Estimation of Conditional Density Functions by Conformal Prediction and Model Averaging

Jinhao Zhao, Guangyuan Cui*

***Abstract*—In this research, we put forward a type of model averaging methods for estimating the conditional density function, based on a non-parametric estimation method and two different loss functions. Such methods provide accurate and stable estimation of the conditional density function. In addition, we develop prediction bands for the conditional density function in the case of finite samples by combining conformal prediction and model averaging. Conclusions from computational simulations and real data assessments based on photometric redshift estimation indicate the superiority of our proposed methods in comparison to other alternative methods.**

***Index Terms*—Model Averaging, Conformal Prediction, Photometric Redshift Estimation**

I. Introduction

REDSHIFT is a critical parameter for measuring the distance between galaxies and Earth, and it plays an important role in inferring cosmological model parameters. The goal of estimating photometric redshift is to infer the redshift Z of a galaxy from the observed photometric measurements $\boldsymbol{X}$. By establishing a linear regression model, researchers study the correspondence between $\boldsymbol{X}$ and Z, and take $\mathbb{E}(Z \mid \boldsymbol{x})$ as the appraisal result. However, since the conditional density $f(z \mid \boldsymbol{x})$ is often non-symmetric and multimodal, instead of a Gaussian distribution, and the noise often possesses heteroscedasticity, there is little useful information contained in point estimation results of the redshift [1], [2], [3], [4]. Furthermore, due to the possibility of two galaxies showing diverse redshifts exhibiting analogous photometric features and the presence of complex observational noise, the conditional density $f(z \mid \boldsymbol{x})$ actually provides a better description than $\mathbb{E}(Z \mid \boldsymbol{x})$. The estimation of $f(z \mid \boldsymbol{x})$ significantly reduces systematic errors in downstream cosmological analyses [1], [2], [4], [5].

There is a growing number of studies that focus on the estimation of $f(z \mid \boldsymbol{x})$. To illustrate, in the situation of low-dimensional covariates, researchers propose several non-parametric methods to estimate $f(z \mid \boldsymbol{x})$. Among these methods, the first step is to assess $f(z, \boldsymbol{x})$ and $f(\boldsymbol{x})$ separately using divergent techniques and then combine them by $f(z \mid \boldsymbol{x}) = f(z, \boldsymbol{x})/f(\boldsymbol{x})$ [6], [7]. Such methods include local polynomial regression [8], least squares estimation [9], and quantile estimation [10]. For covariates with a moderate dimension, Hall et al. [11] developed a technique for optimizing the parameters of a kernel density estimator, and this method identifies that elements of the covariates $\boldsymbol{X}$ are related to $f(z \mid \boldsymbol{x})$. However, since this approach heavily relies on selecting different bandwidths for each component, it is computationally cumbersome. Moreover, such methods typically require a pre-set dimension reduction step, which can lead to significant information loss. Efromovich [12] proposed an orthogonal series estimator of $f(z \mid \boldsymbol{x})$. This approach automatically performs a dimension reduction procedure on the covariates. However, this method is incapable of handling high-dimensional covariates, because it requires the calculation of a tensor product. To deal with this problem, Izbicki and Lee [13] proposed the Flexible non-parametric conditional density estimation via regression (FlexCode) method, which transforms the high-dimensional conditional density estimation problem into an expansion coefficient estimation problem for an orthogonal series.

It is crucial to acknowledge that the data generating process remains unknown in practical problems. Thus, researchers usually build candidate models using the observed data to better approximate this process. We eventually build multiple candidate models. It is crucial to make sure that we select the true model among all candidate models, since subsequent statistical inferences and asymptotic properties based on incorrect models may be biased. This would further lead to incorrect judgments and decisions for practical problems. In this case, model selection is a helpful tool to select the model achieving the best performance outcomes given a certain criterion. Common model selection methods include AIC [14], BIC [15] and cross-validation [16]. However, there are certain potential drawbacks about model selection methods [17]: 1) High inferential risk. If the selected model prove to be incorrect, then the corresponding parameter estimates and statistical inference results may be biased, and researchers tend to underestimate the variance of parameter estimates and overestimate the actual coverage probability of a confidence interval [18], [19]. 2) Estimation instability. Yuan and Yang [20] define a measurement named Perturbation instability in estimation (PIE) to assess the model selection instability. They point out that when the PIE value is large, the model selection method exhibits greater instability, which may lead to unstable estimation or prediction results. Similar opinions are found in Leung and Barron [19]. 3) Loss of information. After obtaining the candidate model with the best performance, researchers may proceed with the subsequent analysis based on this selected model, while discarding the rest of the candidate models. This approach ignores the useful information contained in these discarded models [21].

To avoid potential problems caused by model selection methods, researchers utilize an alternative method called model averaging. This method combines the estimation or prediction results from each candidate model with a specific model weight. Intuitively, model averaging can avoid select-

Manuscript received Jan 26, 2024; revised Jun 19, 2024.

Jinhao Zhao is a postgraduate student at School of Management, University of Science and Technology of China, Hefei 230026, China (email: zjh3608@mail.ustc.edu.cn).

Guangyuan Cui is a PhD student at Department of Management Sciences, College of Business, City University of Hong Kong, Hong Kong (corresponding author to provide email: guangcui-c@my.cityu.edu.hk).

ing inferior models, thereby reducing the risk of estimation and prediction [19], [22]. Moreover, as a smooth extension of model selection, model averaging typically reduces loss of information and produces more stable estimates. According to the weight choice criterion, model averaging is chiefly divided into two categories: Bayesian model averaging (BMA) and Frequentist model averaging (FMA). See [23], [24], [25], [26] for detailed overviews. By treating the model structure as random, BMA assigns prior probabilities to both the model structure and the key parameter in each candidate model. Afterward, this method calculates the posterior probabilities of each candidate model, which are later used to construct model weights. However, BMA has several distinct drawbacks. First, there is no standard protocol for selecting prior distributions for both candidate models and the parameters of interest. Second, the calculation of posterior probabilities may involve complicated integrations, which can be computationally inefficient. Given this circumstance, FMA is an alternative method that establishes the weight choice criterion using frequentist methods, without the need to specify prior distributions.

Based on the different strategies for assigning model weights, FMA can be categorized into three types: information criterion model averaging, adaptive model averaging, and optimal model averaging. Buckland et al. [27] constructed model weights based on two types of information criteria and proposed Smoothed-AIC and Smoothed-BIC, respectively. This study primarily concentrates on the optimal model averaging method. Optimal model averaging is that under certain conditions, its estimator is asymptotically optimal, as it attains the minimal achievable bound of the loss function. Hansen [28] first proposed an optimal model averaging method named Mallows model averaging (MMA), which selects the model weights based on the Mallows criterion. The related model averaging estimator is demonstrated to possess asymptotic optimality under specified regularity conditions. However, this strategy bounds the model weights to a particular discrete subset, and all candidate models are required to be strictly nested. Wan et al. [29] extended the framework of Hansen [28] to a non-nested setting of candidate models and a continuous weight space, and showed asymptotic optimality under this setting. Hansen and Racine [30] developed Jackknife model averaging (JMA) for linear regression models. The JMA method develops the weight choice criterion, and this resultant estimator still maintains asymptotic optimality. In comparison to MMA, JMA extends the applicability to models characterized by heteroscedasticity and non-nested candidate models. Lin et al. [31] first developed a novel model averaging approach for density functions under the parametric framework. The idea of a model selection criterion, named Takeuchi information criterion (TIC) [32], is used to modify the weight choice criterion. Theoretical properties, including asymptotic optimality and the consistency of model weights, are provided. However, this method is still within the framework of parametric models and cannot be extended to the non-parametric framework, which is the motivation of our work.

Regarding the challenge of conditional density estimation, this paper proposes two types of model averaging methods based on different loss functions, i.e., the Kullback-Leibler (KL) divergence and conditional density estimation (CDE) loss. The weight choice criterion is set to be the cross-validation method to determine model weights for different candidate models. Furthermore, since point estimation may not contain enough information under certain special cases, we intend to provide prediction bands for the conditional density function in the finite sample case by combining optimal model averaging with Highest predictive density (HPD) conformal prediction. Moreover, we also combine MMA method with Inductive conformal prediction (ICP) to develop a prediction interval for point estimation. Simulation studies and factual data analysis based on photometric redshift estimation are carried out to assess the performance of our proposed model averaging methods in comparison to other alternative methods.

The subsequent sections of the research are arranged in the following manner. Section 2 details the foundational setup of the research problem and introduces how to construct candidate models. Then, we introduce the KL divergence and CDE loss, and propose the model averaging estimator. Section 3 briefly reviews the marginal validity and asymptotic properties of conformal prediction, and introduces the ICP and HPD conformal prediction. Building on the model averaging estimates in the previous section, two algorithms that combine model averaging and conformal prediction are presented. In Section 4, we conduct simulation experiments under two model settings. Section 5 employs the recommended technique to the practical problem of photometric redshift estimation, and the experimental results confirm the superiority of the two model averaging methods proposed in this research. Finally, in Section 6, we summarize the content of the whole paper and discuss related issues.

II. Cross-validation model averaging

In this section, we first outline the basic setup of the research problem and then detail the process of constructing candidate models. Next, we establish the procedure of model averaging estimation, and introduce the KL divergence and CDE loss. By utilizing cross-validation, we determine the weight choice criteria for both loss functions. Finally, we present the formulae of the model averaging estimators.

A. Model setup

We consider a set of independent and identically distributed samples $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$. Our goal is to determine the conditional density of Y_i based on the value of $\boldsymbol{X}_i$. Since the conditional density function of the true model is unknown, we construct an array of candidate models as approximations to the actual conditional density function. Each candidate model uses a different subset of the covariates $\boldsymbol{X}$ to establish a connection with the response variable Y, which forms M non-nested candidate models.

There have been various approaches to estimating the conditional density function [33], [34], [35], [36], [37]. In this study, we choose FlexCode as an example. First, this method specifies an orthogonal basis $(\phi_i)_{i\in\mathbb{N}}$ in $\mathcal{L}^2(\mathbb{R})$, to outline the conditional density $f(y \mid \boldsymbol{x})$ with respective to y, it is,

$$f(y \mid \boldsymbol{x}) = \sum_{i\in\mathbb{N}} \beta_i(\boldsymbol{x})\phi_i(y),$$

where

$$\begin{aligned}\beta_i(\boldsymbol{x}) &= \langle f(\cdot \mid \boldsymbol{x}), \phi_i \rangle \\ &= \int_{\mathbb{R}} \phi_i(y) f(y \mid \boldsymbol{x}) dy \\ &= \mathbb{E}\left[\phi_i(Y) \mid \boldsymbol{x}\right].\end{aligned}$$

The conditional density estimation under various candidate models are denoted by $\hat{f}_1(y \mid \boldsymbol{x}), \ldots, \hat{f}_M(y \mid \boldsymbol{x})$. By combining these estimates with a set of model weights $\boldsymbol{w} = (w_1, \ldots, w_M)^{\mathrm{T}}$, we obtain the model averaging estimation for the conditional density function as

$$\hat{f}_{\boldsymbol{w}}(y \mid \boldsymbol{x}) = \sum_{m=1}^{M} w_m \hat{f}_m(y \mid \boldsymbol{x}), \tag{1}$$

where

$$\boldsymbol{w} \in \mathcal{W} = \{\boldsymbol{w} \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\}.$$

a Since the weight w_m in Equation (1) remains unknown, we select the weight w_m based on the criterion of minimizing cross-validation after estimating $f_m(\cdot)$ by non-parametric methods.

B. Loss functions

We evaluate the difference between the model averaging estimator in (1) and the actual conditional density function $f(y \mid \boldsymbol{x})$ based on two types of loss functions: KL divergence and CDE loss. The KL divergence is used to assess the information loss between two distributions, while the CDE loss is used to measure the spatial proximity between two conditional distributions. The specific forms of the KL divergence and CDE loss are shown below:

$$\begin{aligned}KL(\hat{f}_{\boldsymbol{w}}, f) &= \mathbb{E}\left[\log\{f(y \mid \boldsymbol{x})\}\right] - \mathbb{E}\left[\log\{\hat{f}_{\boldsymbol{w}}(y \mid \boldsymbol{x})\}\right] \\ &= -\mathbb{E}\left[\log\left\{\sum_{m=1}^{M} w_m \hat{f}_m(y \mid \boldsymbol{x})\right\}\right] + C_1, \end{aligned} \tag{2}$$

$$\begin{aligned}&CDE(\hat{f}_{\boldsymbol{w}}, f) \\ &= \iint (\hat{f}_{\boldsymbol{w}}(y \mid \boldsymbol{x}) - f(y \mid \boldsymbol{x}))^2 dP(\boldsymbol{x}) dy \\ &= \iint \hat{f}^2_{\boldsymbol{w}}(y \mid \boldsymbol{x}) dP(\boldsymbol{x}) dy \\ &\quad - 2\iint \hat{f}_{\boldsymbol{w}}(y \mid \boldsymbol{x}) f(y, \boldsymbol{x}) d\boldsymbol{x} dy + C_2 \\ &= \iint \left(\sum_{m=1}^{M} w_m \hat{f}_m(y \mid \boldsymbol{x})\right)^2 dP(\boldsymbol{x}) dy \\ &\quad - 2\sum_{m=1}^{M} w_m \iint \hat{f}_m(y \mid \boldsymbol{x}) f(y, \boldsymbol{x}) d\boldsymbol{x} dy + C_2, \end{aligned} \tag{3}$$

where $C_1 = \mathbb{E}\left[\log\{f(y \mid \boldsymbol{x})\}\right]$ and $C_2 = \iint f^2(y \mid \boldsymbol{x}) dP(\boldsymbol{x}) dy$ are fixed values unrelated to estimation.

It is noteworthy that the KL divergence is a universal method applicable to estimation of the density and the conditional density function [38], while the CDE loss proposed by Izbicki et al. [13] is only applicable to estimation of the conditional density function.

Given the formulae of (2) and (3), we use the plug-in method to estimate them separately, and obtain

$$\widehat{KL}(\hat{f}_{\boldsymbol{w}}, f) = -\frac{1}{n}\sum_{i=1}^{n}\left(\log\{\sum_{m=1}^{M} w_m \hat{f}_m(Y_i \mid \boldsymbol{X}_i)\}\right),$$

$$\begin{aligned}\widehat{CDE}(\hat{f}_{\boldsymbol{w}}, f) &= \frac{1}{n}\sum_{i=1}^{n}\left(\int\left(\sum_{m=1}^{M} w_m \hat{f}_m(y \mid \boldsymbol{X}_i)\right)^2 dy\right) \\ &\quad - \sum_{i=1}^{n}\sum_{m=1}^{M} 2 w_m \hat{f}_m(Y_i \mid \boldsymbol{X}_i).\end{aligned}$$

C. Weight choice criteria

Similar to Hansen and Racine [30] and Zhang and Liu[39], we also consider using cross-validation methods to construct weight choice criteria. In the J-fold cross-validation, we randomly divide n samples into J folds, with each fold containing H samples (we take as a given that $H = n/J$ is a non-fractional number). Under the mth candidate model for the conditional density function, we select the lth fold sample as the test set and use the remaining $J-1$ fold samples for estimation of the conditional density function, resulting in the estimated conditional density $\hat{f}_{m,[-l]}(y \mid \boldsymbol{x})$. Next, we propose two types of weight choice criteria based on Equation (2) and (3). The weight choice criterion corresponding to the CDE loss function is shown as follows:

$$\begin{aligned}&CV_{CDE}(\boldsymbol{w}) \\ &= \sum_{l=1}^{J}\sum_{h=1}^{H}\left\{\int\left(\sum_{m=1}^{M} w_m \hat{f}_{m,[-l]}\left(y \mid \boldsymbol{X}_{(l-1)H+h}\right)\right)^2 dy\right\} \\ &\quad - \sum_{l=1}^{J}\sum_{h=1}^{H}\sum_{m=1}^{M} 2 w_m \hat{f}_{m,[-l]}\left(Y_{(l-1)H+h} \mid \boldsymbol{X}_{(l-1)H+h}\right).\end{aligned} \tag{4}$$

By minimizing Equation (4), we obtain the model weights under the CDE loss, as shown in Equation (5),

$$\hat{\boldsymbol{w}}_{CDE} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} CV_{CDE}(\boldsymbol{w}). \tag{5}$$

The weight choice criterion for the KL divergence is shown in Equation (6),

$$\begin{aligned}&CV_{KL}(\boldsymbol{w}) \\ &= \sum_{l=1}^{J}\sum_{h=1}^{H}\log\sum_{m=1}^{M} w_m \hat{f}_{m,[-l]}\left(Y_{(l-1)H+h} \mid \boldsymbol{X}_{(l-1)H+h}\right).\end{aligned} \tag{6}$$

By maximizing this equation, we obtain the model weights under the KL divergence, as shown in Equation (7),

$$\hat{\boldsymbol{w}}_{KL} = \arg\max_{\boldsymbol{w}\in\mathcal{W}} CV_{KL}(\boldsymbol{w}). \tag{7}$$

By substituting the optimal weights (5) and (7) into Equation (1), we can obtain the model averaging estimates corresponding to the CDE loss function and the KL divergence function:

$$\hat{f}_{\hat{\boldsymbol{w}}_{CDE}}(y \mid \boldsymbol{x}) = \sum_{m=1}^{M} \hat{w}_{CDE,m} \hat{f}_m(y \mid \boldsymbol{x}),$$

$$\hat{f}_{\hat{\boldsymbol{w}}_{KL}}(y \mid \boldsymbol{x}) = \sum_{m=1}^{M} \hat{w}_{KL,m} \hat{f}_m(y \mid \boldsymbol{x}).$$

The above method employs a strategy of selecting optimal weights by minimizing the loss function between the estimation of the conditional density function and its true value. The CDE loss function enables the model averaging estimate to be as close as possible to the actual conditional density function, and the KL divergence effectively reduces information loss. By combining conformal prediction, we conduct simulation experiments and demonstrate the effectiveness of this method in the practical application of photometric redshift estimation. First, the length of the prediction bands generated by HPD conformal prediction is still directly affected by the estimation quality. The cross-validation model averaging methods proposed in this paper can provide stable predictions for HPD conformal prediction. Using the HPD conformal prediction bands as an evaluation and comparison technique for the conditional density estimation is meaningful, as conformal prediction can provide prediction bands for model averaging estimates under finite samples. Both complement each other.

III. Conformal prediction based on model averaging

In the following section, we provide a brief to review the theoretical properties of conformal prediction, such as marginal validity, conditional validity, and asymptotic conditional validity. Then, we introduce the ICP and HPD conformal prediction with asymptotic conditional validity. Furthermore, we combine them with model averaging techniques, resulting in two prediction algorithms for model averaging estimators in the case of finite samples.

A. Review of conformal prediction

Point estimators usually contain relatively limited information, while interval estimators provide more comprehensive information about the range of possible values for the prediction. This helps us understand the concentration and dispersion of the prediction results. Conformal prediction is a type of method that calculates a prediction interval with exact coverage probability in the finite sample case. Typically, this method generates a prediction interval $C_\alpha(\boldsymbol{X}_{n+1})$ for the response Y_{n+1} simply based on the interchangeable assumption, the training samples $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$ and the covariates of the target sample $\boldsymbol{X}_{n+1}$ [40], [41]. This interchangeability only requires the order of sample data to be interchangeable (the joint probability distribution remains invariant after the order of the samples is permutated [42]). This provides a theoretical guarantee for the marginal coverage of the prediction interval [43], which is called marginal validity:

$$\mathbb{P}\left(Y_{n+1} \in C_\alpha\left(\boldsymbol{X}_{n+1}\right)\right) \geq 1-\alpha, \tag{8}$$

where $1-\alpha$ represents the corresponding confidence level.

In addition to studying the marginal validity of the prediction set as a whole, researchers intend to find theoretical guarantees that are stronger than Equation (8). Consequently, a guarantee called conditional validity is established:

$$\mathbb{P}\left(Y_{n+1} \in C_\alpha\left(\boldsymbol{X}_{n+1}\right) \mid \boldsymbol{X}_{n+1}\right) \geq 1-\alpha.$$

Unfortunately, we can only ensure conditional validity by posing strong assumptions regarding the joint distribution of (X, Y) [44], [45]. Given this limitation, Lei and Wasserman [45], Guan [46] and Barber et al. [47] studied local validity as an intermediate result, which is shown as follows:

$$\mathbb{P}\left(Y_{n+1} \in C_\alpha\left(\boldsymbol{X}_{n+1}\right) \mid \boldsymbol{X}_{n+1} \in A\right) \geq 1-\alpha,$$

where A is the local neighborhood where the target sample $\boldsymbol{X}_{n+1}$ falls. These methods use the training samples that fall into the same local neighborhood A to complete conformal prediction. The intention is to approximate conditional validity by achieving validity in a sufficiently small local neighborhood. However, when dealing with high-dimensional data, these methods often need to create larger local neighborhoods to contain enough training samples, which contradicts their original intention.

Furthermore, under weaker conditions, other researchers establish asymptotic conditional validity using different techniques, such as the quantile regression [48], [49], cumulative distribution function estimators [50], [51] and density estimators [52]. Note that the asymptotic conditional validity requires that there exists a set Λ_n satisfying $\mathbb{P}(\boldsymbol{X}_{n+1} \in \Lambda_n \mid \Lambda_n) = 1 - o_{\mathbb{P}}(1)$, such that

$$\inf_{\boldsymbol{X}_{n+1} \in \Lambda_n} \left|\mathbb{P}\left(Y_{n+1} \in C\left(\boldsymbol{X}_{n+1}\right) \mid \boldsymbol{X}_{n+1}\right) - (1-\alpha)\right| = o_{\mathbb{P}}(1).$$

B. Conformal prediction based on Mallows model averaging

We now describe how to construct the prediction interval that satisfy marginal validity for conformal prediction. We focus on the combination of inductive conformal prediction and Mallows model averaging.

The inductive conformal prediction first randomly partitions the sample $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$ into two subsets, namely the training set $\mathbb{D}$ and the calibration set $\mathbb{D}'$. Then, the model is trained taking advantage of samples from $\mathbb{D}$. The algorithm uses an arbitrary real-valued function $S\left(\boldsymbol{X}, Y, \mathbb{D}\right)$ as the conformity score to generate an effective prediction interval. The conformity score measures the degree of difference between the target sample and the samples in the subset $\mathbb{D}'$. A smaller score indicates that the sample is more consistent with the calibration set $\mathbb{D}'$. Hence, $S\left(\boldsymbol{X}, Y, \mathbb{D}\right)$ is also called a nonconformity score. If the conformity score is a good measurement of the degree of difference between the target sample and the samples in the subset $\mathbb{D}'$, then the prediction bands can be effective (small). Furthermore, after separately calculating $S\left(\boldsymbol{X}_{n+1}, y, \mathbb{D}\right)$ and $S\left(\boldsymbol{X}_i, Y_i, \mathbb{D}\right), i \in \mathbb{D}'$, we define

$$\begin{aligned}\pi(y) &= \frac{\mathbb{I}\left(S\left(\boldsymbol{X}_{n+1}, y, \mathbb{D}\right) \leq S\left(\boldsymbol{X}_{n+1}, y, \mathbb{D}\right)\right)}{1+\#\mathbb{D}'} \\ &\quad + \frac{\sum_{i \in \mathbb{D}'} \mathbb{I}\left(S\left(\boldsymbol{X}_{n+1}, y, \mathbb{D}\right) \leq S\left(\boldsymbol{X}_i, Y_i, \mathbb{D}\right)\right)}{1+\#\mathbb{D}'} \\ &= \frac{1+\sum_{i \in \mathbb{D}'} \mathbb{I}\left(S\left(\boldsymbol{X}_{n+1}, y, \mathbb{D}\right) \leq S\left(\boldsymbol{X}_i, Y_i, \mathbb{D}\right)\right)}{1+\#\mathbb{D}'}\end{aligned}$$

where $\mathbb{1}(\cdot)$ is an indicator function. Assuming that the sample is interchangeable, the scores $S\left(\boldsymbol{X}_{n+1}, y, \mathbb{D}\right)$ and $S\left(\boldsymbol{X}_i, Y_i, \mathbb{D}\right), i \in \mathbb{D}'$ are symmetric. When $Y_{n+1} = y$ holds, it follows that $\pi(Y_{n+1})$ obeys a uniform distribution, which ensures the marginal validity. By traversing the grid points of y values, we can construct prediction bands as

$$\left\{y : S\left(\boldsymbol{X}_i, y, \mathbb{D}\right) \geq U_{\lfloor \alpha \rfloor}\right\},$$

where $U_{\lfloor\alpha\rfloor}$ is the upper α-quantile of the conformity score $S\left(\boldsymbol{X}_i, Y_i, \mathbb{D}\right), i \in \mathbb{D}'$.

In many cases, we only need to use the $\mathcal{L}^1$ distance as a conformity score, which is defined as

$$S\left(\boldsymbol{X}_i, Y_i, \mathbb{D}\right) = \left|Y_i - \hat{\mu}\left(\boldsymbol{X}_i\right)\right|. \tag{9}$$

When the conformity score $S\left(\boldsymbol{X}_i, Y_i, \mathbb{D}\right)$ is monotonically transformed, the prediction interval generated by conformal prediction remains unchanged. For example, if S is non-negative, replacing S with S^2 makes no difference. Therefore, the choice of distance measure is relatively unimportant. The crucial step in determining the conformity score $S\left(\boldsymbol{X}_i, Y_i, \mathbb{D}\right)$ is the choice of point estimator $\hat{\mu}\left(\boldsymbol{X}_i\right)$ [42], [45]. First, the length of the prediction interval generated by conformal prediction still depends on the quality of the estimation. The MMA method, as the foundation of optimal model averaging, provides asymptotically optimal estimation, which can provide stable predictions for conformal prediction. Using the conformal prediction interval as an assessment and benchmarking method for the regression function estimator is meaningful. Conformal prediction can provide a prediction interval for model averaging estimates under finite samples. Both aspects complement each other.

The MMA method considers the linear model

$$y = \mu + e = \boldsymbol{X}\boldsymbol{\beta} + e,$$

where $\boldsymbol{\beta}$ is the unknown regression coefficient and e is the stochastic error term with a mean of 0 and a variance of σ^2. By calculating the regression residuals $\hat{e}_1, \ldots, \hat{e}_M$ in each candidate model and the variance estimate $\hat{\sigma}^2 = (n-p)^{-1}\hat{e}_M'\hat{e}_M$ in the full model, we can obtain the weight selection criterion of MMA method:

$$C(\boldsymbol{w}) = \boldsymbol{w}'\hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}\boldsymbol{w} + 2\hat{\sigma}^2\boldsymbol{w}'\boldsymbol{P},$$

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathcal{W}}{\arg\min}\, C(\boldsymbol{w}),$$

where $\hat{\boldsymbol{E}} = (\hat{e}_1, \ldots, \hat{e}_M)$ and $\boldsymbol{P} = (p_1, \ldots, p_M)^{\mathrm{T}}$ are the dimensions of the regression coefficients in each candidate model.

We take MMA method as an example to combine ICP method with the model averaging method, as shown in Algorithm 3.1.

Algorithm 3.1 ICP-MMA

Input Confidence level $1-\alpha \in (0,1)$, training samples $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$, target covariates $\boldsymbol{X}_{n+1}$.

Output Prediction bands $C_\alpha\left(\boldsymbol{X}_{n+1}\right)$ for the target response variable Y_{n+1}.

Step 1 Divide the sample into two subsets, $\mathbb{D}$ and $\mathbb{D}'$, at random.

Step 2 Train $\hat{\mu}_{\hat{\boldsymbol{w}}}(y \mid \boldsymbol{x})$ using samples from the subset $\mathbb{D}$:

$$\hat{\mu}_{\hat{\boldsymbol{w}}}(\boldsymbol{x}) = \sum_{m=1}^{M} \hat{w}_m \hat{\mu}_m(\boldsymbol{x}).$$

Step 3 Calculate the conformity score $S\left(\boldsymbol{X}_i, Y_i, \mathbb{D}\right)$ for each $i \in \mathbb{D}'$ using Equation (9). Then, obtain the upper α quantile $U_{\lfloor\alpha\rfloor}$.

Step 4 For the given $\boldsymbol{X}_{n+1}$, traverse the grid points of y values, and provide the prediction bands

$$\left\{y : S\left(\boldsymbol{X}_{n+1}, y, \mathbb{D}\right) \geq U_{\lfloor\alpha\rfloor}\right\}.$$

This can be directly represented as a prediction interval

$$C_\alpha\left(\boldsymbol{X}_{n+1}\right) = [\hat{\mu}(\boldsymbol{X}_{n+1}) - U_{\lfloor\alpha\rfloor}, \hat{\mu}(\boldsymbol{X}_{n+1}) + U_{\lfloor\alpha\rfloor}]. \tag{10}$$

C. Conformal prediction based on cross-validation model averaging

When evaluating the validity of conformal prediction, we also pay attention to the length of the prediction interval. When the conditional density function presents multimodal characteristics, conformal prediction methods should generate prediction bands rather than a prediction interval. Izbicki et al. [53] proposed HPD conformal prediction, based on inductive conformal prediction and the estimation of the conditional density function, addresses the aforementioned issues.

The basic idea of HPD conformal prediction can be summarized as follows: Let the conformity score $H(Z \mid \boldsymbol{X})$ be the conditional cumulative distribution function of the random variable $Z = f(Y \mid \boldsymbol{X})$ given $\boldsymbol{X}$. Then, the conditional distribution of $H(Z|\boldsymbol{X})$ given $\boldsymbol{X}$ obeys a uniform distribution that is independent of $\boldsymbol{X}$. If $\hat{f}(Y \mid \boldsymbol{X})$ is sufficiently close to $f(Y \mid \boldsymbol{X})$, then $H(\hat{f}(Y \mid \boldsymbol{X}) \mid \boldsymbol{X})$ is adequately approximate to a uniform distribution independent of $\boldsymbol{X}$. Finally, as long as the order of sample data is interchangeable, the conformity score $H(\hat{f}(Y \mid \boldsymbol{X}) \mid \boldsymbol{X})$ can be guaranteed to be exchangeable, which ensures validity. In addition to possessing marginal validity and asymptotic conditional validity, this method also ensures theoretical convergence to the highest predictive density set:

$$\mathbb{P}\left(Y_{n+1} \in C_\alpha^*\left(\boldsymbol{X}_{n+1}\right) \Delta C_\alpha\left(\boldsymbol{X}_{n+1}\right)\right) = o(1),$$

where $C_\alpha^*(\boldsymbol{x})$ represents the highest predictive density set, and Δ represents the symmetric difference operation.

By combining the HPD conformal prediction with the two previously mentioned model averaging methods, we obtain the prediction bands corresponding to the model averaging estimation of the conditional density function, as shown in Algorithm 3.2.

Algorithm 3.2 HPDCP-MA

Input Confidence level $1-\alpha \in (0,1)$, Training samples $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$, target covariates $\boldsymbol{X}_{n+1}$.

Output Prediction bands $C_\alpha\left(\boldsymbol{X}_{n+1}\right)$ for the target response variable Y_{n+1}.

Step 1 Divide the sample into two subsets, $\mathbb{D}$ and $\mathbb{D}'$, at random.

Step 2 Train $\hat{f}_{\hat{\boldsymbol{w}}}(y \mid \boldsymbol{x})$ using samples from the subset $\mathbb{D}$.

Step 3 Using samples from the subset $\mathbb{D}'$, calculate the upper α quantile $U_{\lfloor\alpha\rfloor}$ of the conformity score

$$\begin{aligned}&\{\hat{H}\left(\hat{f}_{\hat{\boldsymbol{w}}}\left(Y_i \mid \boldsymbol{X}_i\right) \mid \boldsymbol{X}_i\right)\\ &= \int_{\left\{y:\hat{f}_{\hat{\boldsymbol{w}}}(y|\boldsymbol{X}_i)\leq\hat{f}_{\hat{\boldsymbol{w}}}(Y_i|\boldsymbol{X}_i)\right\}} \hat{f}_{\hat{\boldsymbol{w}}}\left(y \mid \boldsymbol{X}_i\right) dy\}.\end{aligned} \tag{11}$$

Step 4 For the given $\boldsymbol{X}_{n+1}$, traverse the grid points of y values, and provide the prediction bands

$$\{y : \hat{H}\left(\hat{f}_{\hat{\boldsymbol{w}}}\left(y \mid \boldsymbol{X}_{n+1}\right) \mid \boldsymbol{X}_{n+1}\right) \geq U_{\lfloor\alpha\rfloor}\}.$$

IV. Simulation

Throughout this section, we first study the predictive performance of the ICP-MMA algorithm under different sample sizes and different population R^2. Then, we investigate the prediction bands derived from the HPDCP-MA algorithm under two different model settings. In both settings, we select the simplest model and the full model from all candidate models, along with the Equal-weighted model averaging (EWMA) method as alternative methods, and compare them with the KLMA and CDEMA methods.

A. ICP-MMA

The first simulation setting follows the setup of Hurvich and Tsai [54] and adds an endogenous variable that is not included in all candidate models. Therefore, the following outlines the data generating process:

$$y = \boldsymbol{X}\boldsymbol{\beta} + e,$$

where

$$\boldsymbol{X} \sim N\left(\mathbf{0}_{5\times 1}, \begin{pmatrix} 1 & 0 & 0 & 0 & 0.3 \\ 0 & 1 & 0 & 0 & 0.3 \\ 0 & 0 & 1 & 0 & 0.3 \\ 0 & 0 & 0 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 1 \end{pmatrix}\right),$$

and $e \sim N(0,1)$.

Please be aware that the last variable in $\boldsymbol{X}$ is absent from any candidate model. Thus, by setting the value of the last coefficient to be zero or non-zero, we can determine if there occur correctly specified models in the candidate model set. Each candidate model is established by building a connection between the response variable Y and a different subset of the covariates $\boldsymbol{X}$. Eventually, we obtain 15 non-nested candidate models.

As the benchmark method, the LM represents the full model without an intercept, whose prediction interval is determined based on the t-statistics. The ICP-LM method calculates the prediction interval by fitting the linear model and utilizing inductive conformal prediction. Stock and Watson [55] found that the ideal estimated weights may fall short compared to the EWMA method in terms of mean square prediction error. This phenomenon is referred to as the "forecast combination problem". Moreover, Smith and Wallis [56] confirmed that when the optimal weights are nearly identical to equal weights, EWMA can outperform the forecast combination with estimated weights. Therefore, it is natural that we include the EWMA method by way of comparison.

We separately generate $n = 50, 100, 200$ samples, of which 50% are utilized to estimate the regression coefficients using the least squares method. The remaining 50% are used to calculate the prediction interval with significance level $\alpha = 0.1$ according to Equation (10). When $\beta = (1,2,3,0,0)^{\mathrm{T}}$, the population $R^2 = 14/\left(14+\sigma^2\right)$ is controlled by the parameter σ^2. In this study, we control the population R^2 at 0.1 to 0.9 grid points by varying the variance σ^2. We repeat the generation of 100 random number seeds and conduct 100 simulation experiments. In each trial, we provide 100 new test data with conformal prediction bands and record the average coverage probability and length. The results are shown in Figures 1 – 6.

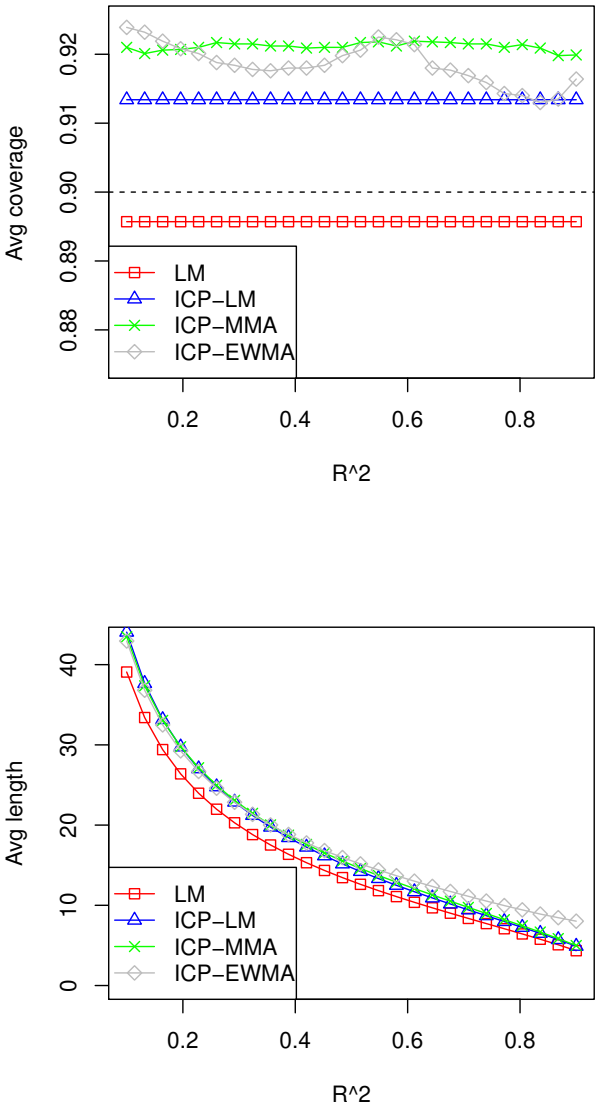
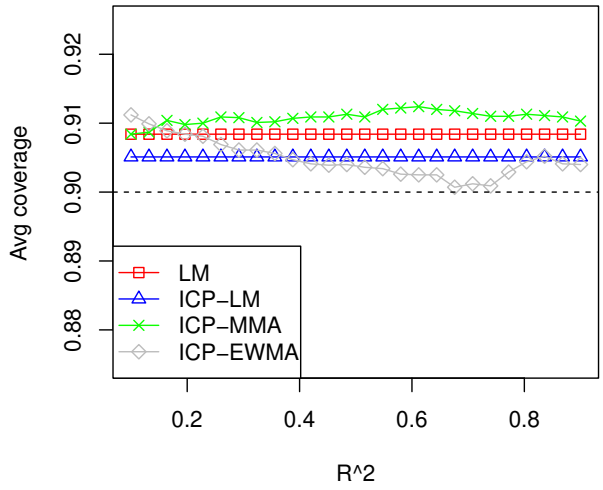


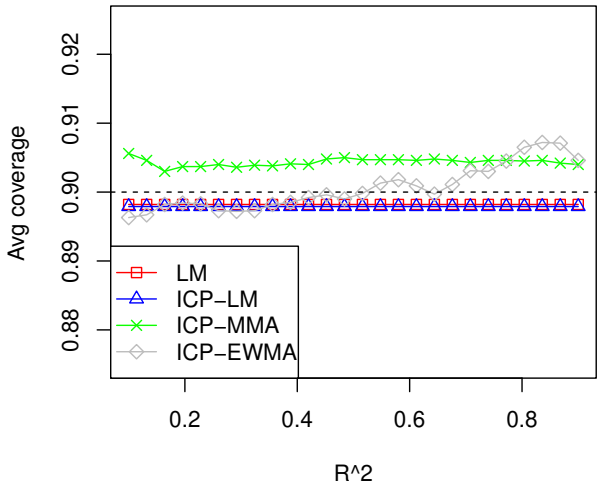
Fig. 1. Outcomes of the average coverage probability and average length of the conformal prediction interval for $n = 50$, $\beta = (1,2,3,0,0)^{\mathrm{T}}$, with varying R^2 values.

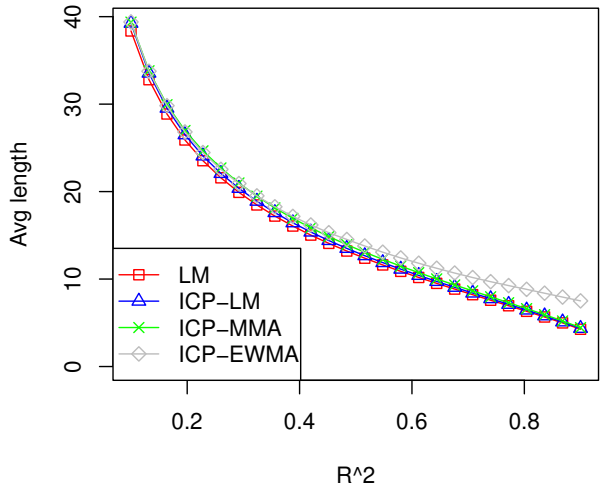
In simulation experiments, the average length performance of the ICP-EWMA method is poor. As the ICP-EWMA method always incorrectly assigns the same weight to the worst and the best models, a shortcoming that is amplified as the population R^2 increases. Additionally, the ICP-MMA method, which approximates other methods in average length, outperforms other methods in average coverage probability, reflecting the stability of this method.

B. HPDCP-MA

1) The fixed number of covariates: Following the setup from the previous section, we set $\sigma^2 = 1$ and generate 1000 samples, of which 50% are utilized to estimate the conditional density function using the FlexCode method. The remaining 50% are used to calculate the conformity score according to Equation (11). Additionally, we set the significance level $\alpha = 0.1$ and the number of folds $J = 10$.







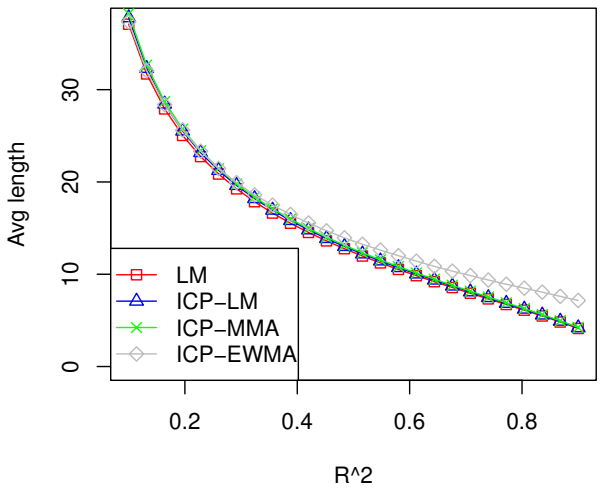


Fig. 2. Outcomes of the average coverage probability and average length of the conformal prediction interval for $n = 100$, $\beta = (1, 2, 3, 0, 0)^{\mathrm{T}}$, with varying R^2 values.

Fig. 3. Outcomes of the average coverage probability and average length of the conformal prediction interval for $n = 200$, $\beta = (1, 2, 3, 0, 0)^{\mathrm{T}}$, with varying R^2 values.

Each candidate model is established by connecting $\boldsymbol{X}$ to Y through different subsets of the covariates, ultimately resulting in 15 non-nested candidate models.

As for the comparison methods, the simplest model represents a candidate model that only includes the coefficients of the variables that are not zero in the true data generation process (except for the unaccounted endogenous variable), while the full model consists of all observed covariates. The EWMA method assigns the equal weight to each candidate model. We repeat the generation of 100 random number seeds and conduct 100 simulation experiments. In each trial, we provide 100 new test data points with conformal prediction bands. The results are shown in Tables I – IX.

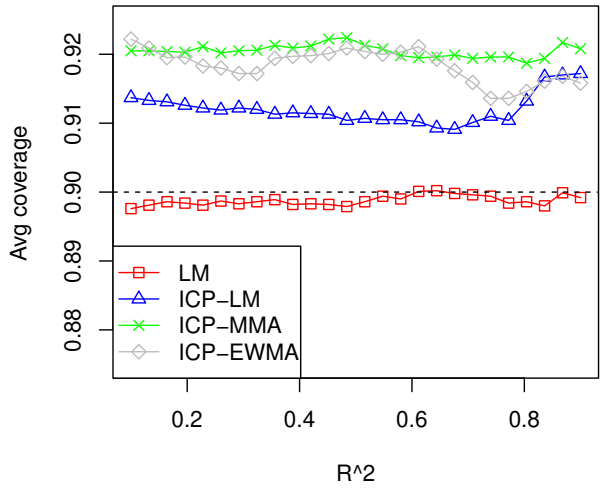
We observe from the result that both CDEMA and KLMA methods exhibit higher average coverage probability compared to the simplest model, while the average length is also close to that of the simplest model. The EWMA approach, which serves as the benchmark in this setting, is relatively poor. Compared to the KLMA method, the CDEMA method

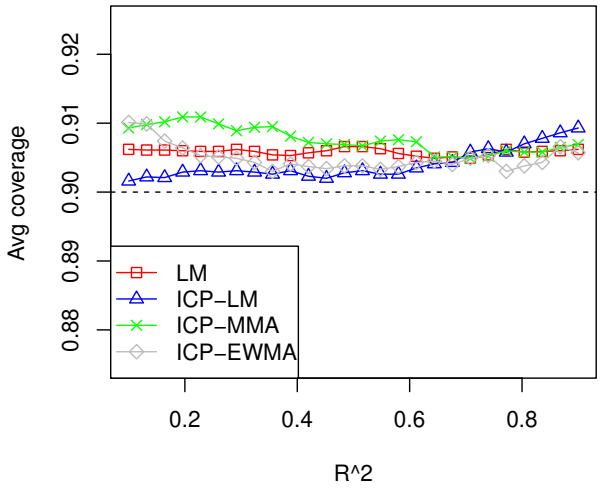
TABLE I
$\beta = (1, 0, 0, 0, 0)^{\mathrm{T}}$

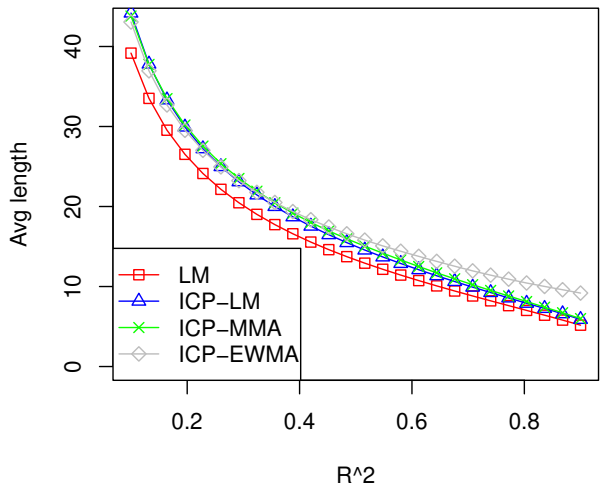
	Simp	Full	EWMA	CDEMA	KLMA
Coverage	0.895	0.894	0.900	0.898	0.896
Length	3.869	3.914	3.573	3.430	3.341

TABLE II
$\beta = (1, 2, 0, 0, 0)^{\mathrm{T}}$

	Simp	Full	EWMA	CDEMA	KLMA
Coverage	0.892	0.888	0.900	0.895	0.894
Length	4.306	3.848	4.023	3.575	3.610







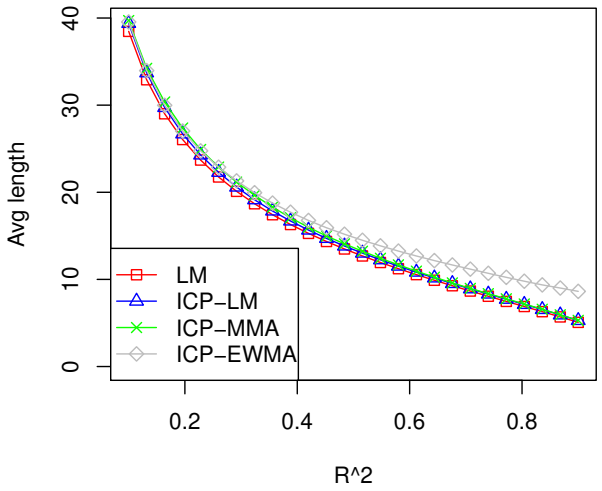


Fig. 4. Outcomes of the average coverage probability and average length of the conformal prediction interval for $n = 50$, $\beta = (1, 2, 3, 0, 1)^{\mathrm{T}}$, with varying R^2 values.

Fig. 5. Outcomes of the average coverage probability and average length of the conformal prediction interval for $n = 100$, $\beta = (1, 2, 3, 0, 1)^{\mathrm{T}}$, with varying R^2 values.

TABLE III
$\beta = (1, 2, 3, 0, 0)^{\mathrm{T}}$

	Simp	Full	EWMA	CDEMA	KLMA
Coverage	0.881	0.890	0.898	0.897	0.896
Length	4.039	4.351	4.868	4.005	4.048

TABLE IV
$\beta = (1, 2, 3, 4, 0)^{\mathrm{T}}$

	Simp	Full	EWMA	CDEMA	KLMA
Coverage	0.887	0.887	0.899	0.898	0.899
Length	4.854	4.854	6.547	4.802	4.887

shows a slight advantage.

TABLE V
$\beta = (1, 0, 0, 0, 1)^{\mathrm{T}}$

	Simp	Full	EWMA	CDEMA	KLMA
Coverage	0.901	0.891	0.899	0.899	0.898
Length	5.519	5.090	4.663	4.484	4.508

TABLE VI
$\beta = (1, 2, 0, 0, 1)^{\mathrm{T}}$

	Simp	Full	EWMA	CDEMA	KLMA
Coverage	0.891	0.882	0.894	0.895	0.895
Length	5.269	4.901	5.215	4.639	4.701

2) Synthetic procedure: We adopt another simulation setting from Lei and Wasserman [45] and Izbicki et al. [53].

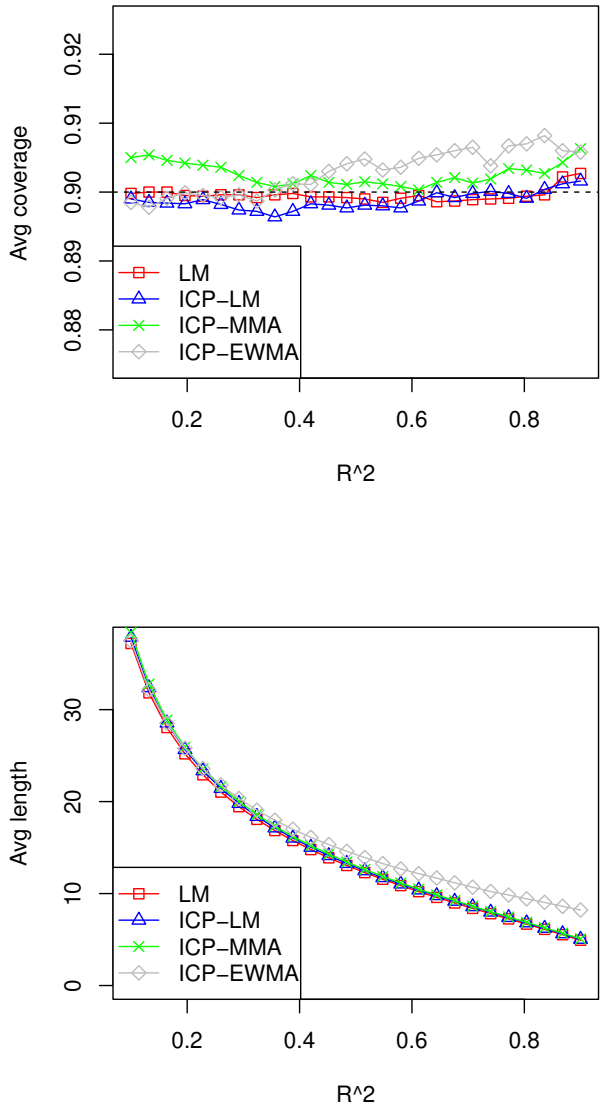


Fig. 6. Outcomes of the average coverage probability and average length of the conformal prediction interval for $n = 200$, $\beta = (1, 2, 3, 0, 1)^{\mathrm{T}}$, with varying R^2 values.

TABLE VII
$\beta = (1, 2, 3, 0, 1)^{\mathrm{T}}$

	Simp	Full	EWMA	CDEMA	KLMA
Coverage	0.884	0.884	0.894	0.899	0.895
Length	5.184	5.323	6.165	5.115	5.283

TABLE VIII
$\beta = (1, 2, 3, 4, 1)^{\mathrm{T}}$

	Simp	Full	EWMA	CDEMA	KLMA
Coverage	0.889	0.889	0.892	0.896	0.895
Length	5.902	5.902	7.799	5.998	5.984

The details about the data generating process are as follows:

$$\boldsymbol{X} = (X_1, \ldots, X_4),$$

TABLE IX
$\beta = (1, 2, 3, 4, 2)^{\mathrm{T}}$

	Simp	Full	EWMA	CDEMA	KLMA
Coverage	0.885	0.885	0.897	0.892	0.893
Length	7.942	7.942	10.108	8.048	8.063

TABLE X
SYNTHETIC PROCEDURE

	Simp	Full	EWMA	CDEMA	KLMA
Coverage	0.854	0.912	0.965	0.886	0.888
Length	5.967	7.282	5.881	5.479	5.546

$$\begin{aligned} Y \mid \boldsymbol{X} = \boldsymbol{x} \sim & 0.5N\left(f(\boldsymbol{x}) - g(\boldsymbol{x}), \sigma^2(\boldsymbol{x})\right) \\ & + 0.5N\left(f(\boldsymbol{x}) + g(\boldsymbol{x}), \sigma^2(\boldsymbol{x})\right), \end{aligned}$$

where

$$\begin{aligned} X_i &\overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[-1.5, 1.5], \\ f(\boldsymbol{x}) &= (x_1 - 1)^2(x_1 + 1), \\ g(\boldsymbol{x}) &= 2\sqrt{x_1 + 0.5}\mathbb{1}(x_1 \geqslant -0.5), \\ \sigma^2(\boldsymbol{x}) &= 1/4 + |x_1|. \end{aligned}$$

When $x_1 \leqslant -0.5$, $(Y \mid \boldsymbol{X} = \boldsymbol{x})$ is a Gaussian distribution with expectation $f(\boldsymbol{x})$ and variance $\sigma^2(\boldsymbol{x})$. However, when $x_1 \geqslant -0.5$, the distribution of $(Y \mid \boldsymbol{X} = \boldsymbol{x})$ is a mixture of two Gaussian distributions. When x_1 increases, the overlap between the two Gaussian distributions decreases. Other simulation experiments remain the same as the former experiment, and the results are shown in Table X.

Compared to the previous simulation settings, CDEMA and KLMA methods have better average coverage probability as well as shorter average lengths than the simplest model. Furthermore, the CDEMA method shows a slight advantage compared to the KLMA method.

V. PHOTOMETRIC REDSHIFT ESTIMATION

In this section, we first introduce the measurement methods for photometric and spectroscopic data, thereby introducing the dataset used in this study. Then we demonstrate the superiority of the model averaging estimation of the conditional density function and the prediction algorithm through actual data analysis.

A. Background introduction

Redshift is a concept in physics and astronomy, describing the phenomenon of frequency decrease in electromagnetic radiation emitted by an object due to its movement away from the observer. The change in frequency gives rise to a change in color, with spectral lines shifting. By comparing the positions of absorption lines under laboratory light sources, we can understand the movement of the star relative to the Earth. Therefore, redshift is a crucial role in inferring parameters of cosmological models.

Currently, astronomical data can be categorized into two types: photometric and spectroscopic. Although spectroscopy can accurately estimate redshift, more than 99% of galaxy observations currently rely on photometric techniques due

TABLE XI
PHOTOMETRIC REDSHIFT ESTIMATION

	Full	EWMA	CDEMA	KLMA
Coverage	0.897	0.893	0.896	0.895
Length	0.150	0.153	0.140	0.144

to cost and time constraints. Photometry is an efficient low-resolution measurement method [57], which roughly records the radiation of an astronomical object through $5-10$ broadband filters. Only the covariates $\boldsymbol{X}$ are available, as the exact measurement of redshift Z is not obtainable. However, in spectroscopy, both the covariates $\boldsymbol{X}$ and redshift Z can be measured with negligible error.

Photometric redshift estimation is achieved by using galaxy samples with confirmed redshifts from spectroscopic data. The galaxy's redshift Z can be inferred from the observed photometric characteristics $\boldsymbol{X}$. However, due to the lack of spectroscopic data for galaxies with more extreme colors and weaker brightness in the SDSS-DR12 [58], we use the Happy A and B datasets [59] to construct conformal prediction bands for redshift. These two datasets provide more comprehensive spectroscopic data, aimed at evaluating the applicability of photometric redshift estimation methods in more realistic scenarios. In this case, all photometric data comes from DR12, while the spectroscopic data is extended from the DR12 photometric data by Bayesian cross-matching with other sources, respectively containing 74950 galaxies and 74900 galaxies.

B. Experimental results

This paper uses the same covariates $\boldsymbol{X}$ as Sheldon et al. [4], namely r-magnitude and four color magnitudes. In this case, the simplest model is also the full model. We use the samples in Happy A as the training set, taking out 1000 samples each time, of which 50% are employed in estimating the conditional density function using the FlexCode method. The remaining 50% are used to calculate the conformity score according to Equation (11) and set the significance level $\alpha = 0.1$. Finally, we evaluate a total of 31 non-nested candidate models and set the number of cross-validation folds $J = 5$ for the weight choice in model averaging. We repeat the generation of 100 random number seeds, conducting 100 simulation experiments. In each trial, we provide conformal prediction bands for 100 samples from the test set Happy B. The results are shown in Table XI.

The results show that, despite all methods achieving similar average coverage probability close to the nominal level, the two model averaging methods proposed in this paper can achieve smaller average length. It is noteworthy that, by using HPD conformal prediction with conditional density estimation, the prediction bands can ensure conditional validity for a single new sample. Given that the density function of dark galaxies redshift usually has multimodal characteristics [33], [60], [61], [62], and the results of conformal prediction are highly sensitive to density estimation, this further confirms the superiority of the cross-validation model averaging method in this paper.

VI. CONCLUSION

Throughout this study, we develop two methods for model averaging based on KL divergence and CDE loss, taking FlexCode as an example. The cross-validation model averaging method considers different combinations of covariates as candidate models to address the issue of information loss in variable selection, thereby reducing estimation risk and enhancing estimation stability. To our understanding, this paper is the inaugural work to develop a model averaging method using cross-validation to select weights based on KL divergence and CDE loss in the field of conditional density estimation. Additionally, due to the limited information provided by point estimation, we combine model averaging estimation with HPD conformal prediction for simulation studies for the first time. We also conduct real data analysis for photometric redshift estimation, and discover that the conformal prediction bands generated by the two methods significantly outperform those produced by several other methods. This provides more support for using our proposed model averaging methods for conditional density estimation in real-world examples.

While we have demonstrated the superiority of our methods through simulation experiments, it remains challenging to prove the asymptotic optimality of the proposed model averaging estimators. Moreover, further studies on how to shorten the length of prediction bands are of great importance.

REFERENCES

[1] M. Carrasco Kind and R. J. Brunner, "Tpz: Photometric redshift pdfs and ancillary information by using prediction trees and random forests," *Monthly Notices of the Royal Astronomical Society*, vol. 432, no. 2, pp. 1483–1501, 2013.

[2] P. E. Freeman, R. Izbicki, and A. B. Lee, "A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting," *Monthly Notices of the Royal Astronomical Society*, vol. 468, no. 4, pp. 4556–4565, 2017.

[3] R. Izbicki, A. B. Lee, and P. E. Freeman, "Photo-z estimation: An example of nonparametric conditional density estimation under selection bias," *Annals of Applied Statistics*, vol. 11, no. 2, p. 698, 2017.

[4] E. S. Sheldon, C. E. Cunha, R. Mandelbaum, J. Brinkmann, and B. A. Weaver, "Photometric redshift probability distributions for galaxies in the sdss dr8," *The Astrophysical Journal Supplement Series*, vol. 201, no. 2, p. 32, 2012.

[5] M. M. Rau, S. Seitz, F. Brimioulle, E. Frank, O. Friedrich, D. Gruen, and B. Hoyle, "Accurate photometric redshift probability density estimation–method comparison and application," *Monthly Notices of the Royal Astronomical Society*, vol. 452, no. 4, pp. 3710–3725, 2015.

[6] R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald, "Estimating and visualizing conditional densities," *Journal of Computational and Graphical Statistics*, vol. 5, no. 4, pp. 315–336, 1996.

[7] T. Ichimura and D. Fukuda, "A fast algorithm for computing least-squares cross-validations for nonparametric conditional kernel density functions," *Computational Statistics & Data Analysis*, vol. 54, no. 12, pp. 3404–3410, 2010.

[8] J. Fan, Q. Yao, and H. Tong, "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems," *Biometrika*, vol. 83, no. 1, pp. 189–206, 1996.

[9] M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara, "Conditional density estimation via least-squares density ratio estimation," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. PMLR, 13–15 May 2010, pp. 781–788.

[10] I. Takeuchi, K. Nomura, and T. Kanamori, "Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression," *Neural Computation*, vol. 21, no. 2, pp. 533–559, 2009.

[11] P. Hall, J. Racine, and Q. Li, "Cross-validation and the estimation of conditional probability densities," *Journal of the American Statistical Association*, vol. 99, no. 468, pp. 1015–1026, 2004.

[12] S. Efromovich, "Dimension reduction and adaptation in conditional density estimation," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 761–774, 2010.

[13] R. Izbicki and A. B. Lee, "Converting high-dimensional regression to high-dimensional conditional density estimation," *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 2800–2831, 2017.

[14] H. Akaike, "Maximum likelihood identification of gaussian autoregressive moving average models," *Biometrika*, vol. 60, no. 2, pp. 255–265, 1973.

[15] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[16] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," *Technometrics*, vol. 16, no. 1, pp. 125–127, 1974.

[17] X. Zhang and G. Zou, "Model averaging method and its application in forecast," *Statistical Research*, vol. 28, no. 06, pp. 97–102, 2011.

[18] N. L. Hjort and G. Claeskens, "Frequentist model average estimators," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 879–899, 2003.

[19] G. Leung and A. R. Barron, "Information theory and mixing least-squares regressions," *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3396–3410, 2006.

[20] Z. Yuan and Y. Yang, "Combining linear regression models: When and how?" *Journal of the American Statistical Association*, vol. 100, no. 472, pp. 1202–1214, 2005.

[21] J. M. Bates and C. W. Granger, "The combination of forecasts," *Journal of the Operational Research Society*, vol. 20, no. 4, pp. 451–468, 1969.

[22] B. E. Hansen, "Model averaging, asymptotic risk, and regressor groups," *Quantitative Economics*, vol. 5, no. 3, pp. 495–530, 2014.

[23] G. Claeskens and N. L. Hjort, *Model selection and model averaging*. New York: Cambridge University Press, 2008.

[24] E. Moral-Benito, "Model averaging in economics: An overview," *Journal of Economic Surveys*, vol. 29, no. 1, pp. 46–75, 2015.

[25] D. Fletcher, *Model averaging*. Berlin: Springer, 2018.

[26] M. F. J. Steel, "Model averaging and its use in economics," *Journal of Economic Literature*, vol. 58, no. 3, pp. 644–719, 2020.

[27] S. T. Buckland, K. P. Burnham, and N. H. Augustin, "Model selection: An integral part of inference," *Biometrics*, vol. 53, no. 2, pp. 603–618, 1997.

[28] B. E. Hansen, "Least squares model averaging," *Econometrica*, vol. 75, no. 4, pp. 1175–1189, 2007.

[29] A. T. Wan, X. Zhang, and G. Zou, "Least squares model averaging by mallows criterion," *Journal of Econometrics*, vol. 156, no. 2, pp. 277–283, 2010.

[30] B. E. Hansen and J. S. Racine, "Jackknife model averaging," *Journal of Econometrics*, vol. 167, no. 1, pp. 38–46, 2012.

[31] P. Lin, J. Liao, Z. Lu, K. You, and G. Zou, "Optimal averaging estimation for density functions," 2024, working paper.

[32] K. Takeuchi, "Distribution of information statistics and validity criteria of models," *Mathematical Sciences*, vol. 153, pp. 12–18, 1976.

[33] N. Dalmasso, T. Pospisil, A. Lee, R. Izbicki, P. Freeman, and A. Malz, "Conditional density estimation tools in python and r with applications to photometric redshifts and likelihood-free cosmological inference," *Astronomy and Computing*, vol. 30, p. 100362, 2020.

[34] R. Izbicki and A. B. Lee, "Nonparametric conditional density estimation in a high-dimensional regression setting," *Journal of Computational and Graphical Statistics*, vol. 25, no. 4, pp. 1297–1316, 2016.

[35] J.-M. Lueckmann, P. J. Goncalves, G. Bassetto, K. Öcal, M. Nonnenmacher, and J. H. Macke, "Flexible statistical inference for mechanistic models of neural dynamics," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[36] G. Papamakarios, T. Pavlakou, and I. Murray, "Masked autoregressive flow for density estimation," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[37] T. Pospisil and A. B. Lee, "Rfcde: Random forests for conditional density estimation," *arXiv:1804.05753*, 2018.

[38] Q. Li and J. S. Racine, *Nonparametric econometrics: Theory and practice*. Princeton: Princeton University Press, 2023.

[39] X. Zhang and C.-A. Liu, "Model averaging prediction by k-fold cross-validation," *Journal of Econometrics*, vol. 235, no. 1, pp. 280–301, 2023.

[40] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. New York: Springer, 2005.

[41] V. Vovk, I. Nouretdinov, and A. Gammerman, "On-line predictive linear regression," *The Annals of Statistics*, vol. 37, no. 3, pp. 1566–1590, 2009.

[42] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, no. 12, pp. 371–421, 2008.

[43] O. Kallenberg, *Probabilistic symmetries and invariance principles*. New York: Springer, 2005.

[44] V. Vovk, "Conditional validity of inductive conformal predictors," in *Proceedings of the Asian Conference on Machine Learning*. PMLR, 04–06 Nov 2012, pp. 475–490.

[45] J. Lei and L. Wasserman, "Distribution-free prediction bands for non-parametric regression," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 76, no. 1, pp. 71–96, 2014.

[46] L. Guan, "Localized conformal prediction: A generalized inference framework for conformal prediction," *Biometrika*, vol. 110, no. 1, pp. 33–50, 2023.

[47] R. Foygel Barber, E. Candès, A. Ramdas, and R. J. Tibshirani, "The limits of distribution-free conditional predictive inference," *Information and Inference: A Journal of the IMA*, vol. 10, no. 2, pp. 455–482, 2021.

[48] M. Sesia and E. Candès, "A comparison of some conformal quantile regression methods," *Stat*, vol. 9, no. 1, p. e261, 2020.

[49] Y. Romano, E. Patterson, and E. Candès, "Conformalized quantile regression," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[50] V. Chernozhukov, K. Wüthrich, and Y. Zhu, "Distributional conformal prediction," *Proceedings of the National Academy of Sciences*, vol. 118, no. 48, p. e2107794118, 2021.

[51] R. Izbicki, G. Shimizu, and R. Stern, "Flexible distribution-free conditional predictive bands using density estimators," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR, 26–28 Aug 2020, pp. 3068–3077.

[52] M. Sesia and Y. Romano, "Conformal prediction using conditional histograms," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6304–6315, 2021.

[53] R. Izbicki, G. Shimizu, and R. B. Stern, "Cd-split and hpd-split: Efficient conformal regions in high dimensions," *Journal of Machine Learning Research*, vol. 23, no. 87, pp. 1–32, 2022.

[54] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.

[55] J. H. Stock and M. W. Watson, "Combination forecasts of output growth in a seven-country data set," *Journal of Forecasting*, vol. 23, no. 6, pp. 405–430, 2004.

[56] J. Smith and K. F. Wallis, "A simple explanation of the forecast combination puzzle," *Oxford Bulletin of Economics and Statistics*, vol. 71, no. 3, pp. 331–355, 2009.

[57] H. Hildebrandt, C. Wolf, and N. Benítez, "A blind test of photometric redshifts on ground-based data," *Astronomy & Astrophysics*, vol. 480, no. 3, pp. 703–714, 2008.

[58] S. Alam, F. D. Albareti, C. A. Prieto, F. Anders, S. F. Anderson, T. Anderton, B. H. Andrews, E. Armengaud, É. Aubourg, S. Bailey *et al.*, "The eleventh and twelfth data releases of the sloan digital sky survey: final data from sdss-iii," *The Astrophysical Journal Supplement Series*, vol. 219, no. 1, p. 12, 2015.

[59] R. Beck, C.-A. Lin, E. Ishida, F. Gieseke, R. de Souza, M. Costa-Duarte, M. Hattab, A. Krone-Martins, and C. Collaboration, "On the realistic validation of photometric redshifts," *Monthly Notices of the Royal Astronomical Society*, vol. 468, no. 4, pp. 4323–4339, 2017.

[60] D. Wittman, "What lies beneath: Using p(z) to reduce systematic photometric redshift errors," *The Astrophysical Journal*, vol. 700, no. 2, p. L174, 2009.

[61] S. D. Kügler, N. Gianniotis, and K. L. Polsterer, "A spectral model for multimodal redshift estimation," in *2016 IEEE Symposium Series on Computational Intelligence*. IEEE, 13–15 May 2016, pp. 1–8.

[62] K. L. Polsterer, "Dealing with uncertain multimodal photometric redshift estimations," *Proceedings of the International Astronomical Union*, vol. 12, no. S325, pp. 156–165, 2016.