(1,2)}$ be the corresponding data set without the first two observations. Define $A_{l,(1,2)}$ and $A_{l,(2,1)}$ by

$$A_{l,(1,2)}=\frac{Y_1-\widehat{m}_{\boldsymbol{Z}^l_{(1,2)}}(\boldsymbol{X}_1)}{\sqrt{\widehat{v}_{\widehat{\boldsymbol{Z}}^l_{(1,2)}}(\boldsymbol{X}_1)}},\ A_{l,(2,1)}=\frac{Y_2-\widehat{m}_{\boldsymbol{Z}^l_{(1,2)}}(\boldsymbol{X}_2)}{\sqrt{\widehat{v}_{\widehat{\boldsymbol{Z}}^l_{(1,2)}}(\boldsymbol{X}_2)}},$$

and define $\widetilde{A}_1$ and $\widetilde{A}_2$ by

$$\widetilde{A}_1=\left[A_{l,1},\,A_{l,2}\right],\ \widetilde{A}_2=\left[A_{l,(1,2)},\,A_{l,(2,1)}\right].$$

Let $\widetilde{F}_{l,1}$ and $\widetilde{F}_{l,2}$ be the CDFs of $\widetilde{A}_1$ and $\widetilde{A}_2$, respectively, i.e.,

$$\widetilde{F}_{l,1}(b_1,\,b_2)=P\left\{A_{l,1}\le b_1,\,A_{l,1}\le b_2\right\},$$

and

$$\widetilde{F}_{l,2}(b_1,\,b_2)=P\left\{A_{l,(1,2)}\le b_1,\,A_{l,(1,2)}\le b_2\right\}.$$

For $\widetilde{F}_{l,2}$, we have

$$\widetilde{F}_{l,2}(b_1,\,b_2)=E\left[F_{\boldsymbol{Z}^l_{(1,2)}}(b_1)F_{\boldsymbol{Z}^l_{(1,2)}}(b_2)\right]. \tag{13}$$

Since $F_{\boldsymbol{Z}^l_{(1,2)}}(b_1)$ and $F_{\boldsymbol{Z}^l_{(1,2)}}(b_2)$ are bounded random variables, which converge to $F(b_1)$ and $F(b_2)$, respectively, due to formula (9), we have

$$\lim_{l\to\infty}\widetilde{F}_{l,2}(b_1,\,b_2)=F(b_1)F(b_2).$$

As Lemma 1 holds and μ satisfies Condition 2, we can deduce that $A_{l,1}$, $A_{l,2}$, $A_{l,(1,2)}$ and $A_{l,(2,1)}$ are all convergent in probability to ζ_0, which implies that $A_{l,1}-A_{l,(1,2)}\to_p 0$ and $A_{l,2}-A_{l,(2,1)}\to_p 0$. Therefore, from Lemma 2.8 in [29], there holds

$$\lim_{l\to\infty}\widetilde{F}_{l,1}(b_1,\,b_2)=F(b_1)F(b_2).$$

Furthermore, with

$$\begin{aligned}
\operatorname{cov}\left(J_{l,1},\,J_{l,1}\right) &= \operatorname{cov}\left(1-J_{l,1},\,1-J_{l,1}\right)\\
&= \tilde{F}_{l,1}\left(q^{(\alpha)}+\in,\,q^{(\alpha)}+\in\right)-F_l\left(q^{(\alpha)}+\in\right)F_l\left(q^{(\alpha)}+\in\right)
\end{aligned}$$

and formula (11), we have

$$\lim_{l\to\infty}cov\left(J_{l,1},\,J_{l,1}\right)=0.$$

Based on the formula above and Eq. (12), we have

$$\lim_{l\to\infty} P\left\{\widehat{q}_{\mathbf{Z}^l}^{(\alpha)} > q^{(\alpha)} + \epsilon\right\} = 0.$$

Similarly, we can also prove that

$$\lim_{l\to\infty} P\left\{\widehat{q}_{\mathbf{Z}^l}^{(\alpha)} < q^{(\alpha)} - \epsilon\right\} = 0.$$

Thus, since the fact that

$$P\left\{\left|\widehat{q}_{\mathbf{Z}^l}^{(\alpha)} - q^{(\alpha)}\right| > \epsilon\right\} = P\left\{\widehat{q}_{\mathbf{Z}^l}^{(\alpha)} > q^{(\alpha)} + \epsilon\right\} + P\left\{\widehat{q}_{\mathbf{Z}^l}^{(\alpha)} < q^{(\alpha)} - \epsilon\right\}.$$

and the two limit equations above hold, we have

$$\widehat{q}_{\mathbf{Z}^l}^{(1-\alpha)} \to_p q^{(1-\alpha)}. \tag{14}$$

The following two theorems describe the statistical compatibility of the predictive distributions output by LW-JPS with observations in the asymptotic setting. Theorem 1 proves the asymptotic version of formula (1) and Theorem 2 proves a sufficient condition of the asymptotic version of formula (2), where quantiles can be set arbitrarily.

Theorem 1 *Fix $\alpha \in (0, 1)$. If Assumption 1holds, μ satisfying Condition 1, 2 and 3, and vsatisfying Condition 1, 2 and 4, we have*

$$P\left\{\widehat{Q}_{\mathbf{Z}^l, \boldsymbol{X}_0}(Y_0) \le \alpha | \mathbf{Z}^l\right\} \to_p \alpha.$$

Proof Based on Assumption 1, we have $F\left(q^{(\alpha)}\right) = \alpha$. Therefore,

$$\begin{aligned}\left|P\left\{\Gamma_{\mathbf{Z}^l} \le \hat{q}_{\mathbf{Z}^l}^{(\alpha)} | \mathbf{Z}^l\right\} - \alpha\right| &= \left|F_{\mathbf{Z}^l}\left(\hat{q}_{\mathbf{Z}^l}^{(\alpha)}\right) - F\left(q^{(\alpha)}\right)\right| \\ &\le \left|F_{\mathbf{Z}^l}\left(\hat{q}_{\mathbf{Z}^l}^{(\alpha)}\right) - F\left(\hat{q}_{\mathbf{Z}^l}^{(\alpha)}\right)\right| + \left|F\left(\hat{q}_{\mathbf{Z}^l}^{(\alpha)}\right) - F\left(q^{(\alpha)}\right)\right| \\ &\le \sup_{b\in\boldsymbol{R}} \left|F_{\mathbf{Z}^l}(b) - F(b)\right| + \left|F\left(\hat{q}_{\mathbf{Z}^l}^{(\alpha)}\right) - F\left(q^{(\alpha)}\right)\right|.\end{aligned}$$

From Lemma 2 and $F(b)$ being continuous, we have $\left|F\left(\widehat{q}_{\mathbf{Z}^l}^{(\alpha)}\right) - F\left(q^{(\alpha)}\right)\right| \to_p 0$ using Theorem 1.10 in [26] and $sup \to_p 0$. Thus, we can conclude that
$b\in\boldsymbol{R}$

$$F_{\mathbf{Z}^l}\left(\widehat{q}_{\mathbf{Z}^l}^{(\alpha)}\right) = P\left\{\Gamma_{\mathbf{Z}^l} \le \widehat{q}_{\mathbf{Z}^l}^{(\alpha)} | \mathbf{Z}^l\right\} \to_p \alpha, \tag{15}$$

which is equivalent to

$$P\left\{Y_0 \le \widehat{q}_{\mathbf{Z}^l, \boldsymbol{X}_0}^{(\alpha)} | \mathbf{Z}^l\right\} \to_p \alpha, \tag{16}$$

since for every $\alpha \in (0, 1)$, there holds

$$F_{\mathbf{Z}^l}\left(\widehat{q}_{\mathbf{Z}^l}^{(\alpha)}\right) = P\left\{\Gamma_{\mathbf{Z}^l} \le \widehat{q}_{\mathbf{Z}^l}^{(\alpha)} | \mathbf{Z}^l\right\} = P\left\{Y_0 \le \widehat{q}_{\mathbf{Z}^l, \boldsymbol{X}_0}^{(\alpha)} | \mathbf{Z}^l\right\}.$$

For every ϵ such that $0 < \epsilon < max\{\alpha, 1-\alpha\}$, by the definition of quantiles, we have

$$P\left\{\widehat{Q}_{\mathbf{Z}^l, \boldsymbol{X}_0}(Y_0) \le \alpha | \mathbf{Z}^l\right\} \le F_{\mathbf{Z}^l}\left(\widehat{q}_{\mathbf{Z}^l}^{(\alpha+\epsilon)}\right), \tag{17}$$

and

$$P\left\{\widehat{Q}_{\mathbf{Z}^l, \boldsymbol{X}_0}(Y_0) \le \alpha | \mathbf{Z}^l\right\} \ge F_{\mathbf{Z}^l}\left(\widehat{q}_{\mathbf{Z}^l}^{(\alpha-\epsilon)}\right). \tag{18}$$

Based on formula (15), for every $\delta > 0$, we have

$$P\left\{\left|F_{\mathbf{Z}^l}\left(\widehat{q}_{\mathbf{Z}^l}^{(\alpha+\epsilon)}\right) - (\alpha+\epsilon)\right| > \epsilon\right\} < \delta,$$

which combing formula (17) lead to

$$P\left\{P\left\{\widehat{Q}_{\mathbf{Z}^l, \boldsymbol{X}_0}(Y_0) \le \alpha | \mathbf{Z}^l\right\} - (\alpha+\epsilon) > \epsilon\right\} < \delta.$$

Similarly, with formula (18), there holds

$$P\left\{P\left\{\widehat{Q}_{\mathbf{Z}^l, \boldsymbol{X}_0}(Y_0) \le \alpha | \mathbf{Z}^l\right\} - (\alpha-\epsilon) < -\epsilon\right\} < \delta.$$

Then, we have

$$P\left\{\left|P\left\{\widehat{Q}_{\mathbf{Z}^l, \boldsymbol{X}_0}(Y_0) \le \alpha | \mathbf{Z}^l\right\} - \alpha\right| > 2\epsilon\right\} < 2\delta.$$

Since ϵ and δ are arbitrary, the conclusion of Theorem 1 can be drawn.

Based on the deduction of Theorem 1, we can obtain the following coverage guarantee for derived prediction intervals from $\widehat{Q}_{\boldsymbol{z}^l, \boldsymbol{x}_0}$, which is desirable for practitioners for interval prediction.

Theorem 2 *Fix η_1 and $\eta_2$such that $0 < \eta_1 < \eta_2 < 1$. If the conditions of Theorem 1hold, we have*

$$P\left\{q_{\mathbf{Z}^l, \boldsymbol{X}_0}^{(\eta_1)} \le Y_0 \le q_{\mathbf{Z}^l, \boldsymbol{X}_0}^{(\eta_2)} | \mathbf{Z}^l\right\} \to_p \eta_2 - \eta_1.$$

Proof For every ϵ such that $0 < \epsilon < \max\{\eta_1, \Delta_\eta\}$, based on formula (16), we have

$$P\left\{Y_0 \le \widehat{q}_{\mathbf{Z}^l, \boldsymbol{X}_0}^{(\eta_1-\epsilon)} | \mathbf{Z}^l\right\} \to_p \eta_1 - \epsilon,$$

which leads to

$$P\left\{q_{\mathbf{Z}^l, \boldsymbol{X}_0}^{(\eta_1-\epsilon)} < Y_0 \le q_{\mathbf{Z}^l, \boldsymbol{X}_0}^{(\eta_2)} | \mathbf{Z}^l\right\} \to_p \Delta_\eta - \epsilon,$$

where $\Delta_\eta = \eta_2 - \eta_1$. Then, for every $\delta > 0$, there holds

$$P\left\{\left|P\left\{q^{(\eta_1-\epsilon)}_{\mathbf{Z}^l,\mathbf{X}_0} < Y_0 \le q^{(\eta_2)}_{\mathbf{Z}^l,\mathbf{X}_0}|\mathbf{Z}^l\right\} - \left(\Delta_\eta + \epsilon\right)\right| > \epsilon\right\} < \delta.$$

Thus, we have

$$P\left\{P\left\{q^{(\eta_1)}_{\mathbf{Z}^l,\mathbf{X}_0} \le Y_0 \le q^{(\eta_2)}_{\mathbf{Z}^l,\mathbf{X}_0}|\mathbf{Z}^l\right\} - \left(\Delta_\eta + \epsilon\right) > \epsilon\right\} < \delta.$$

Similarly, there holds

$$P\left\{P\left\{q^{(\eta_1)}_{\mathbf{Z}^l,\mathbf{X}_0} \le Y_0 \le q^{(\eta_2)}_{\mathbf{Z}^l,\mathbf{X}_0}|\mathbf{Z}^l\right\} - \left(\Delta_\eta - \epsilon\right) < -\epsilon\right\} < \delta.$$

Therefore, we have

$$P\left\{\left|P\left\{q^{(\eta_1)}_{\mathbf{Z}^l,\mathbf{X}_0} \le Y_0 \le q^{(\eta_2)}_{\mathbf{Z}^l,\mathbf{X}_0}|\mathbf{Z}^l\right\} - \Delta_\eta\right| > 2\epsilon\right\} < 2\delta,$$

which proves the conclusion of Theorem 2, since ϵ and δ are arbitrary.

Experiments

In this section, to test LW-JPS empirically, randomized kernel ridge regression with random Fourier features [21] is used as μ and k-nearest neighbor regression is used as v, respectively, since they satisfy the conditions we assumed in "Asymptotic analysis of locally weighted jackknife predictive system". Following [38], the number of random features were set to 1000 and $k = \sqrt{l}$ for k-nearest neighbor regression. The ridge parameter with the least leave-one-out errors was chosen for LW-JPS. The comparison predictive systems are SPCS with support vector regression (SCPS–SVR), SCPS with random forests (SCPS–RF), CCPS with support vector regression (CCPS–SVR), CCPS with random forests (CCPS–RF) and CPS with random forests with out-of-bag errors as conformity scores (OOB–CPS–RF). All the comparison algorithms employ formula (5) as the conformity measure based on the recently empirical evaluation research in [40]. SCPS–SVR, SCPS–RF, CCPS–SVR and CCPS–RF use the same normalization for conformity measure as LW-JPS, whereas OOB–CPS–RF uses the standard deviation of out-of-bag predictions for normalization based on the approach in [40]. OOB–CPS–RF was first proposed in [40], which extends the idea of the state-of-the-art conformal regressor with random forests [7]. Following [37], for all SCPSs, 40 percent of the training data was used as calibration set and for all CCPSs, the number of folds was 5. In addition, the meta-parameters of all comparison algorithms were chosen using threefold cross-validation on the training set based on R^2 scores. SVR with Gaussian kernel was employed, whose regularization parameter C was chosen from $\left\{10^{-5}, 10^{-4}, \cdots, 10^4, 10^5\right\}$. For random forests, the number of trees was chosen from {100, 300, 500, 1000} and the minimum number of samples per tree leaf was chosen from {1, 3, 5}, respectively.

Table 1 Data sets

Short name	Examples	Dimensionality	Source
Abalone	4177	8	UCI
Bank8fh	8192	8	Delve
Bank8fm	8192	8	Delve
Bank8nh	8192	8	Delve
Bank8nm	8192	8	Delve
Boston	506	13	UCI
Cooling	768	8	UCI
Heating	768	8	UCI
Istanbul	536	7	UCI
Kin8fh	8192	8	Delve
Kin8fm	8192	8	Delve
Kin8nh	8192	8	Delve
Kin8nm	8192	8	Delve
Laser	993	4	KEEL
Puma8fh	8192	8	Delve
Puma8fm	8192	8	Delve
Puma8nh	8192	8	Delve
Puma8nm	8192	8	Delve
Stock	950	9	KEEL
Treasury	1048	15	Delve

The experiments were conducted on 20 public data sets, which are from Delve [22], KEEL [4] and UCI [5] repositories whose detailed information is summarized in Table 1. The features and labels were all normalized to [0, 1] with min–max normalization. Tenfold cross-validation was used to test the algorithms, i.e., each data set was randomly split into tenfolds, where each fold was used to evaluate the algorithms trained from the other ninefolds and the mean of ten results for each algorithm was reported. All the algorithms in this section were coded with python based on numpy and scikit-learn library and the experimental results were collected from the computer with 3.5 GHz CPU and 32 GB RAM.

Test the validity of LW-JPS

This section tests whether LW-JPS is a valid predictive system by the definition of formula (1). To do so, the values of the CDF of LW-JPS on the test data were collected and the frequency of the values being not more than α was calculated, whose results are shown in Table 2 with mean representing the mean value of each column. Table 2 demonstrates that

Table 2 Validity test of LW-JPS

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Abalone	0.099	0.202	0.302	0.401	0.495	0.601	0.698	0.800	0.900
Bank8fh	0.099	0.200	0.299	0.397	0.498	0.599	0.700	0.798	0.900
Bank8fm	0.099	0.201	0.301	0.401	0.498	0.601	0.699	0.801	0.900
Bank8nh	0.100	0.198	0.300	0.399	0.500	0.601	0.702	0.801	0.899
Bank8nm	0.100	0.202	0.300	0.403	0.499	0.601	0.698	0.798	0.898
Boston	0.099	0.200	0.293	0.419	0.508	0.601	0.713	0.808	0.911
Cooling	0.091	0.202	0.294	0.396	0.503	0.603	0.707	0.814	0.908
Heating	0.104	0.207	0.315	0.398	0.479	0.592	0.703	0.797	0.901
Istanbul	0.099	0.196	0.297	0.399	0.508	0.597	0.700	0.797	0.899
Kin8fh	0.101	0.200	0.298	0.401	0.498	0.600	0.700	0.800	0.900
Kin8fm	0.099	0.201	0.300	0.401	0.501	0.600	0.698	0.798	0.902
Kin8nh	0.101	0.200	0.300	0.396	0.501	0.600	0.704	0.802	0.901
Kin8nm	0.101	0.199	0.299	0.401	0.501	0.598	0.700	0.798	0.903
Laser	0.099	0.203	0.292	0.397	0.493	0.602	0.704	0.793	0.903
Puma8fh	0.100	0.200	0.301	0.401	0.501	0.598	0.699	0.798	0.899
Puma8fm	0.100	0.200	0.299	0.400	0.501	0.601	0.701	0.801	0.899
Puma8nh	0.100	0.201	0.301	0.400	0.500	0.600	0.700	0.800	0.900
Puma8nm	0.100	0.200	0.298	0.405	0.499	0.600	0.699	0.802	0.901
Stock	0.101	0.201	0.293	0.394	0.502	0.602	0.702	0.795	0.900
Treasury	0.097	0.203	0.303	0.397	0.496	0.602	0.692	0.797	0.900
Mean	0.099	0.201	0.299	0.400	0.499	0.600	0.701	0.800	0.901

the frequencies are compatible with corresponding α, which empirically proves the validity of LW-JPS.

As we analyze in "Conformal predictive systems and locally weighted jackknife predictive system", the validity property of formula (1) implies the coverage guarantee by formula (2), which will be shown in the next experiment.

Comparison with the other CPSs

This section compares the performance of LW-JPS with SCP-S–SVR, SCPS–RF, CCPS–SVR, CCPS–RF and OOB–CP-S–RF. To compare the quality of the predictive distributions, the widely used continuous ranked probability score (CRPS) are employed whose definition can be found in [34]. The lower the CRPS is, the better the predictive distribution is. The barplots of the mean of continuous ranked probability scores for different data sets are shown in Fig. 1, which demonstrates that LW-JPS performs better in most cases. Table 3 records the mean CRPS of all algorithms, with the least one of each data set shown in bold. For each data set, the rank of an algorithm is obtained and the mean rank in Table 3 is the mean value of all ranks for each algorithm. From Table 3, we can see that the LW-JPS performs better than the other predictive systems, which indicates the effectiveness of LW-JPS.

We also test the derived prediction intervals from the predictive distributions of all predictive systems. For a significance levelη, which is the expected coverage rate preset by practitioners, the derived prediction interval is based on formula (2) with the help of $\eta/2$ and $1-\eta/2$ quantiles. Two indicators are employed to describe the quality of prediction intervals. One is the prediction error rate, which is the frequency of the true label being out of the prediction intervals. The other is the average interval size, which measures the information efficiency of the prediction intervals. The smaller the average interval size, the more information the prediction intervals contain. We set the significance levels as 0.2, 0.1 and 0.05 and show the experimental results in Tables 4, 5 and 6 for error rates and in Tables 7, 8 and 9 for average interval sizes. We also summarize the error rates, the means and mean ranks of average interval sizes in Figs. 2, 3, and 4.

From Tables 4, 5 and 6, we can see that all predictive systems are empirically valid for the data sets, which also empirically proves the coverage guarantee of LW-JPS. Besides, it is shown in Tables 7, 8 and 9 that prediction intervals of LW-JPS are more informationally efficient than those of the other algorithms, which is demonstrated in Figs. 3 and 4. The box plots of average interval size are also shown in

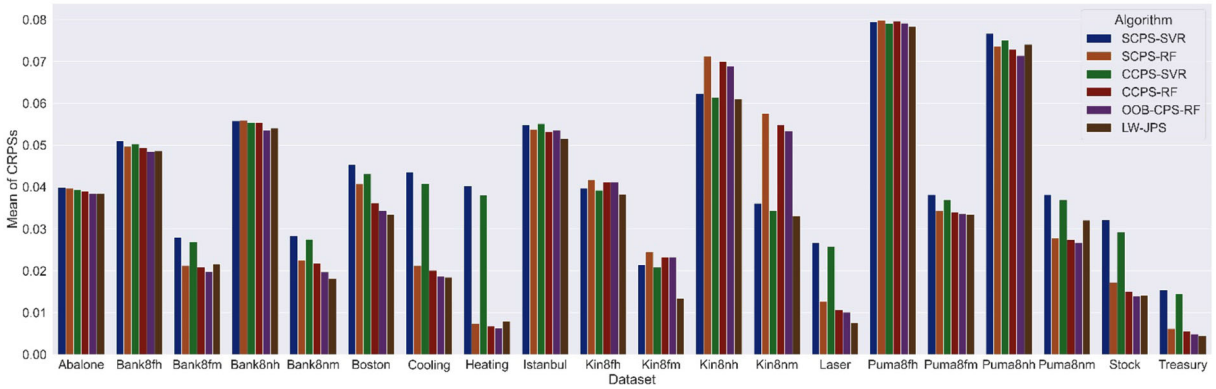

Fig. 1 Mean of continuous ranked probability scores for different algorithms trained on different data sets

Table 3 The mean CRPS of all algorithms

	SCPS–SVR	SCPS–RF	CCPS–SVR	CCPS–RF	OOB–CPS–RF	LW-JPS
Abalone	0.0399	0.0396	0.0393	0.0390	0.0385	**0.0384**
Bank8fh	0.0510	0.0496	0.0502	0.0494	**0.0484**	0.0486
Bank8fm	0.0280	0.0212	0.0270	0.0209	**0.0198**	0.0217
Bank8nh	0.0558	0.0559	0.0554	0.0554	**0.0536**	0.0540
Bank8nm	0.0283	0.0226	0.0276	0.0218	0.0197	**0.0182**
Boston	0.0453	0.0407	0.0431	0.0363	0.0345	**0.0335**
Cooling	0.0436	0.0212	0.0408	0.0201	0.0187	**0.0184**
Heating	0.0402	0.0074	0.0380	0.0068	**0.0063**	0.0080
Istanbul	0.0549	0.0538	0.0551	0.0532	0.0535	**0.0515**
Kin8fh	0.0396	0.0417	0.0392	0.0411	0.0411	**0.0382**
Kin8fm	0.0214	0.0246	0.0208	0.0234	0.0232	**0.0134**
Kin8nh	0.0623	0.0713	0.0614	0.0699	0.0689	**0.0610**
Kin8nm	0.0361	0.0576	0.0344	0.0548	0.0534	**0.0332**
Laser	0.0267	0.0128	0.0258	0.0107	0.0101	**0.0076**
Puma8fh	0.0795	0.0799	0.0791	0.0796	0.0792	**0.0784**
Puma8fm	0.0381	0.0343	0.0370	0.0340	0.0337	**0.0335**
Puma8nh	0.0767	0.0736	0.0751	0.0729	**0.0714**	0.0741
Puma8nm	0.0381	0.0279	0.0369	0.0275	**0.0268**	0.0321
Stock	0.0322	0.0172	0.0293	0.0150	**0.0140**	0.0142
Treasury	0.0155	0.0062	0.0145	0.0056	0.0049	**0.0045**
Mean	0.0427	0.0380	0.0415	0.0369	0.0360	**0.0341**
Mean rank	5.2000	4.4500	4.1500	3.3000	2.1500	**1.7500**

Bold indicates the mean CRPS of all algorithms, with the least one of each data set shown in bold

Fig. 5, which also demonstrates that JPS performs better than other CPSs.

We also conducted Wilcoxon test [10] to answer the question of whether LW-JPS performs better than other comparison algorithms significantly. Table 10 demonstrates the p values of the experimental results about CRPS and average interval sizes with $\eta \in \{0.2, 0.1, 0.05\}$ and the bold values are less than 0.05, which shows the significant differences. From Table 10, we can see that LW-JPS significantly performs better than SCPS–SVR, SCPS–RF, CCPS–SVR and CCPS–RF, and the differences between LW-JPS and OOB–CPS–RF are not significant in most cases. Since OOB–CPS–RF represents the state-of-the-art process using conformal approach for regression problems, the statistical tests confirm the effectiveness of LW-JPS for probabilistic prediction.

For training speed, all of the algorithms are computationally efficient versions of CPSs and the mean values of

Table 4 Error rate ($\eta = 0.2$)

	SCPS–SVR	SCPS–RF	CCPS–SVR	CCPS–RF	OOB–CPS–RF	LW-JPS
Abalone	0.197	0.204	0.191	0.200	0.197	0.200
Bank8fh	0.204	0.202	0.199	0.198	0.196	0.199
Bank8fm	0.196	0.203	0.202	0.200	0.197	0.199
Bank8nh	0.199	0.196	0.202	0.202	0.197	0.202
Bank8nm	0.201	0.199	0.202	0.200	0.194	0.201
Boston	0.216	0.198	0.212	0.208	0.209	0.196
Cooling	0.185	0.195	0.221	0.208	0.202	0.184
Heating	0.184	0.211	0.203	0.203	0.198	0.204
Istanbul	0.207	0.196	0.207	0.194	0.202	0.200
Kin8fh	0.202	0.202	0.198	0.199	0.197	0.201
Kin8fm	0.199	0.205	0.201	0.198	0.195	0.197
Kin8nh	0.198	0.198	0.204	0.200	0.195	0.199
Kin8nm	0.198	0.200	0.205	0.192	0.199	0.198
Laser	0.201	0.213	0.224	0.180	0.205	0.198
Puma8fh	0.204	0.196	0.198	0.201	0.200	0.200
Puma8fm	0.199	0.195	0.198	0.198	0.198	0.201
Puma8nh	0.205	0.201	0.197	0.201	0.197	0.199
Puma8nm	0.205	0.199	0.202	0.200	0.197	0.200
Stock	0.223	0.198	0.188	0.198	0.173	0.201
Treasury	0.207	0.174	0.239	0.208	0.202	0.197
Mean	0.202	0.199	0.205	0.199	0.198	0.199
Mean rank	4.200	3.425	4.250	3.375	2.350	3.400

Table 5 Error rate ($\eta = 0.1$)

	SCPS–SVR	SCPS–RF	CCPS–SVR	CCPS–RF	OOB–CPS–RF	LW-JPS
Abalone	0.101	0.103	0.100	0.101	0.097	0.100
Bank8fh	0.106	0.102	0.102	0.102	0.097	0.101
Bank8fm	0.094	0.101	0.101	0.096	0.102	0.100
Bank8nh	0.098	0.100	0.100	0.102	0.097	0.099
Bank8nm	0.102	0.100	0.100	0.099	0.093	0.100
Boston	0.109	0.095	0.097	0.097	0.089	0.105
Cooling	0.086	0.107	0.118	0.095	0.098	0.095
Heating	0.104	0.103	0.102	0.106	0.090	0.096
Istanbul	0.101	0.097	0.097	0.103	0.103	0.104
Kin8fh	0.097	0.103	0.097	0.102	0.099	0.100
Kin8fm	0.098	0.102	0.099	0.098	0.100	0.102
Kin8nh	0.100	0.098	0.101	0.103	0.100	0.098
Kin8nm	0.100	0.100	0.098	0.094	0.095	0.099
Laser	0.090	0.102	0.098	0.089	0.103	0.101
Puma8fh	0.096	0.098	0.100	0.101	0.095	0.101
Puma8fm	0.104	0.103	0.099	0.099	0.098	0.101
Puma8nh	0.096	0.100	0.097	0.100	0.098	0.102
Puma8nm	0.102	0.095	0.099	0.098	0.097	0.103
Stock	0.112	0.097	0.089	0.100	0.106	0.096
Treasury	0.112	0.099	0.119	0.104	0.106	0.108
Mean	0.100	0.100	0.101	0.099	0.098	0.101
Mean rank	3.650	3.800	3.400	3.600	2.750	3.800

Table 6 Error rate ($\eta = 0.05$)

	SCPS–SVR	SCPS–RF	CCPS–SVR	CCPS–RF	OOB–CPS–RF	LW-JPS
Abalone	0.052	0.056	0.048	0.050	0.047	0.051
Bank8fh	0.053	0.050	0.051	0.051	0.050	0.050
Bank8fm	0.045	0.049	0.050	0.050	0.049	0.050
Bank8nh	0.047	0.050	0.052	0.050	0.049	0.049
Bank8nm	0.048	0.051	0.048	0.051	0.043	0.049
Boston	0.051	0.049	0.037	0.047	0.045	0.051
Cooling	0.050	0.062	0.061	0.053	0.046	0.048
Heating	0.050	0.057	0.057	0.057	0.039	0.056
Istanbul	0.047	0.060	0.056	0.056	0.050	0.047
Kin8fh	0.047	0.048	0.049	0.049	0.049	0.050
Kin8fm	0.049	0.049	0.047	0.050	0.052	0.049
Kin8nh	0.051	0.048	0.049	0.051	0.048	0.050
Kin8nm	0.050	0.049	0.048	0.046	0.047	0.049
Laser	0.046	0.047	0.042	0.048	0.045	0.054
Puma8fh	0.050	0.048	0.050	0.050	0.048	0.051
Puma8fm	0.051	0.053	0.051	0.050	0.048	0.048
Puma8nh	0.049	0.051	0.049	0.049	0.047	0.049
Puma8nm	0.050	0.047	0.049	0.049	0.048	0.052
Stock	0.058	0.055	0.049	0.048	0.044	0.049
Treasury	0.058	0.053	0.059	0.049	0.055	0.051
Mean	0.050	0.052	0.050	0.050	0.048	0.050
Mean rank	3.700	3.950	3.725	3.700	2.050	3.875

Table 7 Average interval size ($\eta = 0.2$)

	SCPS–SVR	SCPS–RF	CCPS–SVR	CCPS–RF	OOB–CPS–RF	LW-JPS
Abalone	0.175	0.170	0.171	0.169	0.170	**0.167**
Bank8fh	0.226	0.219	0.225	0.220	**0.208**	0.215
Bank8fm	0.129	0.095	0.122	0.094	**0.088**	0.097
Bank8nh	0.243	0.242	0.239	0.236	**0.223**	0.231
Bank8nm	0.126	0.086	0.122	0.082	0.086	**0.073**
Boston	0.195	0.185	0.195	0.158	0.152	**0.151**
Cooling	0.199	0.101	0.177	0.091	**0.085**	0.087
Heating	0.176	0.031	0.163	**0.030**	0.031	0.035
Istanbul	0.241	0.237	0.243	0.233	0.230	**0.227**
Kin8fh	0.181	0.189	0.180	0.188	0.190	**0.175**
Kin8fm	0.096	0.108	0.094	0.105	0.106	**0.061**
Kin8nh	0.283	0.328	**0.278**	0.316	0.317	**0.278**
Kin8nm	0.163	0.259	0.155	0.249	0.246	**0.146**
Laser	0.119	0.041	0.112	0.042	0.041	**0.031**
Puma8fh	0.354	0.358	0.353	0.353	0.353	**0.350**
Puma8fm	0.171	0.154	0.167	0.152	0.154	**0.149**
Puma8nh	0.343	0.328	0.336	0.323	**0.313**	0.329
Puma8nm	0.170	0.126	0.166	0.123	**0.121**	0.143
Stock	0.141	0.076	0.133	0.069	0.068	**0.064**
Treasury	0.096	0.025	0.100	0.021	**0.020**	**0.020**
Mean	0.191	0.168	0.187	0.163	0.160	**0.152**
Mean rank	5.200	4.100	4.400	2.900	2.600	**1.800**

Bold indicates the mean CRPS of all algorithms, with the least one of each data set shown in bold

Table 8 Average interval size ($\eta = 0.1$)

	SCPS–SVR	SCPS–RF	CCPS–SVR	CCPS–RF	OOB–CPS–RF	LW-JPS
Abalone	0.237	0.230	0.234	0.230	0.230	**0.228**
Bank8fh	0.296	0.297	0.295	0.294	**0.284**	0.290
Bank8fm	0.165	0.129	0.155	0.127	**0.113**	0.126
Bank8nh	0.336	0.345	0.324	0.336	**0.319**	0.325
Bank8nm	0.158	0.134	0.154	0.126	0.114	**0.104**
Boston	0.257	0.255	0.263	0.218	0.208	**0.207**
Cooling	0.231	0.134	0.202	0.116	0.115	**0.110**
Heating	0.200	0.041	0.188	**0.038**	0.044	0.044
Istanbul	0.355	0.337	0.325	0.319	0.325	**0.313**
Kin8fh	0.234	0.242	0.231	0.240	0.245	**0.223**
Kin8fm	0.124	0.138	0.120	0.133	0.139	**0.078**
Kin8nh	0.364	0.412	**0.357**	0.400	0.401	**0.357**
Kin8nm	0.209	0.325	0.200	0.307	0.306	**0.188**
Laser	0.148	0.060	0.143	0.056	0.057	**0.040**
Puma8fh	0.459	0.464	0.449	0.452	0.448	**0.447**
Puma8fm	0.220	0.201	0.216	0.198	0.199	**0.195**
Puma8nh	0.458	0.443	0.442	0.433	**0.406**	0.439
Puma8nm	0.218	0.173	0.213	0.167	**0.160**	0.188
Stock	0.174	0.102	0.161	0.092	0.088	**0.084**
Treasury	0.130	0.044	0.124	0.036	0.028	**0.027**
Mean	0.249	0.225	0.240	0.216	0.211	**0.201**
Mean rank	5.150	4.450	4.000	3.050	2.750	**1.600**

Bold indicates the mean CRPS of all algorithms, with the least one of each data set shown in bold

Table 9 Average interval size ($\eta = 0.05$)

	SCPS–SVR	SCPS–RF	CCPS–SVR	CCPS–RF	OOB–CPS–RF	LW-JPS
Abalone	0.292	0.293	0.294	0.294	0.295	**0.287**
Bank8fh	0.358	0.375	0.359	0.374	0.364	**0.361**
Bank8fm	0.195	0.161	0.182	0.156	**0.136**	0.150
Bank8nh	0.426	0.453	**0.417**	0.443	0.427	0.431
Bank8nm	0.190	0.177	0.184	0.165	0.140	**0.136**
Boston	0.331	0.333	0.324	0.281	0.264	**0.251**
Cooling	0.253	0.159	0.224	0.139	0.143	**0.129**
Heating	0.230	0.050	0.209	**0.046**	0.058	0.053
Istanbul	0.470	0.433	0.425	0.419	0.430	**0.398**
Kin8fh	0.277	0.291	0.275	0.287	0.293	**0.265**
Kin8fm	0.149	0.162	0.142	0.156	0.169	**0.092**
Kin8nh	0.432	0.482	0.424	0.469	0.470	**0.423**
Kin8nm	0.249	0.377	0.239	0.357	0.358	**0.228**
Laser	0.170	0.086	0.166	0.073	0.074	**0.049**
Puma8fh	0.552	0.562	0.539	0.548	0.544	**0.535**
Puma8fm	0.265	0.242	0.259	0.241	0.241	**0.236**
Puma8nh	0.552	0.539	0.537	0.537	**0.503**	0.532
Puma8nm	0.261	0.216	0.255	0.207	**0.197**	0.228
Stock	0.206	0.125	0.181	0.114	0.110	**0.102**
Treasury	0.159	0.062	0.142	0.053	0.040	**0.034**
Mean	0.301	0.279	0.289	0.268	0.263	**0.246**
Mean rank	4.650	4.600	3.650	3.250	3.250	**1.600**

Bold indicates the mean CRPS of all algorithms, with the least one of each data set shown in bold

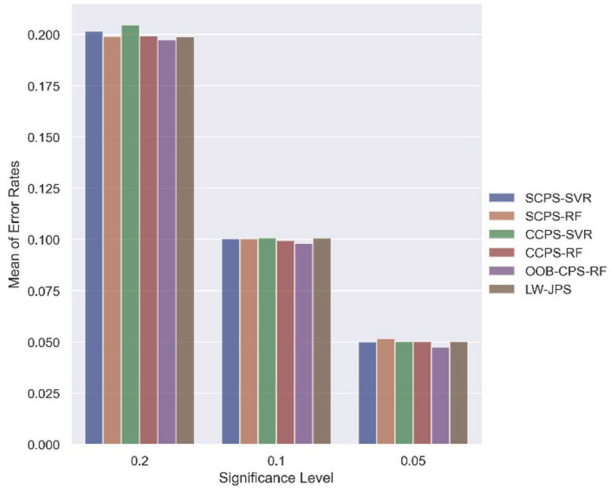

Fig. 2 Mean of prediction error rates of the prediction intervals derived from the predictive distributions

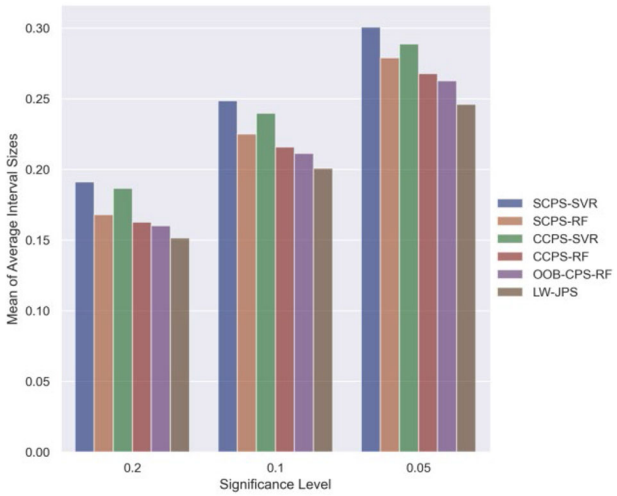

Fig. 3 Mean of average interval sizes of the prediction intervals derived from the predictive distributions

the training times of SCPS–SVR, SCPS–RF, CCPS–SVR, CCPS–RF, OOB–CPS–RF and LW-JPS on 20 data sets are 0.293 s, 8.704 s, 1.940 s, 59.336 s, 15.393 s and 1.443 s, respectively, indicating that the LW-JPS used in this paper is also computationally efficient.

In summary, the experimental results in this section not only verifies the empirical validity of LW-JPS, but also shows its better performance than the other comparison algorithms, which indicates the effectiveness and efficiency of LW-JPS for probabilistic prediction.

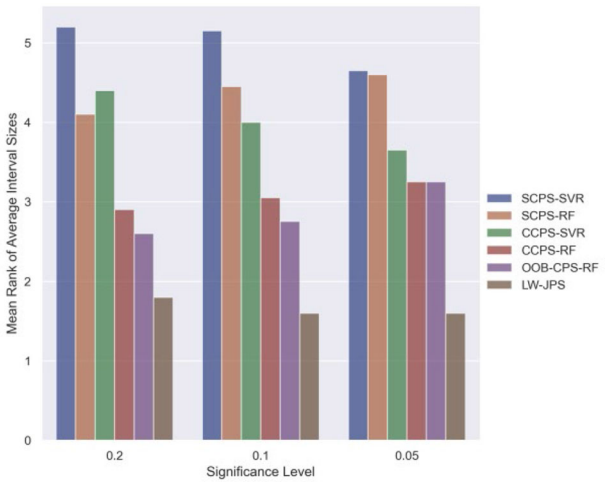

Fig. 4 Mean rank of average interval sizes of the prediction intervals derived from the predictive distributions

Conclusion

This paper proposes a predictive system based on the idea of jackknife prediction, which is inspired by the leave-one-out cross-conformal predictive system. The proposed LW-JPS can transform any regression algorithm for point prediction to probabilistic prediction, which can describe the uncertainty of test labels. The asymptotic validity of LW-JPS is proved with some regularity assumptions and conditions. Based on the analysis, the empirical testing of LW-JPS with randomized kernel ridge regression and k-nearest neighbor regression was conducted. The empirical validity of LW-JPS was demonstrated in the experiments and its performance for probabilistic prediction compared favourably with the other comparison algorithms, which demonstrates the effectiveness and efficiency of LW-JPS for probabilistic prediction.

Although our method is empirically valid and shows better performance when compared with other comparison CPSs, we only employ two representative regression algorithms satisfying the related conditions in this paper. Therefore, future work about empirical studies with a wider range of regression algorithms needs to be done. Moreover, the approach of LW-JPS we proposed in this paper cannot be built on deep learning models efficiently for complex learning problems, such as image segmentation or image-to-image regression problems, since in those cases, there are no efficient ways to compute leave-one-out predictions on training data. Thus, future work about approximately computing leave-one-out predictions for deep neural networks is worth exploring, in order to make the jackknife prediction approach more tractable for complex problems.

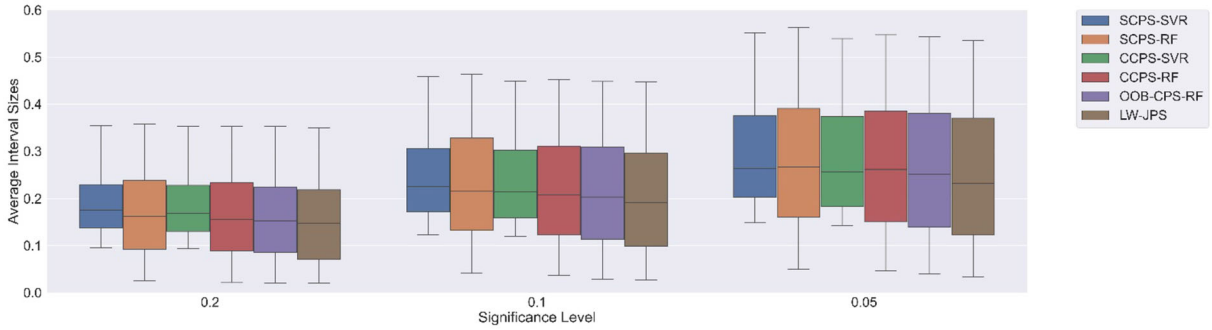

Fig. 5 Box plots of average interval sizes of the prediction intervals derived from the predictive distributions

Table 10 The p values of Wilcoxon tests

	CRPS	$\eta = 0.2$	$\eta = 0.1$	$\eta = 0.05$
LW-JPS vs SCPS–SVR	**1.907E-06**	**1.907E-06**	**1.907E-06**	**9.537E-06**
LW-JPS vs SCPS–RF	**7.076E-04**	**0.001**	**6.294E-05**	**3.624E-05**
LW-JPS vs CCPS–SVR	**1.907E-06**	**3.814E-06**	**9.537E-06**	**8.202E-05**
LW-JPS vs CCPS–RF	**0.009**	**0.021**	**0.004**	**3.223E-04**
LW-JPS vs OOB–CPS–RF	0.154	0.312	0.143	**0.024**

Bold indicates the mean CRPS of all algorithms, with the least one of each data set shown in bold

Acknowledgements The authors would like to thank the anonymous editor and reviewers for their valuable comments and suggestions which improved this work.

Funding This work was supported by the National Natural Science Foundation of China under Grant 62106169 and 61972282.

Data availability The data used during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. MT Abdulkhaleq TA Rashid A Alsadoon 2022 Harmony search: current studies and uses on healthcare systems Artif Intell Med 131 102348 https://doi.org/10.1016/j.artmed.2022.102348
2. MT Abdulkhaleq TA Rashid BA Hassan 2023 Fitness dependent optimizer with neural networks for COVID-19 patients Comput Methods Programs Biomed Update 3 100090 https://doi.org/10.1016/j.cmpbup.2022.100090
3. JM Abdullah TA Rashid BB Maaroof 2023 Multi-objective fitness-dependent optimizer algorithm Neural Comput Appl https://doi.org/10.1007/s00521-023-08332-3
4. J Alcala-Fdez A Fernandez J Luengo 2011 KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework J Mult-Valued Log Soft Comput 17 2–3 255 287 https://doi.org/10.1016/j.jlap.2009.12.002
5. Asuncion A, Newman D (2007) UCI machine learning repository Irvine. Retrieved from http://www.ics.uci.edu/~mlearn/MLRepository.html. Accessed 01 Jan 2013
6. V Balasubramanian S-S Ho V Vovk 2014 Conformal prediction for reliable machine learning: theory, adaptations and applications Morgan Kaufmann Publishers Inc. Newnes
7. Boström H, Linusson H, Löfström T et al (2016) Evaluation of a variance-based nonconformity measure for regression forests. Paper presented at the 5th International Symposium on Conformal and Probabilistic Prediction with Applications, COPA 2016, Madrid, Spain, 9653, pp 75–89
8. O Bousquet A Elisseeff 2002 Stability and generalization J Mach Learn Res 2 3 499 526 https://doi.org/10.1162/153244302760200704
9. F Cucker DX Zhou 2007 Learning theory: an approximation theory viewpoint Cambridge monographs on applied and computational mathematics Cambridge University Press

10. J Derrac S García D Molina 2011 A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms Swarm Evol Comput 1 1 3 18 https://doi.org/10.1016/j.swevo.2011.02.002
11. T Gneiting M Katzfuss 2014 Probabilistic forecasting Ann Rev Stat Appl 1 1 125 151 https://doi.org/10.1146/annurev-statistics-062713-085831
12. G Huang H Zhou X Ding 2012 Extreme learning machine for regression and multiclass classification IEEE Trans Syst Man Cybernetics Part B (Cybernetics) 42 2 513 529 https://doi.org/10.1109/TSMCB.2011.2168604
13. Y LeCun Y Bengio G Hinton 2015 Deep learning Nature 521 7553 436 444 https://doi.org/10.1038/nature14539
14. J Lei M G'Sell A Rinaldo 2018 Distribution-free predictive inference for regression J Am Stat Assoc 113 523 1094 1111 https://doi.org/10.1080/01621459.2017.1307116
15. Z Li F Liu W Yang 2022 A survey of convolutional neural networks: analysis, applications, and prospects IEEE Trans Neural Netw Learn Syst 33 12 6999 7019 https://doi.org/10.1109/TNNLS.2021.3084827
16. BB Maaroof TA Rashid JM Abdulla 2022 Current studies and applications of shuffled frog leaping algorithm: a review Arch Computat Methods Eng 29 5 3459 3474 https://doi.org/10.1007/s11831-021-09707-2
17. A Mahmoodzadeh HR Nejati M Mohammadi 2022 Forecasting tunnel boring machine penetration rate using LSTM deep neural network optimized by grey wolf optimization algorithm Expert Syst Appl 209 118303 https://doi.org/10.1016/j.eswa.2022.118303
18. Melluish T, Saunders C, Nouretdinov I et al (2001) Comparing the Bayes and typicalness frameworks. Paper presented at the 12th European Conference on Machine Learning, ECML 2001, Freiburg, Germany, 2167, pp 360–371
19. MR Mohebbian HR Marateb KA Wahid 2023 Semi-supervised active transfer learning for fetal ECG arrhythmia detection Comput Methods Programs Biomed Update 3 100096 https://doi.org/10.1016/j.cmpbup.2023.100096
20. H Papadopoulos 2008 Inductive conformal prediction: theory and application to neural networks P Fritzsche Eds Tools in artificial intelligence IntechOpen London https://doi.org/10.5772/6078
21. Rahimi A, Recht B (2007) Random features for large-scale kernel machines. Paper presented at the Advances in Neural Information Processing Systems 20 (NIPS 2007), Vancouver, British Columbia, Canada, 3 (4): 1177–1184
22. Rasmussen CE, Neal RM, Hinton G et al. Delve data for evaluating learning in valid experiments, 1995–1996. Retrieved from: https://www.cs.toronto.edu/~delve/. Accessed 01 Mar 2003
23. A Sayeed Y Choi J Jung 2023 A deep convolutional neural network model for improving WRF simulations IEEE Trans Neural Netw Learn Syst 34 2 750 760 https://doi.org/10.1109/TNNLS.2021.3100902
24. J Schmidhuber 2015 Deep learning in neural networks: an overview Neural Netw 61 85 117 https://doi.org/10.1016/j.neunet.2014.09.003
25. T Schweder NL Hjort 2016 Confidence, likelihood, probability 41 Cambridge University Press Cambridge
26. J Shao 2003 Mathematical statistics Springer Science and Business Media New York
27. J Shen RY Liu M-g Xie 2018 Prediction with confidence—a general framework for predictive inference J Statist Plan Inference 195 126 140 https://doi.org/10.1016/j.jspi.2017.09.012
28. Steinberger L, Leeb H (2016) Leave-one-out prediction intervals in linear regression models with many variables. arXiv e-prints, arXiv:1602.05801. https://doi.org/10.48550/arXiv.1602.05801
29. AW Vaart van der 2000 Asymptotic statistics 3 Cambridge University Press Cambridge
30. V Vovk 2015 Cross-conformal predictors Ann Math Artif Intell 74 1 9 28 https://doi.org/10.1007/s10472-013-9368-4
31. Vovk V (2019) Universally consistent conformal predictive distributions. Paper presented at the Proceedings of the Eighth Symposium on Conformal and Probabilistic Prediction and Applications, 105, pp 105-122
32. V Vovk A Gammerman G Shafer 2005 Algorithmic learning in a random world Springer Science and Business Media New York
33. Vovk V, Nouretdinov I, Manokhin V et al (2018) Conformal Predictive Distributions with Kernels. Paper presented at the International Conference Commemorating the 40th Anniversary of Emmanuil Braverman's Decease, Boston, MA, USA, 11100, pp 103–121
34. Vovk V, Nouretdinov I, Manokhin V et al (2018) Cross-conformal predictive distributions. Paper presented at the Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications, 91, pp 37–51
35. Vovk V, Petej I, Toccaceli P et al (2020) Conformal calibrators. Paper presented at the Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications, Proceedings of Machine Learning Research, 128, pp 84–99
36. D Wang P Wang J Shi 2018 A fast and efficient conformal regressor with regularized extreme learning machine Neurocomputing 304 1 11 https://doi.org/10.1016/j.neucom.2018.04.012
37. D Wang P Wang Y Yuan 2020 A fast conformal predictive system with regularized extreme learning machine Neural Netw 126 347 361 https://doi.org/10.1016/j.neunet.2020.03.022
38. D Wang P Wang S Zhuang 2020 Asymptotic analysis of locally weighted jackknife prediction Neurocomputing 417 10 22 https://doi.org/10.1016/j.neucom.2020.07.074
39. L Wasserman 2006 All of nonparametric statistics Springer Science and Business Media New York
40. Werner H, Carlsson L, Ahlberg E et al (2020) Evaluating different approaches to calibrating conformal predictive systems. Paper presented at the Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications, 128, pp 134–150