Model Averaging Prediction by K-Fold Cross-Validation

Xinyu Zhang[a,b] and Chu-An Liu[c,*]

April 15, 2022

Abstract

This paper considers the model averaging prediction in a quasi-likelihood framework that allows for parameter uncertainty and model misspecification. We propose an averaging prediction that selects the data-driven weights by minimizing a K-fold cross-validation. We provide two theoretical justifications for the proposed method. First, when all candidate models are misspecified, we show that the proposed averaging prediction using K-fold cross-validation weights is asymptotically optimal in the sense of achieving the lowest possible prediction risk. Second, when the model set includes correctly specified models, we demonstrate that the proposed K-fold cross-validation asymptotically assigns all weights to the correctly specified models. Monte Carlo simulations show that the proposed averaging prediction achieves lower empirical risk than other existing model averaging methods. As an empirical illustration, the proposed method is applied to credit card default prediction.

Keywords: Asymptotic optimality, Cross-validation, Model averaging, Weight convergence.

JEL Classification: C51, C52

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences. Email: xinyu@amss.ac.cn.

[b] International Institute of Finance, School of Management, University of Science and Technology of China.

[c] Institute of Economics, Academia Sinica. Email: caliu@econ.sinica.edu.tw.

[*] Corresponding author: Chu-An Liu, Institute of Economics, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei, 115 Taiwan. Email: caliu@econ.sinica.edu.tw.

1 Introduction

Model prediction is an important topic in economic and statistical analysis. The common challenge faced by researchers is to achieve the best prediction when there is a large set of candidate models available. One popular approach for dealing with this problem is model selection. The model selection methods such as the Akaike information criterion (Akaike, 1973) and Mallows' C_p (Mallows, 1973) aim to choose one best model for prediction. However, the selected model might miss some useful information contained in other models, and ignore the uncertainty across different candidate models. The other popular approach to achieve the best prediction is model averaging. Unlike model selection, model averaging incorporates all available information and constructs a weighted average of the individual prediction from all potential models. The model averaging estimator aims to achieve the best trade-off between bias and variance, and tends to perform better than the model selection estimator in finite samples.

There are two main model averaging methods, Bayesian model averaging and frequentist model averaging; see Claeskens and Hjort (2008), Moral-Benito (2015), and Steel (2020) for a literature review of both methods. Bayesian model averaging has a long history, and has been widely used in statistical and economic studies. In contrast to Bayesian model averaging, there is a rapidly growing development of frequentist model averaging approaches in the past two decades, including information criterion weighting (Buckland et al., 1997; Hjort and Claeskens, 2003; Zhang and Liang, 2011), adaptive regression by mixing models (Yang, 2000, 2001; Yuan and Yang, 2005), optimal model averaging (Hansen, 2007; Hansen and Racine, 2012; Liu and Okui, 2013; Zhang et al., 2014), plug-in averaging (Liu, 2015; Charkhi et al., 2016; Cheng et al., 2019), and others. One important theoretical justification of the frequentist approach is to demonstrate the asymptotic optimality of the model averaging estimator, that is, the model averaging estimator asymptotically achieves the lowest possible squared error. Most existing studies on optimal model averaging, however, establish the asymptotic optimality based on an in-sample squared error loss function instead of the out-of-sample prediction risk function, which limits their applications for prediction.

In this paper, we consider the model averaging prediction in a quasi-likelihood framework.

The main goal of this paper is to construct an averaging prediction based on a large number of candidate models in a quasi-likelihood setting that allows for parameter uncertainty and model misspecification. In our framework, the candidate models could be nested or non-nested, and all of the potential models could be misspecified. For each candidate model, we are uncertain about which model parameters should be included in the model, and we allow for the parameter uncertainty. It is well known that the conditional expectation of a dependent variable given the covariates is the best predictor. We adopt a frequentist model averaging approach to estimate the unknown conditional expectation function and then propose an averaging prediction that selects the data-driven weights by minimizing a K-fold cross-validation criterion. Our method is not limited to linear regression models and can apply to binary, discrete, or continuous dependent variables. For example, consider a binary outcome variable. A usual practice in empirical studies is to choose between the probit and logit models and make a prediction. Instead of choosing between these two non-nested models, we can take the model misspecification into account and construct an averaging prediction based on these two estimates.

The idea of the K-fold cross-validation is to divide the data set into K groups and treat each group as a validation data set to evaluate the model. The proposed K-fold cross-validation criterion is a quadratic function of the model weights, so the solution can be found numerically via quadratic programming. In this paper, we provide two theoretical justifications for the K-fold cross-validation. We first consider a scenario in which all candidate models are misspecified. In this scenario, we show that the proposed averaging prediction using K-fold cross-validation weights is asymptotically optimal in the sense of achieving the lowest possible prediction risk in the class of model averaging prediction estimators. Thus, this optimality property of the prediction risk function provides a complement to existing methods that focus on the in-sample squared error loss function. In the second scenario, we allow for some correctly specified models in the model set. In this case, we demonstrate that the K-fold cross-validation asymptotically assigns all weights to these correctly specified models. This novel result of asymptotically selecting the correctly specified models corresponds to the consistency property in model selection.

In simulations, we examine the finite sample performance of the proposed model averaging

prediction using K-fold cross-validation weights. Monte Carlo simulations show that the proposed method generally produces lower empirical risk than other existing model selection and model averaging methods when all candidate models are misspecified. In the other scenario, where the model set includes correctly specified models, simulations show that the sum of empirical K-fold cross-validation weights placed on the correctly specified models is monotonically increasing and generally converges to one as the sample size increases, which is consistent with our theoretical finding. As an empirical illustration, we apply the model averaging approach to credit card default prediction. Our empirical results show that the proposed model averaging prediction generally achieves lower mean squared prediction error than other existing methods.

We now discuss the related literature. The cross-validation method was introduced by Allen (1974), Stone (1974), and Geisser (1975) for model selection in regression models. Its asymptotic optimality is demonstrated by Li (1987) and Andrews (1991) for homoskedastic and heteroskedastic regression, respectively, and its consistency property is investigated by Shao (1993) and Shao (1997). In recent years, cross-validation has been used for selection of model weights in various model setups, including the heteroskedastic linear regression model (Hansen and Racine, 2012), the linear regression model with lagged dependent variables (Zhang et al., 2013), the high-dimensional linear regression model (Ando and Li, 2014), the factor-augmented regression model (Cheng and Hansen, 2015), the quantile regression model (Lu and Su, 2015), the longitudinal data model (Gao et al., 2016), the high-dimensional generalized linear model (Ando and Li, 2017), the vector autoregressive model (Liao and Tsay, 2020), and the time-varying parameter regression model (Sun et al., 2021). There are two main differences between this paper and these studies. First, we provide both optimality and consistency for the proposed K-fold cross-validation, while most of these studies only focus on the optimality property. Second, our goal is to make predictions using the optimal model averaging approach and we establish the asymptotic optimality based on an out-of-sample prediction risk function instead of an in-sample squared error loss function. To demonstrate the asymptotic optimality from the prediction aspect, we provide a new strategy to bound the difference between the K-fold cross-validation and prediction risk function.

Consistency and optimality are two important properties of model averaging estimators.

It is well known that the Bayesian information criterion is consistent in selecting the true model, and this consistency property is shared by Bayesian model averaging; see Steel (2020) for a detailed discussion. Fernández et al. (2001) demonstrate the consistency of Bayesian model averaging under general conditions. Fernández-Villaverde and Rubio-Ramírez (2004) show that Bayesian estimates converge to their pseudo-true values, and the model that is best under the Kullback-Leibler measure has the highest posterior probability as the sample size goes to infinity. In addition, Bayesian model averaging achieves the optimal shrinkage in high-dimensional linear regression models. Castillo et al. (2015) show that Bayesian model averaging has an optimal rate for recovery of the true model from the data. However, as pointed out by Yang (2005), Bayesian model averaging is suboptimal in terms of the minimax-rate estimation of the regression function, and hence it might not be asymptotically optimal in the prediction risk function. Unlike Bayesian model averaging, the existing literature on the consistency of frequentist model averaging is comparatively small. Hansen (2014) and Liu (2015) show that the least squares averaging estimators are root-n consistent in a local asymptotic framework. Zhang (2015) investigates the consistency of model averaging estimators including a smoothed Akaike or Bayesian information criterion under the standard asymptotics with a fixed parameters setup. These studies, however, focus on the consistency of parameter estimates instead of the selecting the correctly specified models.

This paper is related to ensemble learning in the machine learning literature. The idea of ensemble learning is to combine multiple learning algorithms to provide better predictions. The common ensemble learning methods include boosting (Schapire, 1990), stacking (Wolpert, 1992), bagging (Breiman, 1996), and AdaBoost (Freund and Schapire, 1997); see Zhang and Ma (2012) and Brownlee (2018) for more discussion and applications. Our method is closely related to stacking, which splits the data into training and validation sets and then uses outputs from the validation set to combine the predictions of different learning algorithms. Recently, Qiu et al. (2020) propose Mallows-type criteria to combine machine learning techniques, and investigate their asymptotic and finite sample behavior. The theoretical properties of K-fold cross-validation in machine learning applications would be an important research topic to study.

The outline of the paper is as follows. Section 2 presents the model framework and the

prediction procedure. Section 3 introduces the K-fold cross-validation. Section 4 presents the asymptotic properties of the proposed averaging prediction. Section 5 examines the finite sample properties of the proposed method. Section 6 provides the empirical study, and Section 7 concludes the paper. Proofs are included in the Appendix.

2 Prediction procedure

Suppose we have n independent and identically distributed (i.i.d.) observations $\{(Y_i, \mathbf{X}_i); i = 1, \ldots, n\}$, where Y_i is a scalar dependent variable and $\mathbf{X}_i$ is a vector of predictors. Let the likelihood function be

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(Y_i|\mathbf{X}_i, \boldsymbol{\theta}), \tag{2.1}$$

where f is an unknown conditional probability density function and $\boldsymbol{\theta}$ is a vector of unknown parameters. The number of predictors could be different from the number of parameters, but we do not let the numbers of predictors and parameters increase with the sample size n. The dependent variable is allowed to be binary, discrete, or continuous. Hence, this framework could apply to a linear regression model with the standard Gaussian likelihood, categorical regression models, and nonlinear regression models.

Our goal is to make predictions given the observed data $(Y_i, \mathbf{X}_i)$ without imposing any assumptions on the structure of the model or the relationship between the dependent variable and predictors. Consider a sequence of candidate models $m = 1, \ldots, M$, where the mth candidate model uses the following quasi-likelihood function

$$\prod_{i=1}^{n} f_{(m)}(Y_i|\mathbf{X}_i, \boldsymbol{\theta}_{(m)}), \tag{2.2}$$

where the function $f_{(m)}$ is known, but it could be misspecified, and $\boldsymbol{\theta}_{(m)}$ is a vector of the unknown parameters. That is, $f_{(m)}$ could be different from the true conditional probability density function f. Since the true values of $\boldsymbol{\theta}$ could be zeros, we could also restrict some elements of $\boldsymbol{\theta}$ zeros to obtain candidate models and allow for the parameter uncertainty. Therefore, a candidate model could have a misspecified conditional probability density function, a vector of potentially relevant parameters, or both. Furthermore, the set of possible models could be nested or non-nested and M could go to infinity with the sample size n.

Let $\widehat{\boldsymbol{\theta}}_{(m)}$ denote the maximum likelihood estimator of $\boldsymbol{\theta}_{(m)}$ in the mth candidate model. Thus, the prediction of Y_{n+1} associated with the new observation $\mathbf{X}_{n+1}$ from this mth model is

$$\widehat{Y}_{(m),n+1} = \mathrm{E}_{(m)}(Y_{n+1}|\mathbf{X}_{n+1}, \widehat{\boldsymbol{\theta}}_{(m)}) = \int y f_{(m)}(y|\mathbf{X}_{n+1}, \widehat{\boldsymbol{\theta}}_{(m)})dy, \tag{2.3}$$

where $\mathrm{E}_{(m)}$ is the expectation taken under the mth candidate model.

Let $\mathbf{w} = (w_1, \ldots, w_M)^\top$ be a weight vector with $w_m \geq 0$ and $\sum_{m=1}^M w_m = 1$. That is, the weight vector $\mathbf{w}$ belongs to the set $\mathcal{W} = \{\mathbf{w} \in [0,1]^M : \sum_{m=1}^M w_m = 1\}$. Combining all possible predicted values of $\widehat{Y}_{(m),n+1}$, we construct an averaging prediction as

$$\widehat{Y}_{n+1}(\mathbf{w}) = \sum_{m=1}^M w_m \widehat{Y}_{(m),n+1}. \tag{2.4}$$

Notice that the proposed averaging prediction is closely related to finite mixture models. Finite mixture models are widely used to model heterogeneity in a population; see Melnykov and Maitra (2010) and McLachlan et al. (2019) for literature reviews. Suppose that $\sum_{m=1}^M w_m \int |y| f_{(m)}(y|\mathbf{X}_{n+1}, \widehat{\boldsymbol{\theta}}_{(m)})dy < \infty$. Then, by the definition of $\widehat{Y}_{(m),n+1}$ in (2.3), we can rewrite the averaging prediction (2.4) as $\widehat{Y}_{n+1}(\mathbf{w}) = \sum_{m=1}^M w_m \int y f_{(m)}(y|\mathbf{X}_{n+1}, \widehat{\boldsymbol{\theta}}_{(m)})dy = \int y \sum_{m=1}^M w_m f_{(m)}(y|\mathbf{X}_{n+1}, \widehat{\boldsymbol{\theta}}_{(m)})dy$, where the last equality holds by Fubini's theorem. Thus, the weighted function $\sum_{m=1}^M w_m f_{(m)}(y|\mathbf{X}_{n+1}, \widehat{\boldsymbol{\theta}}_{(m)})$ is a mixture density function. Although there are some similarities between the averaging prediction and finite mixture models, there are some important differences. First, $f_{(m)}$ is a true component density in finite mixture models, while it is allowed to be misspecified in our framework. Second, the parameters $\boldsymbol{\theta}_{(m)}$ and w_m are often estimated by expectation-maximization algorithm in finite mixture models, while for the proposed averaging prediction, we first estimate $\boldsymbol{\theta}_{(m)}$ by maximum likelihood estimation, and then select w_m by minimizing a K-fold cross-validation criterion. Third, our goal is to minimize the prediction risk instead of modelling heterogeneity, and we demonstrate the asymptotic optimality of the proposed method, which is not obtained in finite mixture models.

We now provide some examples to illustrate the averaging prediction procedure. The first example considers nested candidate models, while the second example considers non-nested candidate models.

Example 1. Consider a linear regression model: $Y_i = \sum_{j=1}^{\infty} \beta_j X_{ji} + e_i$ with $\mathrm{E}(e_i|\mathbf{X}_i) = 0$ and $\mathrm{E}(e_i^2|\mathbf{X}_i) = \sigma^2$, where Y_i is a scalar dependent variable, and $\mathbf{X}_i = (X_{1i}, X_{2i}, ...)^\top$ is countably infinite. The data $(Y_i, \mathbf{X}_i)$ are i.i.d. and the unknown parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. Suppose that we have a sequence of nested candidate models $m = 1, \ldots, M$, where the mth candidate model uses the first m predictors in $\mathbf{X}_i$ and the standard Gaussian likelihood. Note that under normality, the maximum likelihood estimator of $\boldsymbol{\theta}$ is equivalent to the ordinary least squares estimator. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$, and $\mathbf{X}_{(m)}$ be the regressors in the mth model that includes the first m predictors in $\mathbf{X}$. We assume that $\mathbf{X}_{(m)}$ has full column rank for $m = 1, \ldots, M$. Thus, the least squares estimator of $\boldsymbol{\beta}$ for the mth model is $\widehat{\boldsymbol{\beta}}_{(m)} = (\mathbf{X}_{(m)}^\top \mathbf{X}_{(m)})^{-1} \mathbf{X}_{(m)}^\top \mathbf{Y}$, and σ^2 is estimated by $\widehat{\sigma}_{(m)}^2 = \frac{1}{n} \widehat{\mathbf{e}}_{(m)}^\top \widehat{\mathbf{e}}_{(m)}$, where $\widehat{\mathbf{e}}_{(m)} = \mathbf{Y} - \mathbf{X}_{(m)}^\top \widehat{\boldsymbol{\beta}}_{(m)}$. The prediction of Y_{n+1} from the mth model is $\widehat{Y}_{(m),n+1} = \mathbf{X}_{(m),n+1}^\top \widehat{\boldsymbol{\beta}}_{(m)}$, and the averaging prediction is $\widehat{Y}_{n+1}(\mathbf{w}) = \sum_{m=1}^{M} w_m \widehat{Y}_{(m),n+1} = \sum_{m=1}^{M} w_m \mathbf{X}_{(m),n+1}^\top \widehat{\boldsymbol{\beta}}_{(m)}$; this is the least squares averaging estimator studied in Hansen (2007) and Hansen (2008). In this example, we consider a set of nested candidate models, which implies an ordering of predictors, and we will use this example to verify the high-level assumptions. As shown in Section 4, the nested property is crucial to verify Assumption 2 (iii), but not other assumptions. The ordering of predictors, however, is not relevant to verification of these assumptions. Hence, we do not impose any assumption on the ordering of predictors, and the ordering is not required to be correct in this example.[1]

Example 2. Suppose that we observe a binary dependent variable $Y_i \in \{0, 1\}$ and a vector of predictors $\mathbf{X}_i$. Assume that Y_i is conditionally Bernoulli with $Pr(Y_i = 1|\mathbf{X}_i) = F(\mathbf{X}_i^\top \boldsymbol{\beta})$, where $F(\cdot)$ is an unknown cumulative distribution function. Our goal is to estimate $Pr(Y_{n+1} = 1|\mathbf{X}_{n+1})$ and predict Y_{n+1}. We consider two non-nested candidate models. The first candidate model is a probit model setting $F(u) = \Phi(u)$, where Φ is a standard normal distribution function, while the second candidate model is a logit model setting $F(u) = (1 + e^{-u})^{-1}$. Let $\widehat{\boldsymbol{\beta}}_P$ and $\widehat{\boldsymbol{\beta}}_L$ be the maximum likelihood estimator for the probit and logit model, respectively. The predictions of Y_{n+1} based on these two models are $\widehat{Y}_{(1),n+1} = \Phi(\mathbf{X}_{n+1}^\top \widehat{\boldsymbol{\beta}}_P)$ and $\widehat{Y}_{(2),n+1} = (1 + e^{-\mathbf{X}_{n+1}^\top \widehat{\boldsymbol{\beta}}_L})^{-1}$. Instead of choosing between these

[1]The simulations show that the ordering of predictors has little impact on the finite sample properties of the proposed model averaging estimator. The results are available on request from the authors.

two non-nested models, we can take the model misspecification into account and construct an averaging prediction as $\widehat{Y}_{n+1}(\mathbf{w}) = w_1\widehat{Y}_{(1),n+1} + w_2\widehat{Y}_{(2),n+1}$.

3 K-fold cross-validation

In this section, we propose a K-fold cross-validation criterion to select the model weights for the averaging prediction. The idea of the K-fold cross-validation is to split the sample into K groups and treat each group as a validation sample (or a testing data set) to evaluate the model. We then select the model weights by minimizing the sum of squared prediction errors obtained from all groups. Unlike the Mallows criterion or other information criteria, which require derivation of the penalty term on a case-by-case basis in nonlinear models, the implementation of the K-fold cross-validation is easy and flexible, and it seldom relies on the model structure.

We now describe how to calculate the K-fold cross-validation criterion and construct an averaging prediction with data-driven weights in detail. The proposed method works by the following steps.

Step 1: Divide the data set into K groups with $2 \leq K \leq n$, so that there are $J = n/K$ observations in each group.

Step 2: For $k = 1, \ldots, K$,

(a) Exclude the kth group from the data set and use the remaining $n - J$ observations to calculate the estimator $\widehat{\boldsymbol{\theta}}_{(m)}^{[-k]}$ for each model. That is, $\widehat{\boldsymbol{\theta}}_{(m)}^{[-k]}$ is the estimator of $\boldsymbol{\theta}_{(m)}$ in the mth model without using the observations from the kth group.

(b) Calculate the predictions for observations within the kth group for each model. That is, we calculate the prediction of $Y_{(k-1)\times J+j}$ by

$$\widetilde{Y}_{(m),j}^{[-k]} = \int y f_{(m)}(y|\mathbf{X}_{(k-1)\times J+j}, \widehat{\boldsymbol{\theta}}_{(m)}^{[-k]})dy, \tag{3.1}$$

for $j = 1, \ldots, J$ and $m = 1, \ldots, M$, where the subscript $(k-1) \times J + j$ denotes the observations in the kth group.

Step 3: Compute the predictions for all observations for each model as follows

$$\widetilde{\mathbf{Y}}_{(m)} = (\widetilde{Y}^{[-1]}_{(m),1}, \ldots, \widetilde{Y}^{[-1]}_{(m),J}, \ldots, \widetilde{Y}^{[-K]}_{(m),1}, \ldots, \widetilde{Y}^{[-K]}_{(m),J})^\top, \tag{3.2}$$

and construct the K-fold cross-validation criterion

$$CV_K(\mathbf{w}) = \frac{1}{n}\|\mathbf{Y} - \widetilde{\mathbf{Y}}(\mathbf{w})\|^2, \tag{3.3}$$

where $\|\cdot\|$ is the Euclidean norm, $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, and $\widetilde{\mathbf{Y}}(\mathbf{w}) = \sum_{m=1}^M w_m \widetilde{\mathbf{Y}}_{(m)} = (\widetilde{Y}_1^{[-1]}(\mathbf{w}), \ldots, \widetilde{Y}_J^{[-K]}(\mathbf{w}))^\top$, where $\widetilde{Y}_j^{[-k]}(\mathbf{w}) = \sum_{m=1}^M w_m \widetilde{Y}^{[-k]}_{(m),j}$ is the average prediction of $Y_{(k-1)\times J+j}$.

Step 4: Select the model weights by minimizing the K-fold cross-validation criterion

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{W}}{\operatorname{argmin}} CV_K(\mathbf{w}), \tag{3.4}$$

and construct an averaging prediction for Y_{n+1} as follows

$$\widehat{Y}_{n+1}(\widehat{\mathbf{w}}) = \sum_{m=1}^M \widehat{w}_m \widehat{Y}_{(m),n+1}. \tag{3.5}$$

Notice that the proposed K-fold cross-validation criterion is a quadratic function of the weight vector. Let $\widetilde{\mathbf{e}} = (\widetilde{\mathbf{e}}_{(1)}, \ldots, \widetilde{\mathbf{e}}_{(M)})$, where $\widetilde{\mathbf{e}}_{(m)} = \mathbf{Y} - \widetilde{\mathbf{Y}}_{(m)}$ is the K-fold cross-validation prediction error for the mth model. Then, the proposed criterion (3.3) can be written as a quadratic function of $\mathbf{w}$ as follows

$$CV_K(\mathbf{w}) = \frac{1}{n}\mathbf{w}^\top \widetilde{\mathbf{e}}^\top \widetilde{\mathbf{e}} \mathbf{w}. \tag{3.6}$$

Therefore, the K-fold cross-validation weights can be computed numerically via quadratic programming, and numerical algorithms of quadratic programming are available for most programming languages.

The most common choices of K are 5 and 10, and it corresponds to leave-one-out cross-validation when $K = n$. In the finite sample, the results could be sensitive to the choice of K, especially when K is too small, but the computational cost could be quite heavy when both n and K are large. As suggested by an anonymous referee, we consider the following data-driven selection of K. For a given value of K, let $\widehat{\mathbf{w}}_K$ be the K-fold cross-validation

weights obtained from Step 4, and $CV(K) = n^{-1}\widehat{\mathbf{w}}_K^\top \widetilde{\mathbf{e}}^\top \widetilde{\mathbf{e}}\widehat{\mathbf{w}}_K$ be the K-fold cross-validation criterion evaluated at $\widehat{\mathbf{w}}_K$. We then select K by minimizing $CV(K)$ as follows

$$\widehat{K} = \underset{K\in\{2,\ldots,n\}}{\operatorname{argmin}} CV(K). \tag{3.7}$$

Notice that the minimization problem (3.7) must be solved numerically. In practice, we could conduct a grid search for K to reduce the computational burden. That is, we first consider a grid of values for K: $\{K_1, \ldots, K_J\}$, and we select K by $\widehat{K} = \underset{K\in\{K_1,\ldots,K_J\}}{\operatorname{argmin}} CV(K)$. Another possible way to select K is to follow Arlot and Lerasle (2016) and investigate the non-asymptotic oracle inequality for K-fold cross-validation criterion, and study its variance for each value of K. A theoretical investigation is beyond the scope of the present paper and is left for future research.

4 Theoretical properties

In this section, we present the asymptotic properties of the proposed averaging prediction. Define the risk function as $R(\mathbf{w}) \equiv \mathrm{E}[\{\widehat{Y}_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2]$, where $\mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1}) = \int yf(y|\mathbf{X}_{n+1}, \boldsymbol{\theta})dy$ is the true conditional expectation function. Ideally, one would aim to select the model weights $\mathbf{w}$ to minimize the risk function $R(\mathbf{w})$ with respect to $\mathbf{w}$ in the set $\mathcal{W}$. Unfortunately, this is infeasible because this minimization depends on the unknown conditional probability density function f. Instead of minimizing $R(\mathbf{w})$ directly, we select the data-driven weights by minimizing the K-fold cross-validation criterion and demonstrate that the empirical K-fold cross-validation weights asymptotically minimize the risk function.

We state the assumptions required for asymptotic results, where all limiting processes here and throughout the text are with respect to $n \to \infty$.

Assumption 1. *Suppose that $M \leq n$. There exists a limiting value $\boldsymbol{\theta}^*_{(m)}$ for $\widehat{\boldsymbol{\theta}}_{(m)}$ such that $\widehat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}^*_{(m)} = O_p(M^{1/2}n^{-1/2})$ uniformly for $m = 1, \ldots, M$.*

Assumption 1 is a high-level condition, which ensures that the estimator of $\boldsymbol{\theta}_{(m)}$ in each candidate model has a limit $\boldsymbol{\theta}^*_{(m)}$. Here, $\boldsymbol{\theta}^*_{(m)}$ might be interpreted as a pseudo-true value.

This condition is commonly used to analyze the asymptotic properties of the model averaging estimator in nonlinear models, for example, Zhang et al. (2016) and Ando and Li (2017). When M is fixed, Assumption 1 holds under appropriate primitive assumptions; see conditions in Theorem 3.2 of White (1982). Since $M = 2$ in Example 2, Assumption 1 can be derived by the primitive assumptions in Theorem 3.2 of White (1982). Unlike Example 2, we allow M to increase with the sample size in Example 1. In Appendix B, we provide the primitive conditions for Assumption 1 in Example 1.

Note that Assumption 1 implies that the dimension of $\boldsymbol{\theta}_{(m)}$ is fixed. Although the number of unknown parameters in each candidate model is fixed, there are two reasons why we allow M to go to infinity. First, the number of candidate functions $f_{(m)}$ could go to infinity. Second, the number of predictors, that is, the dimension of $\mathbf{X}_i$, could go to infinity. In the second case, one could construct the candidate models in some way such that the dimension of $\boldsymbol{\theta}_{(m)}$ is fixed, but M goes to infinity. For example, Ando and Li (2014) proposed grouping predictors with similar marginal correlations together to form a candidate model. In this case, the number of predictors used in each candidate model is fixed, but the number of candidate models goes to infinity. An extension to the case with a diverging number of parameters in each candidate model is left for future research.

We now introduce some notation associated with the limiting value $\boldsymbol{\theta}^*_{(m)}$. The prediction of Y_{n+1} calculated based on the limiting value $\boldsymbol{\theta}^*_{(m)}$ in the mth model is

$$Y^*_{(m),n+1} = \int y f_{(m)}(y|\mathbf{X}_{n+1}, \boldsymbol{\theta}^*_{(m)})dy, \tag{4.1}$$

and the averaging prediction calculated based on the limiting value $\boldsymbol{\theta}^*_{(m)}$ is

$$Y^*_{n+1}(\mathbf{w}) = \sum_{m=1}^{M} w_m Y^*_{(m),n+1}. \tag{4.2}$$

Notice that $\widehat{\boldsymbol{\theta}}_{(m)}$ and $\widehat{\boldsymbol{\theta}}_m^{[-k]}$ have the same limiting values $\boldsymbol{\theta}^*_{(m)}$ because n and $n - n/K$ have the same order for any $K \in \{2, \ldots, n\}$. Equations (4.1) and (4.2) correspond to Equations (2.3) and (2.4), except $\widehat{\boldsymbol{\theta}}_{(m)}$ is replaced by $\boldsymbol{\theta}^*_{(m)}$. Similarly, we use $R^*(\mathbf{w})$ to denote the risk function calculated based on the limiting parameter value instead of the estimated parameter value, that is, $R^*(\mathbf{w}) \equiv \mathrm{E}\left[\{Y^*_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2\right]$. Let $\varepsilon_i = Y_i - \mathrm{E}(Y_i|\mathbf{X}_i)$ denote the

error term in the prediction problem and let $\mathcal{O}(\boldsymbol{\theta}^*_{(m)}, \varrho)$ denote a neighborhood of $\boldsymbol{\theta}^*_{(m)}$ for some constant ϱ such that $\|\boldsymbol{\theta}^*_{(m)} - \boldsymbol{\theta}\| \leq \varrho$ for any $\boldsymbol{\theta} \in \mathcal{O}(\boldsymbol{\theta}^*_{(m)}, \varrho)$.

Assumption 2. *For $i = 1, \ldots, n$, $k = 1, \ldots, K$, and $j = (k-1) \times J + 1, \ldots, (k-1) \times J + J$, (i) $E(\varepsilon_i^2)$ and $E(Y_i^2)$ are both $O(1)$; (ii) $\widetilde{Y}^{[-k]}_{(m),j}$ is differentiable with respect to $\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}$; (iii) there exists a constant ϱ such that*

$$E \sup_{\boldsymbol{\theta}^\star \in \mathcal{O}(\boldsymbol{\theta}^*_{(m)}, \varrho)} \left\| \frac{\partial \widetilde{Y}^{[-k]}_{(m),j}}{\partial \widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}} \Big|_{\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)} = \boldsymbol{\theta}^\star} \right\|^2 = O(1),$$

uniformly for $m = 1, \ldots, M$, and $E(Y^*_{(m),i})^2$ is $O(1)$ uniformly for $m = 1, \ldots, M$.

Assumption 2 concerns the boundedness and differentiability. Assumptions 2 (i)-(ii) are straightforward. To illustrate Assumption 2 (iii), we consider the two examples in Section 2. In Example 1, it is easy to show that

$$\sup_{\boldsymbol{\theta}^\star \in \mathcal{O}(\boldsymbol{\theta}^*_{(m)}, \varrho)} \frac{\partial \widetilde{Y}^{[-k]}_{(m),j}}{\partial \widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}} \Big|_{\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)} = \boldsymbol{\theta}^\star} = \mathbf{X}_{(m),(k-1)\times J+j} \quad \text{and} \quad Y^*_{(m),i} = \mathbf{X}^\top_{(m),i} \boldsymbol{\beta}^*_{(m)},$$

where $\boldsymbol{\beta}^*_{(m)}$ is the limiting value of $\widehat{\boldsymbol{\beta}}_{(m)}$. Then, by the nested framework in Example 1, Assumption 2 (iii) holds if $\mathrm{E}\|\mathbf{X}_{(M),i}\|^2 = O(1)$ and $\|\boldsymbol{\beta}^*_{(m)}\|^2$ are all bounded by a fixed constant. Therefore, Assumption 2 (iii) is quite mild. Note that the dimension of parameters in each candidate model is assumed fixed, but M could go to infinity; see the discussion after Assumption 1.

In Example 2, we can show that

$$\begin{aligned}
&Y^*_{(1),i} = \Phi(\mathbf{X}_i^\top \boldsymbol{\beta}^*_P), \qquad Y^*_{(2),i} = (1 + e^{-\mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}^*_L})^{-1}, \\
&\sup_{\boldsymbol{\theta}^\star \in \mathcal{O}(\boldsymbol{\theta}^*_{(1)}, \varrho)} \frac{\partial \widetilde{Y}^{[-k]}_{(1),j}}{\partial \widehat{\boldsymbol{\theta}}^{[-k]}} \Big|_{\widehat{\boldsymbol{\theta}}^{[-k]}_{(1)} = \boldsymbol{\theta}^\star} = \sup_{\boldsymbol{\beta}^\star_P \in \mathcal{O}(\boldsymbol{\beta}^*_P, \varrho)} \phi(\mathbf{X}^\top_{(k-1)\times J+j} \boldsymbol{\beta}^\star_P) \mathbf{X}_{(k-1)\times J+j}, \\
&\sup_{\boldsymbol{\theta}^\star \in \mathcal{O}(\boldsymbol{\theta}^*_{(2)}, \varrho)} \frac{\partial \widetilde{Y}^{[-k]}_{(2),j}}{\partial \widehat{\boldsymbol{\theta}}^{[-k]}} \Big|_{\widehat{\boldsymbol{\theta}}^{[-k]}_{(2)} = \boldsymbol{\theta}^\star} = \sup_{\boldsymbol{\beta}^\star_L \in \mathcal{O}(\boldsymbol{\beta}^*_L, \varrho)} \Big[\Big\{ (1 + e^{-\mathbf{X}^\top_{(k-1)\times J+j} \boldsymbol{\beta}^\star_L})^{-1} - (1 + e^{-\mathbf{X}^\top_{(k-1)\times J+j} \boldsymbol{\beta}^\star_L})^{-2} \Big\} \\
&\times \mathbf{X}_{(k-1)\times J+j} \Big],
\end{aligned}$$

where $\phi(\cdot)$ is a standard normal density function, and $\boldsymbol{\beta}^*_P$ and $\boldsymbol{\beta}^*_L$ are limiting values of $\widehat{\boldsymbol{\beta}}_P$ and $\widehat{\boldsymbol{\beta}}_L$, respectively. Notice that $0 \leq \phi(\mathbf{X}_i^\top \boldsymbol{\beta}^\star_P) \leq 1$, and $0 \leq (1+e^{-\mathbf{X}_i^\top \boldsymbol{\beta}^\star_L})^{-1} - (1+e^{-\mathbf{X}_i^\top \boldsymbol{\beta}^\star_L})^{-2} \leq 1$

for any $\boldsymbol{\beta}_P^\star \in \mathcal{O}(\boldsymbol{\beta}_P^*, \varrho)$ and $\boldsymbol{\beta}_L^\star \in \mathcal{O}(\boldsymbol{\beta}_L^*, \varrho)$. Therefore, Assumption 2 (iii) holds in Example 2 if $\mathrm{E}\|\mathbf{X}_i\|^2 = O(1)$.

Let $\xi_n = \inf_{\mathbf{w}\in\mathcal{W}} R^*(\mathbf{w})$ denote the minimum risk in the class of averaging estimators associated with the limiting value $\boldsymbol{\theta}_{(m)}^*$.

Assumption 3. $\xi_n^{-1} \sup_{w\in\mathcal{W}} \left[\{\widehat{Y}_{n+1}(\mathbf{w}) - E(Y_{n+1}|\mathbf{X}_{n+1})\}^2 - \{Y_{n+1}^*(\mathbf{w}) - E(Y_{n+1}|\mathbf{X}_{n+1})\}^2\right]$ *is uniformly integrable.*

As shown in (C.3) in the proof of Theorem 1, we derive the following result

$$\xi_n^{-1} \sup_{w\in\mathcal{W}} \left[\{\widehat{Y}_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2 - \{Y_{n+1}^*(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2\right] = o_p(1).$$

Although Assumption 3 might not be straightforward, it is only imposed to ensure that the expectation of the above equation is $o(1)$.

Let $\varepsilon_{(m),i} = Y_{(m),i}^* - \mathrm{E}(Y_i|\mathbf{X}_i)$ denote the prediction error based on the limiting value $\boldsymbol{\theta}_{(m)}^*$ in the mth model. The following assumption imposes moment conditions of the error term and prediction error.

Assumption 4. *var*$(\varepsilon_{(m),i}\varepsilon_{(m'),i})$ *and var*$(Y_{(m),i}^*\varepsilon_i)$ *are bounded by a constant uniformly for* $m = 1, \ldots, M$ *and* $m' = 1, \ldots, M$.

We now verify Assumption 4 in Examples 1 and 2. In Example 1, recall that $Y_{(m),i}^* = \mathbf{X}_{(m),i}^\top \boldsymbol{\beta}_{(m)}^*$ and $\boldsymbol{\beta}_{(m)}^*$ is the limiting value of $\widehat{\boldsymbol{\beta}}_{(m)}$. Therefore, a simple sufficient condition for Assumption 4 is that $\mathrm{E}[(\mathbf{X}_{(m),i}^\top \boldsymbol{\beta}_{(m)}^*)^4]$, $\mathrm{E}(Y_i^4)$, and $\mathrm{E}(\varepsilon_i^4)$ are all bounded by a constant. For Example 2, notice that $Y_{(1),i}^* = \Phi(\mathbf{X}_i^\top \boldsymbol{\beta}_P^*)$, $Y_{(2),i}^* = (1 + e^{-\mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}_L^*})^{-1}$, and $\mathrm{E}(Y_i|\mathbf{X}_i) = Pr(Y_i = 1|\mathbf{X}_i)$ are all between 0 and 1, and $Y_i \in \{0, 1\}$. Hence, Assumption 4 holds directly in Example 2.

Assumption 5. $n^{-1/2} M \xi_n^{-1} = o(1)$.

Assumption 5 puts a bound on the number of models relative to the sample size, and it specifies that M grows at a rate no faster than $n^{1/2}\xi_n$. Assumption 5 is similar to Condition 7 of Ando and Li (2014), Condition C.6 of Zhang et al. (2016), and Condition A3 of Ando

and Li (2017). In Appendix B, we verify this assumption in Example 1. In addition, this assumption requires that all candidate models be misspecified. To better understand this condition, suppose that the m^oth model is correctly specified. Thus, we have $f_{(m^o)}(\cdot) = f(\cdot)$ and $\boldsymbol{\theta}^*_{(m^o)} = \boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is the true value defined in (2.1). Then it follows that

$$\xi_n = \inf_{\mathbf{w}\in\mathcal{W}} \mathrm{E}\left[\{Y^*_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2\right] \leq \mathrm{E}\left[\{Y^*_{(m^o),n+1} - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2\right] = 0,$$

and thus Assumption 5 is violated. Therefore, if one of the candidate models is correctly specified, then Assumption 5 does not hold. We will discuss the case where all candidate models are misspecified first, and then discuss the alternative condition for Assumption 5 and the case where some models are correctly specified later.

We now present two theoretical justifications for the K-fold cross-validation criterion. The first justification is that the proposed averaging prediction using K-fold cross-validation weights is asymptotically optimal, and thus the proposed method achieves the lowest possible prediction risk. In other words, the K-fold cross-validation weights asymptotically minimize the prediction risk. The following theorem shows the asymptotic optimality of the K-fold cross-validation criterion.

Theorem 1. *Under Assumptions 1-5, we have*

$$\frac{R(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w})} \to 1 \tag{4.3}$$

in probability.

The asymptotic optimality of the weight selection criterion is a standard but important theoretical justification of the model averaging estimator, and model averaging can achieve a lower optimal risk than model selection; see Peng and Yang (2021). However, most existing work, for example, Hansen (2007), Wan et al. (2010), Hansen and Racine (2012), and Zhang et al. (2013), establish the asymptotic optimality based on an in-sample squared error loss function. Unlike these works, we demonstrate the asymptotic optimality based on the out-of-sample prediction risk function, and hence it is more applicable for model averaging on prediction. The proof of the asymptotic optimality from the prediction aspect is not a trivial extension of already existing results, because we are not able to apply the theory developed

in Li (1987), Andrews (1991), Hansen and Racine (2012), and Zhang et al. (2013) directly. Instead of applying Whittle's inequality, we provide a new strategy to bound the difference between the K-fold cross-validation and prediction risk function.

The second justification is that the proposed averaging prediction asymptotically assigns all weights to the correctly specified models if they are included in the model set. Specifically, let $\mathcal{D}$ be the subset of $\{1, \ldots, M\}$ that consists of the indices of the correctly specified models, and let $\widehat{\tau} = \sum_{m \in \mathcal{D}} \widehat{w}_m$ be the sum of K-fold cross-validation weights given to the correctly specified models. We aim to show that $\widehat{\tau} \to 1$ in probability under some regularity conditions.

We first discuss the alternative condition for Assumption 5. Let $\mathcal{W}_S = \{\mathbf{w} \in \mathcal{W} : \sum_{m \notin \mathcal{D}} w_m = 1\}$ be the subset of $\mathcal{W}$ that assigns all weights to the misspecified models. The following assumption is imposed for the case where some models are correctly specified.

Assumption 6. $n^{-1/2} M \left\{\inf_{\mathbf{w} \in \mathcal{W}_S} R^*(\mathbf{w})\right\}^{-1} = o(1)$.

Assumption 6 imposes the restriction on the growth rate of the minimum risk when we construct the averaging prediction by averaging over all misspecified models. It is easy to see that Assumption 6 is equivalent to Assumption 5 when $\mathcal{D}$ is empty, that is, all candidate models are misspecified. Like Assumption 5, Assumption 6 can also be verified in Example 1 using a setup similar to that discussed in Appendix B.

Theorem 2. *Under Assumptions 1, 2, 4 and 6, if $\mathcal{D}$ is not empty, then we have $\widehat{\tau} \to 1$ in probability.*

Theorem 2 shows that the proposed K-fold cross-validation asymptotically assigns all weights to the correctly specified models when the model set includes correctly specified models. This result corresponds to the consistency property in model selection. If there is only one correctly specified model among the candidate models, then Theorem 2 implies that the proposed K-fold cross-validation would select this correctly specified model asymptotically.

5 Simulation study

In this section, we investigate the finite sample performance of model averaging prediction by K-fold cross-validation in three simulation designs. The first design is the binary choice model, and we consider non-nested candidate models. The second design is the linear regression model, and we consider the averaging prediction between an incorrectly specified model and a correctly specified model. The third design is the nonlinear regression model, and we consider a sequence of nested candidate models.

5.1 Binary choice model

In the first simulation design, we generate a sample of n independent binary variables Y_i from the Bernoulli distribution. Here, Y_i takes the value 1 with probability P_i and the value 0 with probability $1 - P_i$. We set the probability P_i as a cumulative distribution function of the exponential distribution as follows

$$P_i = \begin{cases} 1 - \exp(-\eta_i), & \text{if } \eta_i \geq 0 \\ 0, & \text{if } \eta_i < 0 \end{cases},$$

where $\eta_i = \alpha + \beta X_i$ and X_i is generated from a standard normal distribution. We set $\beta = 0.75$, and the parameter α is varied on a grid from 0 to 2. The sample size is varied between $n = 100$, 200, 500, and 1000.

The goal is to estimate $Pr(Y_{n+1} = 1|X_{n+1})$ and predict Y_{n+1}. Note that Y_i is conditionally Bernoulli with $Pr(Y_i = 1|X_i) = P_i$. Thus, we have $\mathrm{E}(Y_{n+1}|X_{n+1}) = Pr(Y_{n+1} = 1|X_{n+1}) = P_{n+1}$ and the corresponding estimator $\widehat{Y}_{n+1} = \widehat{\mathrm{E}}(Y_{n+1}|X_{n+1}) = \widehat{P}_{n+1}$. For simplicity purposes, we consider two candidate models only, the probit and logit models. These two models are non-nested, and both models are misspecified.

We consider the following estimators: (1) Probit model estimator (labeled Probit); (2) Logit model estimator (labeled Logit); (3) Model averaging estimator with equal weights (labeled Equal); (4) Model averaging estimator with 2-fold cross-validation weights (labeled $K = 2$); (5) Model averaging estimator with 5-fold cross-validation weights (labeled $K = 5$); (6) Model averaging estimator with 10-fold cross-validation weights (labeled $K = 10$); (7) Model averaging estimator with n-fold cross-validation weights (labeled $K = n$); and (8)

Model averaging estimator with data-driven K (labeled K-data). The probit and logit models are estimated by the maximum likelihood method; see Example 2 of Section 2 for details. The Equal estimator assigns 1/2 weight to the probit and logit estimators, respectively. The proposed model averaging estimators with K-fold cross-validation and data-driven K are described in Section 3.[2]

We evaluate the finite sample behavior of each estimator based on the empirical risk function. Let S be the number of simulation replications, and let $^{\{s\}}$ denote the sth replication. The empirical risk function is calculated as follows: $\frac{1}{S}\sum_{s=1}^{S}\{\widehat{Y}_{n+1}^{\{s\}}(\widehat{\mathbf{w}}^{\{s\}}) - \mathrm{E}(Y_{n+1}^{\{s\}}|X_{n+1}^{\{s\}})\}^2 = \frac{1}{S}\sum_{s=1}^{S}\{\widehat{P}_{n+1}^{\{s\}}(\widehat{\mathbf{w}}^{\{s\}}) - P_{n+1}^{\{s\}}\}^2$, where $\widehat{P}_{n+1}^{\{s\}}(\widehat{\mathbf{w}}^{\{s\}})$ is the prediction based on Probit, Logit, Equal, and K-fold cross-validation weights in the sth replication, respectively. More precisely, for each simulation replication, we use $\{X_i^{\{s\}}, Y_i^{\{s\}}\}$ for $i = 1, \ldots, n$ to estimate the probit and logit models and calculate the K-fold cross-validation weights for $K = 2$, 5, 10, n, and data-driven K. We then compute $\widehat{P}_{n+1}^{\{s\}}(\widehat{\mathbf{w}}^{\{s\}})$ associated with the new observation x_{n+1} for each method. The risk function is calculated by averaging across 5000 simulation replications. For easy comparison, we divide the risk of each method by that of 10-fold cross-validation weights and report the relative risk. Lower relative risk means better performance on predictions. When the relative risk exceeds one, it indicates that the specified method performs worse than the model averaging estimator with 10-fold cross-validation weights.

In Figure 1, we present the relative risk for $n = 100$, 200, 500, and 1000 in four panels, and in each panel, the relative risk is displayed for α between 0 and 2. We first compare the finite sample performance between Probit and Logit. Probit has smaller relative risk than Logit for small α, but larger relative risk than Logit for large α. Recall that both models are misspecified, and neither Probit nor Logit uniformly dominates the other. For a fixed value of α, the relative risk of Equal is always between those of Probit and Logit. Furthermore, the relative risk of Equal is above one for most ranges of the parameter space, which implies that there is no efficiency gain by taking a simple equal-weighted average between Probit and Logit.

We next examine the finite sample behavior of the proposed averaging prediction using

[2]To reduce the computational burden, we conduct a grid search for the data-driven K. For example, for $n = 200$, we evaluate $CV(K)$ and then select K from the following the candidate values $K \in \{2, 5, 10, 20, 30, 40, 50, 66, 100, 200\}$.

K-fold cross-validation weights with fixed and data-driven K. The relative risk of K-fold cross-validation is quite similar for $K = 2$, 5, 10, n, and data-driven K, and they dominate the Equal estimator for most situations. When the sample size is small, the K-fold cross-validation performs better than Logit for small α, but the relative risk of the K-fold cross-validation is slightly larger than that of Logit for large α. When the sample size increases, the risk of Logit relative to that of K-fold cross-validation is getting larger for small α, and the relative risk of the K-fold cross-validation is quite similar to that of Logit for large α. The pattern of relative performance between Probit and K-fold cross-validation is quite similar to that of Logit and K-fold cross-validation.

Figure 2 presents the K-fold cross-validation weight placed on the probit model for $n =$ 100, 200, 500, and 1000, respectively. The model weight is calculated by averaging the K-fold cross-validation weights placed on the probit model across 5000 simulation replications. As we expected, the K-fold cross-validation weights are quite similar for $K = 2$, 5, 10, n, and data-driven K, which is consistent with the similar finite sample performance among these choices of K displayed in Figure 1. Notice that the K-fold cross-validation assigns more weights on Probit when α is small, and put less weights on Probit when α is large. Therefore, the weight assignment of the K-fold cross-validation is consistent with the relative performance between Probit and Logit shown in Figure 1.

5.2 Linear regression model

In the second simulation design, we consider a linear regression model:

$$Y_i = \sum_{j=1}^{q} \beta_j X_{ji} + e_i, \;\; \mathrm{E}(e_i|\mathbf{X}_i) = 0,$$

where $e_i \sim N(0,1)$. We generate $(X_{1i}, ..., X_{qi})^\top$ from a joint normal distribution $N(\mathbf{0}, \mathbf{\Omega})$, where the diagonal elements of $\mathbf{\Omega}$ are 1, and off-diagonal elements are 0.5. The true model has only coefficients $\beta_1, ..., \beta_p$ different from zero. We set $p = 4$ and $q = 20$, and the regression coefficients are $\boldsymbol{\beta} = (1, 1, 1, c, 0, ..., 0)^\top$, where the parameter c is varied on a grid between 0.01 and 0.99. The sample size is varied between $n = 100$, 200, 500, and 1000.

We consider two candidate models only, an incorrectly specified model and a correctly specified model. The incorrectly specified model includes the first $p - 1$ predictors in $\mathbf{X}_i$

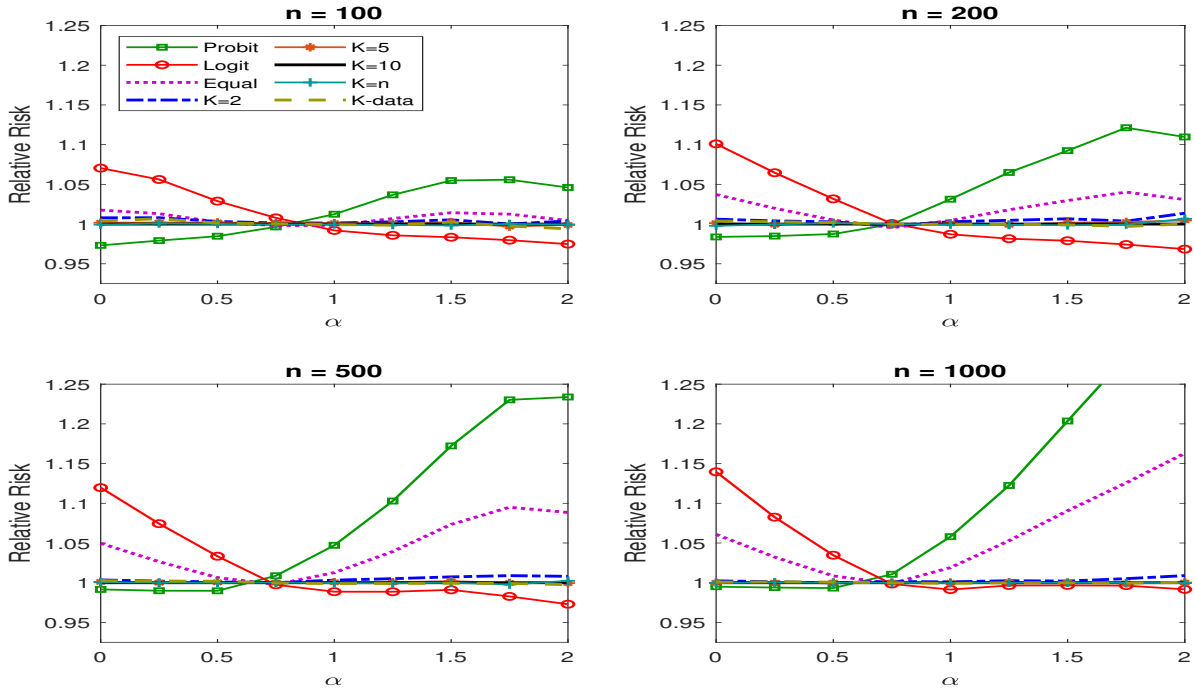

Figure 1: Relative risk in the binary choice model

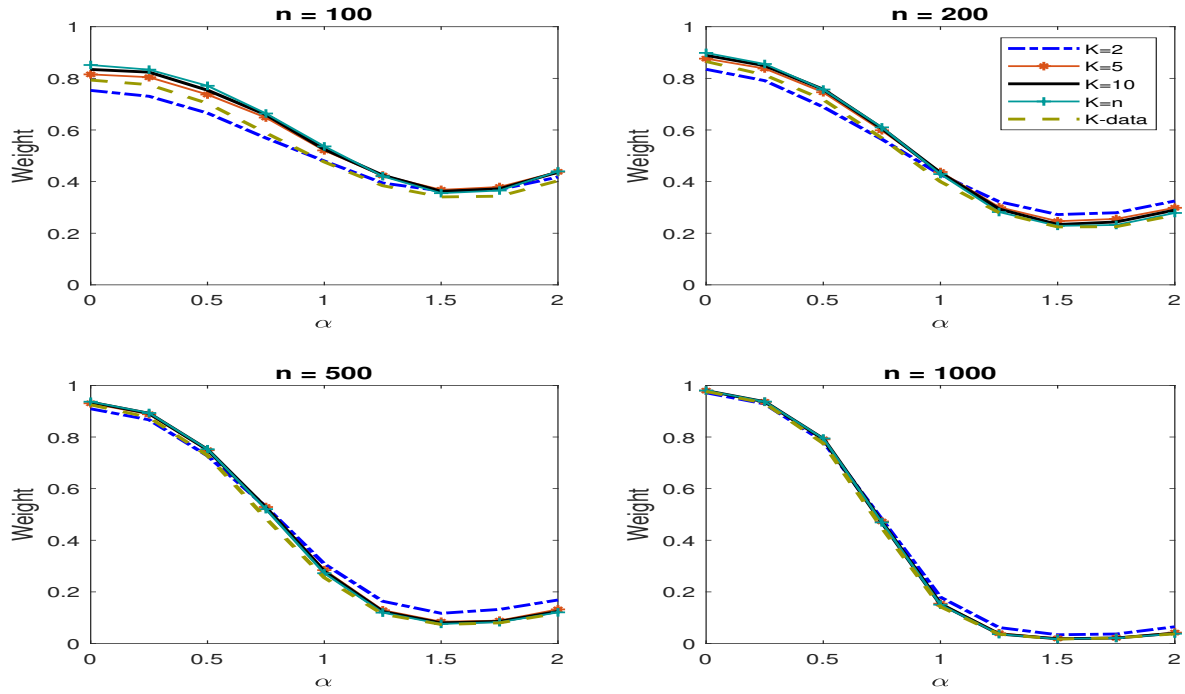

Figure 2: Model weight placed on the probit model

only, that is, $X_{1i}, ..., X_{p-1,i}$, while the correctly specified model includes all predictors in $\mathbf{X}_i$. Both models are estimated by the ordinary least squares estimator; see Example 1 of Section 2 for details. In addition to the incorrectly specified model estimator (labeled Incorrect) and the correctly specified model estimator (labeled Correct), we also consider the proposed averaging prediction using K-fold cross-validation weights with $K = 2$, 5, 10, n, and data-driven K.

We evaluate the finite sample behavior of each method based on the following empirical risk function $\frac{1}{S}\sum_{s=1}^{S}\{\widehat{Y}_{n+1}^{\{s\}}(\widehat{\mathbf{w}}^{\{s\}}) - \mathrm{E}(Y_{n+1}^{\{s\}}|X_{n+1}^{\{s\}})\}^2$, where $\mathrm{E}(Y_{n+1}^{\{s\}}|X_{n+1}^{\{s\}}) = X_{n+1}^{\{s\}^\top}\boldsymbol{\beta}$ and $\widehat{Y}_{n+1}^{\{s\}}(\widehat{\mathbf{w}}^{\{s\}}) = \sum_{m=1}^{M}\widehat{w}_m^{\{s\}}\mathbf{X}_{(m),n+1}^{\{s\}^\top}\widehat{\boldsymbol{\beta}}_{(m)}^{\{s\}}$ is the prediction based on each method in the sth replication. Like the first simulation design, the risk function is calculated by averaging across 5000 simulation replications, and is divided by the risk of 10-fold cross-validation weights.

Figure 3 presents the relative risk for $n = 100$, 200, 500, and 1000, respectively. The results show that the incorrectly specified model has smaller relative risk than the correctly specified model for small c, but larger relative risk than the correctly specified model for large c. The K-fold cross-validation methods, except $K = 2$, perform quite well and dominate both Incorrect and Correct in most situations. The 2-fold cross-validation achieves lower relative risk than Incorrect and Correct in the middle range of the parameter c, but it has larger relative risk than other K-fold cross-validation in most cases.

Figure 4 presents the K-fold cross-validation weight placed on the correctly specified model for $n = 100$, 200, 500, and 1000, respectively. The results show that the K-fold cross-validation weight placed on the correctly specified model is increasing with the parameter c, and approaching one as the sample size increases, which is consistent with Theorem 2. As we expected, the K-fold cross-validation weights are quite similar for $K = 5$, 10, n, and data-driven K, but $K = 2$ assigns less weights on the correctly specified model than other K-fold cross-validation methods.

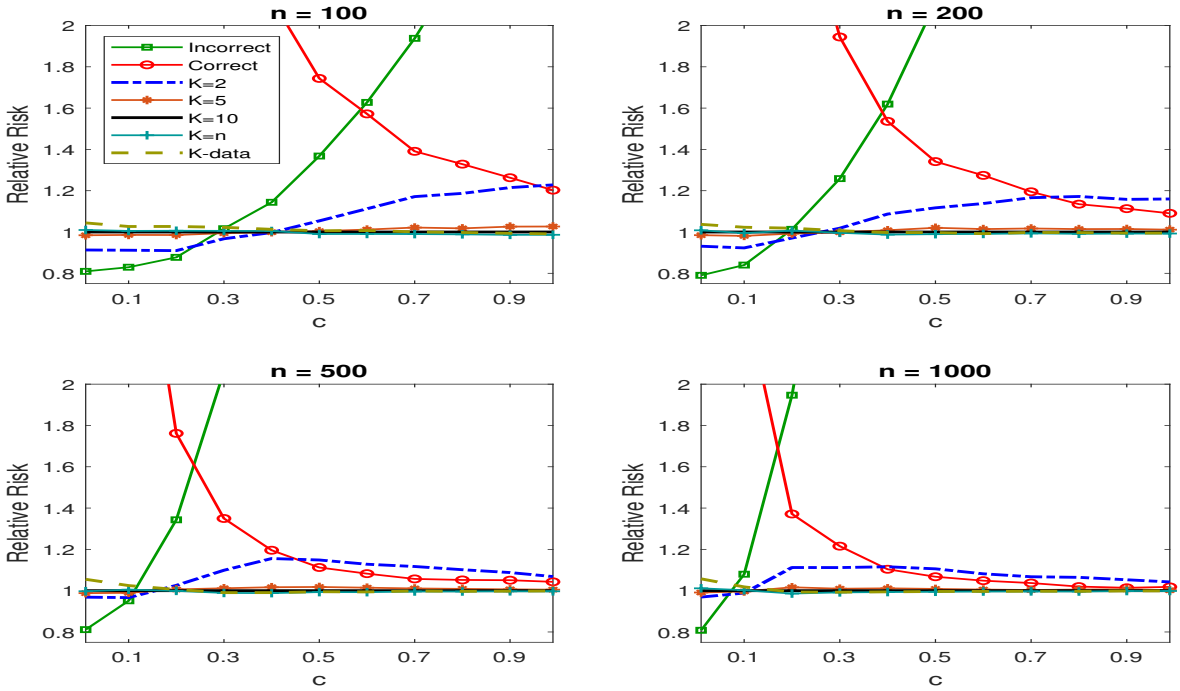

Figure 3: Relative risk in the linear regression model

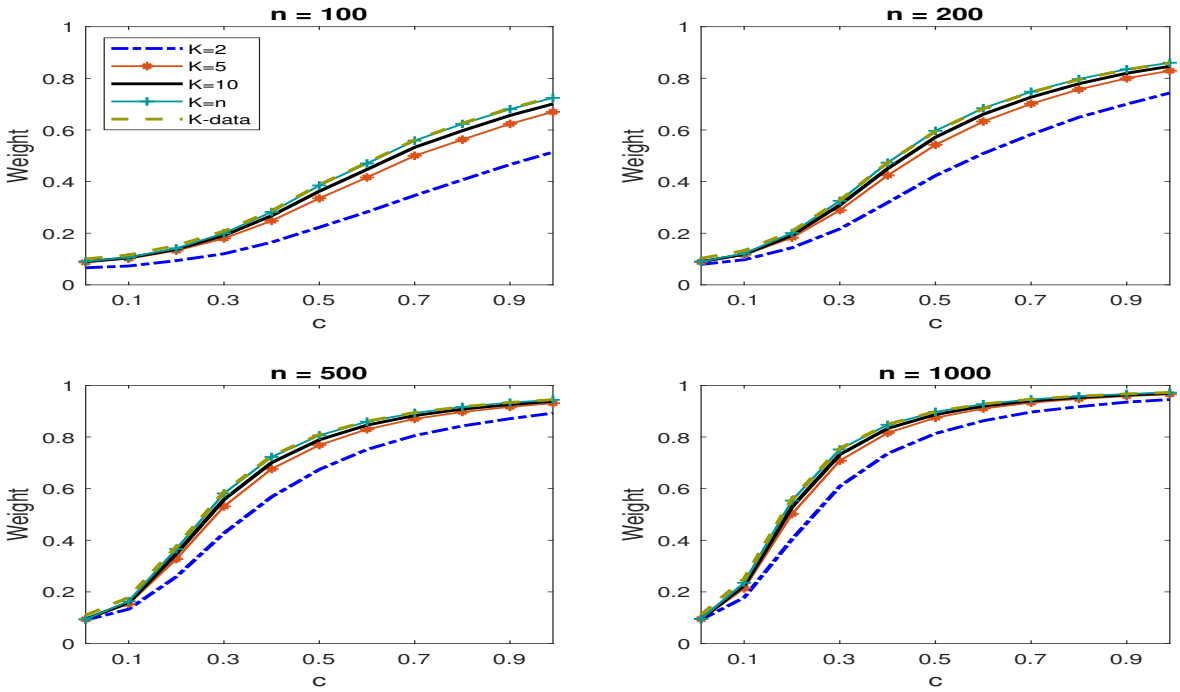

Figure 4: Model weight placed on the correctly specified model

5.3 Nonlinear regression model

In the third simulation design, we consider a nonlinear regression model:

$$Y_i = \mu(\mathbf{X}_i) + e_i = \exp\left(\sum_{j=1}^{p} \beta_j X_{ji}\right) + e_i, \;\; \mathrm{E}(e_i|\mathbf{X}_i) = 0,$$

where $X_{ji} \sim iid$ Uniform$(-1, 1)$. The error term is generated by $e_i = \sigma_i \epsilon_i$, where ϵ_i is generated from a log-normal distribution whose logarithm is normally distributed with mean zero and variance one. For the homoskedastic simulation, we set $\sigma_i = 1$, and for the heteroskedastic simulation, we set $\sigma_i^2 = 0.5 + 0.5x_{pi}^2$. The sample size is varied between $n = 100$, 200, 400, and 800.

Similar to the Example 1 of Section 2, we consider a sequence of nested candidate models, and the mth model uses the first m predictors in $\mathbf{X}_i$. That is, the mth candidate model is $Y_i = \exp\left(\sum_{j=1}^{m} \beta_j X_{ji}\right) + u_i$ for $m = 1, \ldots, M$. Three cases of the regression coefficients are studied:

$$\text{Case 1: } p = 10,\ M = 8,\ \boldsymbol{\beta} = \left(1^\delta, 2^\delta, ..., p^\delta\right)^\top,$$

$$\text{Case 2: } p = 10,\ M = 10,\ \boldsymbol{\beta} = \left(1^\delta, 2^\delta, ..., p^\delta\right)^\top,$$

$$\text{Case 3: } p = 10,\ M = 10,\ \boldsymbol{\beta} = \left(1^\delta, 2^\delta, ..., 6^\delta, 0, 0, 0, 0\right)^\top.$$

We set $\delta = -0.5$ so that the regression coefficient is a decreasing sequence. In Case 1, we exclude the last two predictors from all candidate models to study the scenario where all candidate models are misspecified. In other words, all candidate models have two omitted variables in Case 1. In Cases 2 and 3, we study the scenario where the model set includes correctly specified models. The numbers of correctly specified models are 1 and 5 for Case 2 and 3, respectively. In Case 2, only the model that includes all predictors, the 10th model, is correctly specified. In Case 3, the 6th to 10th models are correctly specified.

We consider the following estimators: (1) Akaike information criterion model selection estimator (labeled AIC); (2) Bayesian information criterion model selection estimator (labeled BIC); (3) Smoothed Akaike information criterion model averaging estimator (labeled SAIC); (4) Smoothed Bayesian information criterion model averaging estimator (labeled

SBIC); (5) Nonlinear model averaging estimator (labeled NMA); (6) Model averaging estimator with equal weights (labeled Equal); and (7) Model averaging estimator using K-fold cross-validation weights with $K = 2$, 5, 10, n, and data-driven K.

We briefly discuss each estimator. The AIC criterion for the mth model is $\text{AIC}_m = n\log(\widehat{\sigma}^2_{(m)}) + 2p_m$, where $\widehat{\sigma}^2_{(m)} = \frac{1}{n}\sum_{i=1}^n \widehat{e}^2_{(m),i}$. Here, $\widehat{e}_{(m),i}$ is the nonlinear least squares residual from the model m, and p_m is the number of parameters in the model m. The BIC criterion for the mth model is $\text{BIC}_m = n\log(\widehat{\sigma}^2_{(m)}) + \log(n)p_m$. For AIC and BIC, we select the model with the smallest AIC_m and AIC_m, respectively. The SAIC estimator is proposed by Buckland et al. (1997) and it uses the exponential AIC as the model weight. The SAIC weight is proportional to the likelihood of the model and is defined as $\widehat{w}_m = \exp(-\frac{1}{2}\text{AIC}_m)/\sum_{j=1}^M \exp(-\frac{1}{2}\text{AIC}_j)$. The SBIC estimator is a simplified form of Bayesian model averaging with diffuse priors, and the SBIC weight is defined as $\widehat{w}_m = \exp(-\frac{1}{2}\text{BIC}_m)/\sum_{j=1}^M \exp(-\frac{1}{2}\text{BIC}_j)$. The NMA is proposed by Feng et al. (2021), and is a generalization of the Mallows model averaging estimator from linear regression models to nonlinear regression models. The NMA estimator selects the model weights by minimizing a nonlinear information criterion $C(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\mu}_i(\mathbf{w}))^2 + 2\widehat{\sigma}^2 \sum_{m=1}^M w_m \pi_m$, where $\widehat{\mu}_i(\mathbf{w})$ is the model averaging estimator of $\mu(\mathbf{X}_i)$ and π_m is a bias-adjusting term based on the first derivatives of $\mu(\mathbf{X}_i)$ and the Hessian matrix. The Equal estimator assigns $\frac{1}{M}$ weight to each candidate model, and the K-fold cross-validation weights and data-driven K are described in Section 3.

We evaluate the finite sample behavior of each method based on the following empirical risk function $\frac{1}{S}\sum_{s=1}^S \{\widehat{Y}^{\{s\}}_{n+1}(\widehat{\mathbf{w}}^{\{s\}}) - \text{E}(Y^{\{s\}}_{n+1}|X^{\{s\}}_{n+1})\}^2$, where $\text{E}(Y^{\{s\}}_{n+1}|X^{\{s\}}_{n+1}) = \mu(\mathbf{X}^{\{s\}}_{n+1}) = \mu^{\{s\}}_{n+1}$ and $\widehat{Y}^{\{s\}}_{n+1}(\widehat{\mathbf{w}}^{\{s\}}) = \widehat{\mu}^{\{s\}}_{n+1}(\widehat{\mathbf{w}}^{\{s\}})$ is the prediction based on each method in the sth replication. Like the first simulation design, the risk function is calculated by averaging across 5000 simulation replications, and is divided by the risk of 10-fold cross-validation weights.

Tables 1 and 2 present the relative risk of each method for the homoskedastic and heteroskedastic setup, respectively. We first compare the finite sample performance of AIC, BIC, SAIC, and SBIC in the homoskedastic setup. In Cases 1 and 2, BIC has smaller relative risk than AIC for $n = 100$, but larger relative risk than AIC when the sample size is large. Unlike Cases 1 and 2, BIC performs better than AIC for all sample sizes in Case 3.

Table 1: Relative risk in the homoskedastic setup

	n	AIC	BIC	SAIC	SBIC	NMA	Equal	K=2	K=5	K=10	K=n	K-data
Case 1	100	2.386	2.254	2.188	1.983	1.603	0.946	1.008	1.025	1.000	1.033	1.037
	200	1.794	2.076	1.773	1.896	1.270	1.094	1.151	1.025	1.000	0.996	1.009
	400	1.457	1.493	1.452	1.473	1.113	1.236	1.082	1.019	1.000	0.996	0.990
	800	1.195	1.203	1.196	1.204	0.989	1.580	1.063	1.012	1.000	0.996	1.006
Case 2	100	3.078	2.713	2.910	2.333	1.935	0.948	1.004	0.972	1.000	0.959	1.006
	200	2.419	2.786	2.345	2.506	1.478	1.170	1.111	1.021	1.000	0.995	1.093
	400	2.109	2.184	2.101	2.138	1.320	1.703	1.194	1.018	1.000	1.008	1.052
	800	1.881	1.836	1.870	1.863	1.219	2.072	1.073	1.008	1.000	1.003	1.025
Case 3	100	4.969	2.490	4.147	2.193	2.902	1.082	0.904	0.929	1.000	1.005	0.979
	200	2.990	2.715	2.786	2.541	1.981	1.068	1.013	1.021	1.000	0.947	1.017
	400	2.526	2.136	2.428	2.074	1.631	0.893	0.987	0.979	1.000	0.996	1.009
	800	1.533	1.361	1.485	1.349	1.156	1.117	1.041	0.991	1.000	1.004	0.999

Table 2: Relative risk in the heteroskedastic setup

	n	AIC	BIC	SAIC	SBIC	NMA	Equal	K=2	K=5	K=10	K=n	K-data
Case 1	100	4.583	4.658	4.407	3.652	2.567	0.958	1.019	0.996	1.000	1.020	0.995
	200	1.587	1.712	1.570	1.593	1.190	0.998	1.009	0.997	1.000	1.007	1.011
	400	1.335	1.354	1.338	1.358	1.066	1.288	0.998	1.009	1.000	1.012	1.018
	800	1.129	1.133	1.130	1.132	0.999	1.603	0.999	1.003	1.000	1.001	1.003
Case 2	100	2.519	2.506	2.182	2.087	1.451	0.894	0.976	1.074	1.000	1.018	0.982
	200	1.908	2.201	1.926	2.042	1.263	1.268	1.162	0.999	1.000	0.964	1.005
	400	1.701	1.726	1.701	1.720	1.116	2.261	1.203	1.011	1.000	1.001	1.050
	800	1.974	1.997	1.976	1.991	1.208	2.986	1.091	1.023	1.000	1.009	1.006
Case 3	100	3.244	2.691	2.960	2.425	2.315	1.038	1.069	0.993	1.000	1.011	1.109
	200	2.471	2.360	2.342	2.081	1.897	0.961	1.013	1.005	1.000	0.999	1.027
	400	2.110	1.821	2.035	1.791	1.491	1.073	1.035	1.000	1.000	0.975	1.014
	800	1.581	1.394	1.513	1.394	1.175	1.287	1.000	0.997	1.000	0.994	1.003

For SAIC and SBIC, the pattern of relative performance between SAIC and SBIC is quite similar to that of AIC and BIC. Notice that SAIC and SBIC achieve lower relative risk than AIC and BIC in most situations, respectively, which implies that there is an efficiency gain by using the model averaging counterparts of the information criteria.

We next compare the finite sample performance of NMA, Equal, and K-fold cross-validation. Both NMA and Equal achieve lower relative risk than AIC, BIC, SAIC, and SBIC in most cases. The relative performance between NMA and Equal is mixed. NMA has better finite sample performance than Equal for $n = 400$ and 800 in Cases 1-2, but Equal

has smaller relative risk than NMA for $n = 100$ and 200 in Cases 1-2 and all sample sizes in Case 3. The proposed averaging prediction using K-fold cross-validation weights with fixed and data-driven K performs quite well, and achieves lower relative risk than other methods in most situations. Similar to the findings in the previous simulation designs, the relative risk is quite similar for K-fold cross-validation with $K = 2$, 5, 10, n, and data-driven K.

We now compare the finite sample performance of each method in the heteroskedastic setup. The relative performance of these estimators depends strongly on regression coefficients and sample sizes. Overall, the ranking of these estimators is quite similar to that in the homoskedastic setup, and K-fold cross-validation weights with fixed and data-driven K still achieve lower relative risk than other methods in most situations. Similar to the homoskedastic results, the relative risk of $K = 5$, 10, n, and data-driven K is quite similar, but the relative risk of $K = 2$ is slightly larger than that of other K-fold cross-validation weights in most cases.[3]

Figure 5 presents the sum of K-fold cross-validation weights placed on the correctly specified models in Cases 2 and 3 for the homoskedastic and heteroskedastic setup. The sum of model weights is calculated by averaging the sum of K-fold cross-validation weights placed on the correctly specified models across 5000 simulation replications. That is, we calculate the sum of weights by $\frac{1}{5000}\sum_{s=1}^{5000}\widehat{\tau}_s$, where $\widehat{\tau}_s$ is the sum of K-fold cross-validation weights given to the correctly specified model in the sth simulation replication. As shown in Figure 5, the sum of model weights is monotonically increasing and generally converges to one as the sample size increases, which is consistent with Theorem 2. The sums of K-fold cross-validation weights are quite similar for $K = 5$, 10, n, and data-driven K, but the sums of $K = 2$ are smaller than those of other K-fold cross-validation weights.

Figure 6 presents the K-fold cross-validation weights placed on the 6th to 10th models in Case 3 for the homoskedastic and heteroskedastic setups. Notice that the 6th model is the just-fitted model that has no omitted predictor and no irrelevant predictor, while the 7th to 10th models are over-fitted models that have no omitted predictor but have irrelevant predictors. The results show that the individual weights on the 6th to 10th models all

[3]The simulation results are consistent with the findings from Arlot and Lerasle (2016) that the performance of K-fold cross-validation increases significantly from $K = 2$ to $K = 5$, but is quite similar for $K \geq 5$.

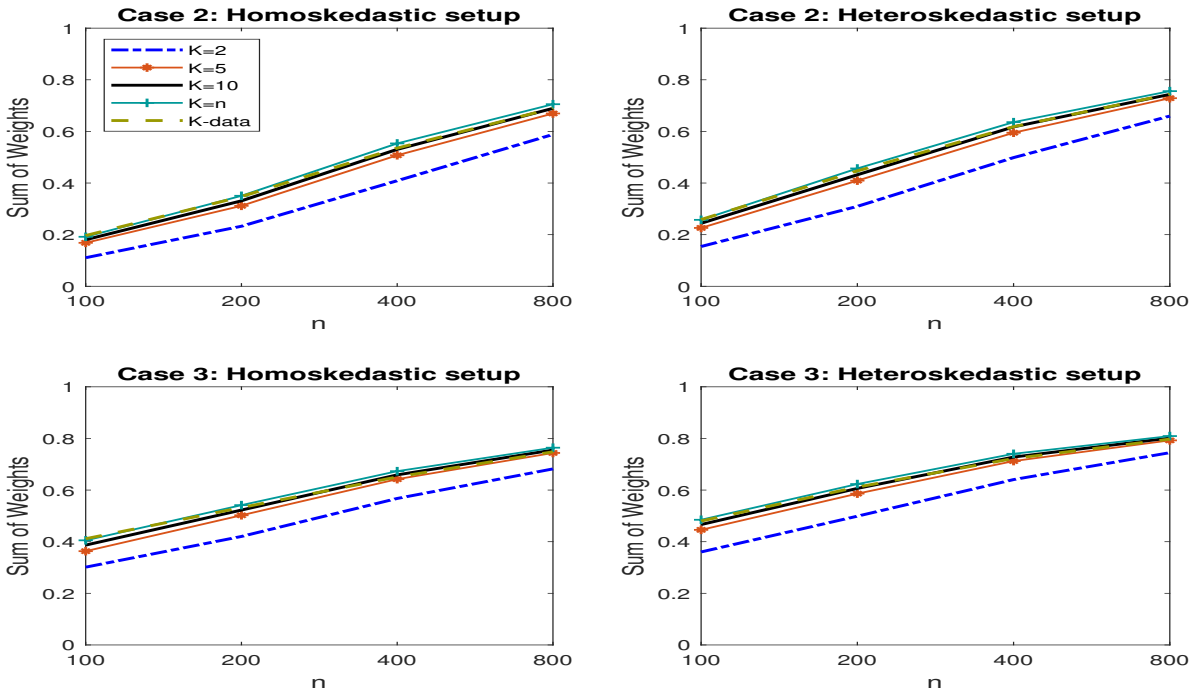

Figure 5: Sum of model weights placed on the correctly specified models

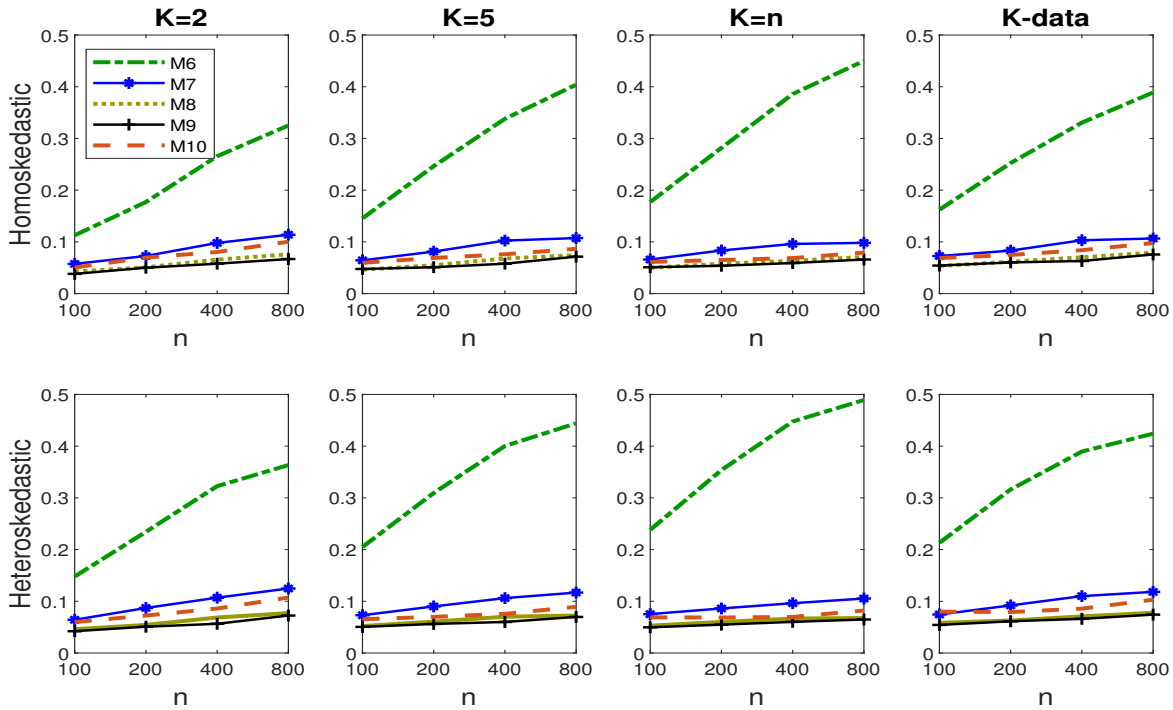

Figure 6: Model weights placed on the correctly specified models in Case 3

increase with the sample size. However, K-fold cross-validation assigns more weights to the just-fitted model than to the over-fitted models, and the weight on the just-fitted model increases much faster than those on the over-fitted models as the sample size increases. Compared with other K-fold cross-validation methods, our results show that $K = 2$ assigns less weights on the just-fitted model.

6 Empirical example

In this section, we apply the model averaging methods to credit card default prediction. We employ Yeh and Lien (2009)'s credit card clients data set to study defaulting on payment by credit card customers in Taiwan. The data consist of 30000 observations and are available at the UC Irvine Machine Learning Repository: https://archive.ics.uci.edu/ml. The dependent variable is the binary response $Y_i \in \{1, 0\}$ for defaulting on payment or not by a credit card client of a major bank in Taiwan in October 2005. There are 23 potential predictors including the amount of the given credit in NT dollars, gender (1 = female, 0 = male), education (1 = university or above, 0 = other), marital status (1 = married, 0 = other), age, history of payment status in the past six months (from April to September 2005), total amount of bill statements in NT dollars in the past six months (from April to September 2005), and total amount of previous payments in the past six months (from April to September 2005); see Yeh and Lien (2009) for a detailed description of the data. For convenience, we changed the coding for gender, education, and marital status from the original data.

We follow Yeh and Lien (2009) and Fang and Chen (2019) and estimate the probability of default by credit card clients by logistic regression. We do not impose any assumption on the distribution of the regression error. Thus, the logit model could be misspecified. Furthermore, we allow for uncertainty about the predictors in each candidate model. We consider two cases to illustrate the proposed method: (i) a set of 23 nested candidate models, and (ii) a set of 217 non-nested models. In the first case, we consider a sequence of nested candidate models, where the mth model uses the first m predictors. In the second case, we divide the predictors into two groups. The first group is the social background variables that include the amount of the given credit, gender, education, marital status, and age, and

the second group is the historical financial variables that include history of payment status, amount of bill statement, and amount of previous payment in the past six months. For the first group, we consider all possible subsets of variables with 31 possibilities. For the second group, we include these three historical financial variables sequentially in the model, that is, we include no variable, variables in the past month, variables in the past two months and so on. We then consider the interaction between these two groups, and this leads to a total of $217 = 31 \times 7$ non-nested models.

We next randomly select two samples of n_1 and n_2 observations as a training set and an evaluation set, respectively. We use observations in the training set to estimate the default probability and model parameters in each candidate model, and then apply the same model selection and model averaging methods as those in the third simulation design. We then evaluate these methods by computing their mean squared prediction error (MSPE). We follow Hansen (2008) and Hansen and Racine (2012) and use observations in the evaluation set to calculate the MSPE as follows

$$\text{MSPE} = \frac{1}{\sigma^2}\left(\frac{1}{n_2}\sum_{j=1}^{n_2}(Y_{n_1+j} - \widehat{Y}_{n_1+j}(\widehat{\mathbf{w}}))^2 - \sigma^2\right),$$

where σ^2 is estimated by the sample analogue $\widehat{\sigma}^2$ and $\widehat{Y}_{n_1+j}(\widehat{\mathbf{w}})$ is the probability prediction based on each method. As pointed out in Hansen (2008), the error variance σ^2 is the common leading term of the MSPE across all candidate models, and the scaling $\frac{1}{\sigma^2}$ is used to ensure that results are scale-free. Note that we use the same estimator $\widehat{\sigma}^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{e}^2_{(M),i}$ for all methods, and $\widehat{e}_{(M),i}$ is the logistic regression residual from the model that includes all predictors.

We repeat the above procedure 1000 times in Cases (i) and (ii), and calculate the mean of the MSPEs. For easy comparison, we divide the mean of the MSPEs of each method by those of 10-fold cross-validation weights. Thus, an entry greater than one indicates that the specified method performs worse than the model averaging estimator with 10-fold cross-validation weights.

We first consider the setting where the sample sizes n_1 and n_2 are varied between 100, 250, 500, and 1000. Tables 3 and 4 present the relative mean of the MSPEs in Cases (i) and (ii), respectively. The results show that the proposed K-fold cross-validation method achieves

Table 3: Relative mean of the MSPEs in Case (i) of the empirical example

n_1	n_2	AIC	BIC	SAIC	SBIC	NMA	Equal	K=2	K=5	K=10	K=n	K-data
250	100	1.234	1.072	1.165	1.045	1.412	1.021	1.020	1.001	1.000	0.999	1.001
	250	1.230	1.091	1.159	1.063	1.410	1.033	1.024	1.006	1.000	1.001	1.007
	500	1.225	1.090	1.158	1.057	1.423	1.027	1.022	1.001	1.000	0.999	1.005
	1000	1.229	1.094	1.157	1.060	1.411	1.028	1.020	1.001	1.000	0.999	1.005
500	100	1.148	1.041	1.089	1.022	1.283	1.084	1.031	1.007	1.000	1.000	1.003
	250	1.168	1.036	1.103	1.011	1.312	1.081	1.025	1.003	1.000	0.999	1.005
	500	1.166	1.026	1.106	1.006	1.307	1.076	1.021	1.002	1.000	0.999	1.003
	1000	1.167	1.029	1.105	1.010	1.300	1.074	1.025	1.002	1.000	1.000	1.005
1000	100	1.167	1.061	1.103	1.041	1.204	1.166	1.010	1.004	1.000	1.006	1.006
	250	1.140	1.081	1.088	1.057	1.189	1.174	1.022	1.002	1.000	1.002	1.007
	500	1.163	1.085	1.093	1.061	1.204	1.187	1.014	1.000	1.000	1.001	1.002
	1000	1.157	1.086	1.091	1.064	1.201	1.195	1.017	1.001	1.000	1.003	1.006

Table 4: Relative mean of the MSPEs in Case (ii) of the empirical example

n_1	n_2	AIC	BIC	SAIC	SBIC	NMA	Equal	K=2	K=5	K=10	K=n	K-data
250	100	1.259	1.048	1.165	1.000	1.424	0.991	1.003	0.998	1.000	0.999	1.005
	250	1.254	1.043	1.161	0.998	1.427	0.993	1.007	1.001	1.000	1.000	1.011
	500	1.265	1.052	1.169	1.002	1.449	0.997	1.014	1.000	1.000	1.000	1.010
	1000	1.263	1.051	1.169	1.005	1.448	0.999	1.014	1.000	1.000	1.001	1.011
500	100	1.188	1.081	1.104	1.031	1.337	1.013	1.019	1.004	1.000	1.001	1.005
	250	1.200	1.063	1.108	1.017	1.351	1.011	1.017	1.003	1.000	0.997	1.008
	500	1.199	1.068	1.114	1.021	1.345	1.008	1.018	1.001	1.000	1.002	1.012
	1000	1.191	1.076	1.099	1.029	1.334	1.008	1.017	1.000	1.000	1.001	1.007
1000	100	1.142	1.132	1.064	1.086	1.213	1.036	1.009	1.000	1.000	0.995	1.001
	250	1.206	1.179	1.108	1.113	1.265	1.058	1.008	0.998	1.000	1.000	1.008
	500	1.173	1.163	1.089	1.104	1.246	1.048	1.009	1.000	1.000	1.003	1.011
	1000	1.180	1.178	1.090	1.113	1.255	1.051	1.013	1.001	1.000	0.999	1.007

a lower mean of the MSPEs than other model selection and model averaging methods in most scenarios. Furthermore, the prediction performance of K-fold cross-validation is quite similar for $K = 2$, 5, 10, n, and data-driven K. Comparing the results between Case (i) and Case (ii), we find that the relative mean of the MSPEs of AIC, BIC, and NMA in Case (ii) are slightly larger than those of AIC, BIC, and NMA in Case (i).

We next consider the setting where the sample sizes n_1 and n_2 are varied between 2500, 5000, 10000, and 15000. Due to the heavy computational burden, we only compute the proposed model averaging estimator with $K = 2$, 5, and 10, but not $K = n$ and data-driven

Table 5: Relative mean of the MSPEs for larger n_1 and n_2 in Case (i)

n_1	n_2	AIC	BIC	SAIC	SBIC	NMA	Equal	K=2	K=5	K=10
5000	2500	1.146	1.464	1.094	1.441	1.035	2.207	1.034	1.005	1.000
	5000	1.131	1.413	1.088	1.395	1.033	2.088	1.028	1.004	1.000
	10000	1.162	1.533	1.110	1.503	1.040	2.363	1.039	1.006	1.000
	15000	1.133	1.437	1.087	1.414	1.030	2.150	1.028	1.003	1.000
10000	2500	1.108	1.950	1.079	1.894	1.004	3.772	1.040	1.005	1.000
	5000	1.099	1.879	1.070	1.830	1.006	3.580	1.041	1.003	1.000
	10000	1.121	2.101	1.085	2.037	1.003	4.194	1.047	1.004	1.000
	15000	1.099	1.881	1.071	1.829	1.006	3.599	1.042	1.005	1.000
15000	2500	1.079	1.972	1.060	1.881	0.987	4.414	1.033	1.003	1.000
	5000	1.070	2.024	1.049	1.940	0.981	4.543	1.040	1.004	1.000
	10000	1.081	2.108	1.059	2.006	0.980	4.864	1.039	1.003	1.000
	15000	1.094	2.284	1.067	2.172	0.980	5.499	1.045	1.003	1.000

Table 6: Relative mean of the MSPEs for larger n_1 and n_2 in Case (ii)

n_1	n_2	AIC	BIC	SAIC	SBIC	NMA	Equal	K=2	K=5	K=10
5000	2500	1.089	1.316	1.031	1.233	1.044	1.483	1.016	1.000	1.000
	5000	1.104	1.390	1.035	1.287	1.050	1.584	1.018	1.003	1.000
	10000	1.102	1.381	1.034	1.280	1.047	1.573	1.021	1.001	1.000
	15000	1.113	1.416	1.037	1.304	1.052	1.640	1.027	1.004	1.000
10000	2500	1.062	1.466	1.028	1.359	0.990	2.200	1.031	1.007	1.000
	5000	1.076	1.500	1.031	1.390	0.985	2.340	1.029	1.003	1.000
	10000	1.074	1.538	1.032	1.418	0.988	2.436	1.041	1.005	1.000
	15000	1.071	1.513	1.029	1.399	0.985	2.321	1.029	1.006	1.000
15000	2500	1.030	1.412	1.016	1.335	0.964	2.760	1.031	1.002	1.000
	5000	1.033	1.447	1.017	1.369	0.968	2.840	1.038	1.005	1.000
	10000	1.034	1.465	1.017	1.381	0.960	2.948	1.043	1.006	1.000
	15000	1.031	1.443	1.016	1.361	0.966	2.880	1.039	1.005	1.000

K. Tables 5 and 6 present the relative mean of the MSPEs for larger n_1 and n_2 in Cases (i) and (ii), respectively. Unlike the results in Tables 3 and 4, AIC and SAIC perform better than BIC, SBIC, and Equal. Both NMA and K-fold cross-validation have better performance than other methods in most cases. Similar to the findings in the third simulation design, the prediction performance of 5-fold and 10-fold cross-validation is quite similar, but the mean of the MSPEs of $K = 2$ are slightly larger than those of $K = 5$ and $K = 10$.

7 Conclusion

In this paper, we study the model averaging prediction in a quasi-likelihood framework that allows for parameter uncertainty and model misspecification. We propose a K-fold cross-validation to select the data-driven weights for a large number of candidate models. The proposed method is asymptotically optimal for the case where all candidate models are misspecified and has the consistency property for the case where some candidate models are correctly specified. The simulation and empirical results show that the proposed model averaging prediction using K-fold cross-validation weights with fixed and data-driven K generally achieves lower empirical risk than other existing methods. Furthermore, the finite sample performance of the proposed method is not sensitive to the choice of K. While this paper has focused on the optimality of the quadratic prediction risk function, it would be greatly desirable to study the optimal model averaging prediction in other risk functions, for example, Kullback-Leibler divergence or Bregman divergence. Another possible extension would be to extend the proposed method to the model setting with a diverging number of parameters.

Acknowledgments

We thank the editor Xiaohong Chen, the associate editor, and two referees for many constructive comments and suggestions. We also thank the seminar participants of University of California, Riverside and National Taiwan University, and conference participants of MMW 2020, TES 2021, and IEIDC 2021 for their discussions and suggestions. Xinyu Zhang gratefully acknowledges research support from the National Natural Science Foundation of China (71925007, 72091212, 71988101, 11688101), the CAS Project for Young Scientists in Basic Research (YSBR-008), and the Beijing Academy of Artificial Intelligence. Chu-An Liu gratefully acknowledges research support from the Academia Sinica Career Development Award (AS-CDA-110-H02) and the Ministry of Science and Technology of Taiwan (MOST 110-2410-H-001-081-MY3). All errors remain the authors'.

Appendix

A Computing time of K-fold cross-validation

Table A1 presents the computing time of K-fold cross-validation per replication in three simulation designs. The computing time is measured in seconds and varies from 0.0007 to 83.6522 for different numbers of groups K and sample size n. As we expected, the computing time increases with K or n, and it increases rapidly when both K and n are large.

Table A1: Computing time measured in seconds of K-fold cross-validation

	Simulation design I				Simulation design II				Simulation design III			
	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=400	n=800
K=2	0.0067	0.0086	0.0117	0.0132	0.0007	0.0010	0.0015	0.0017	0.0677	0.0754	0.0948	0.1295
K=5	0.0128	0.0169	0.0235	0.0259	0.0008	0.0011	0.0016	0.0021	0.1361	0.1401	0.2397	0.2657
K=10	0.0231	0.0304	0.0428	0.0470	0.0010	0.0015	0.0020	0.0027	0.2502	0.2421	0.4430	0.4939
K=n	0.2102	0.5897	1.9307	4.2536	0.0034	0.0101	0.0342	0.1128	2.2852	5.1944	16.1263	37.0056
K-data	0.5703	1.4071	4.7395	9.4370	0.0135	0.0321	0.0897	0.2551	6.2298	13.1001	40.1847	83.6522

Notes: The simulations are conducted in MATLAB R2021a by a desktop computer with an Intel(R) Core(TM) i9-9980XE CPU (3.00GHz) with 128GB memory.

B Verifications of Assumptions 1 and 5 in Example 1

We first verify Assumption 1 in Example 1. Let $\mu_i = \sum_{j=1}^{\infty} \beta_j X_{ji}$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$, and $\mathbf{e} = (e_1, \ldots, e_n)^\top$. Let p_m denote the dimension of $\mathbf{X}_m$ and $p_{max} = \max_m p_m$. We consider the following primitive conditions: (i) $\lambda_{min}(n^{-1}\mathbf{X}_{(m)}^\top \mathbf{X}_{(m)}) \geq c_{min}$ almost surely for $m = 1, \ldots, M$, where $\lambda_{min}(\cdot)$ is the minimum eigenvalue of a matrix and c_{min} is a positive constant; (ii) $n^{1/2}(\widehat{\boldsymbol{\beta}}_{(m)} - \boldsymbol{\beta}_{(m)}^*)$ converges to a normal distribution for $m = 1, \ldots, M$, where $\boldsymbol{\beta}_{(m)}^*$ is a p_m-dimensional vector. Then, it follows that for any $\delta > 0$,

$$
\begin{aligned}
& \operatorname{pr}\left\{ n^{1/2} M^{-1/2} \max_m \|(n^{-1}\mathbf{X}_{(m)}^\top \mathbf{X}_{(m)})^{-1} n^{-1} \mathbf{X}_{(m)}^\top \mathbf{e}\| > \delta \right\} \\
\leq & \sum_{m=1}^{M} \operatorname{pr}\left\{ n^{1/2} M^{-1/2} \|(n^{-1}\mathbf{X}_{(m)}^\top \mathbf{X}_{(m)})^{-1} n^{-1} \mathbf{X}_{(m)}^\top \mathbf{e}\| > \delta \right\}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{m=1}^{M} M^{-1} n \delta^{-2} \mathrm{E} \|(n^{-1}\mathbf{X}_{(m)}^{\top}\mathbf{X}_{(m)})^{-1} n^{-1}\mathbf{X}_{(m)}^{\top}\mathbf{e}\|^2 \\
&= \sum_{m=1}^{M} M^{-1}\delta^{-2}\sigma^2 \mathrm{E}\left\{\mathrm{trace}(n^{-1}\mathbf{X}_{(m)}^{\top}\mathbf{X}_{(m)})^{-1}\right\} \\
&\leq c_{min}^{-1} p_{max}\sigma^2\delta^{-2}. \qquad \text{(B.1)}
\end{aligned}
$$

Let $\boldsymbol{\zeta}_{(m)} = (n^{-1}\mathbf{X}_{(m)}^{\top}\mathbf{X}_{(m)})^{-1} n^{-1}\mathbf{X}_{(m)}^{\top}\boldsymbol{\mu} - \boldsymbol{\beta}_{(m)}^{*}$ for $m = 1, \ldots, M$. Then, we have

$$
\widehat{\boldsymbol{\beta}}_{(m)} - \boldsymbol{\beta}_{(m)}^{*} = (n^{-1}\mathbf{X}_{(m)}^{\top}\mathbf{X}_{(m)})^{-1} n^{-1}\mathbf{X}_{(m)}^{\top}\mathbf{e} + \boldsymbol{\zeta}_{(m)}. \qquad \text{(B.2)}
$$

Hence, we have $n^{1/2}\boldsymbol{\zeta}_{(m)} = O_p(1)$. We further assume that $\mathrm{E}\|n^{1/2}\boldsymbol{\zeta}_{(m)}\|^2 \leq c p_m$ almost surely, where c is a positive constant. Similar to (B.1), we have

$$
\mathrm{pr}\left\{n^{1/2}M^{-1/2}\max_{m}\|\boldsymbol{\zeta}_{(m)}\| > \delta\right\} \leq c p_{max}\delta^{-2}, \qquad \text{(B.3)}
$$

which, along with (B.1) and (B.2), implies Assumption 1. For simplicity purposes, we can set $\boldsymbol{\theta} = \boldsymbol{\beta}$ in Example 1, since the estimator $\widehat{\sigma}_{(m)}^2$ is not used in the prediction.

We next verify Assumption 5 in a specific setup of Example 1. In particular, we assume that $\mathbf{X}_i$ follows a joint normal distribution with mean zeros, $\mathrm{cov}(X_{j_1 i}, X_{j_2 i}) = 0$, and $\mathrm{var}(X_{ji}) \geq c_1$ for any positive integer j, j_1, and j_2, where c_1 is a positive constant. We further assume that the variables used in candidate models belong to $\{X_{1i}, \ldots, X_{j^* i}\}$, where j^* is a positive integer, and $\sum_{j=j^*+1}^{\infty}\beta_j^2 \geq c_2$, where c_2 is a positive constant.

Let $\mathbf{X}_{(m),n+1}$ be the new observation for the mth candidate model. Then, for any $\mathbf{w} \in \mathcal{W}$,

$$
\begin{aligned}
R^*(\mathbf{w}) &= \mathrm{E}\left[\{Y_{n+1}^*(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2\right] \\
&= \mathrm{E}\left[\left\{\sum_{m=1}^{M} w_m \mathbf{X}_{(m),n+1}^{\top}\boldsymbol{\beta}_{(m)}^{*} - \sum_{j=1}^{\infty}\beta_j X_{j,n+1}\right\}^2\right] \\
&\geq \mathrm{E}\left(\sum_{j=j^*+1}^{\infty}\beta_j X_{ji}\right)^2 \\
&\geq c_1 c_2. \qquad \text{(B.4)}
\end{aligned}
$$

Thus, the sufficient condition for Assumption 5 is $M = o(n^{1/2})$. In practice, when the number of potential candidate models is very large, we could conduct a model screening step to reduce the number of candidate models.

C Proofs of the theorems

The following lemma, which is Lemma 1 in Zhang (2010) and Lemma 1 in Gao et al. (2019), will be used in the proof of Theorem 1.

Lemma 1. *(Zhang, 2010; Gao et al., 2019) Let*

$$\widetilde{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{W}}{argmin}\left\{R(\mathbf{w}) + a_n(\mathbf{w}) + b_n\right\},$$

where $a_n(\mathbf{w})$ is a term related to $\mathbf{w}$, and b_n is a term unrelated to $\mathbf{w}$. If

$$\sup_{\mathbf{w}\in\mathcal{W}} |a_n(\mathbf{w})|/R^*(\mathbf{w}) = o_p(1), \ \sup_{\mathbf{w}\in\mathcal{W}} |R^*(w) - R(w)|/R^*(\mathbf{w}) = o_p(1),$$

and there exists a constant c and a positive integer N^* so that when $n \geq N^*$, $\inf_{\mathbf{w}\in\mathcal{W}} R^*(\mathbf{w}) \geq c > 0$ almost surely, then $R(\widetilde{\mathbf{w}})/\inf_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w}) \to 1$ in probability.

Proof of Theorem 1: Let $CV_K^*(\mathbf{w}) = CV_K(\mathbf{w}) - \{\mathbf{Y} - \mathrm{E}(\mathbf{Y}|\mathbf{X})\}^\top \{\mathbf{Y} + \mathrm{E}(\mathbf{Y}|\mathbf{X})\}$, where the second term is unrelated to $\mathbf{w}$. Therefore,

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{W}}{\mathrm{argmin}} CV_K(\mathbf{w}) = \underset{\mathbf{w}\in\mathcal{W}}{\mathrm{argmin}} CV_K^*(\mathbf{w}).$$

According to Lemma 1, Theorem 1 is valid if the following hold:

$$\sup_{\mathbf{w}\in\mathcal{W}} \frac{|R(\mathbf{w}) - R^*(\mathbf{w})|}{R^*(\mathbf{w})} = o(1) \tag{C.1}$$

and

$$\sup_{\mathbf{w}\in\mathcal{W}} \frac{|CV_K^*(\mathbf{w})/n - R^*(\mathbf{w})|}{R^*(\mathbf{w})} = o_p(1). \tag{C.2}$$

We first consider (C.1). Observe that

$$\begin{aligned}
& \xi_n^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \{\widehat{Y}_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2 - \{Y_{n+1}^*(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2 \right| \\
= \ & \xi_n^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \{\widehat{Y}_{n+1}(\mathbf{w}) - Y_{n+1}^*(\mathbf{w})\}\{\widehat{Y}_{n+1}(\mathbf{w}) - Y_{n+1}^*(\mathbf{w}) + 2Y_{n+1}^*(\mathbf{w}) - 2\mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}\} \right| \\
= \ & \xi_n^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \left\{ \sum_{m=1}^{M} w_m \left(\widehat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{(m)}^* \right)^\top \frac{\partial \widehat{Y}_{(m),n+1}}{\partial \widehat{\boldsymbol{\theta}}_{(m)}} \, |_{\widehat{\boldsymbol{\theta}}_{(m)} = \boldsymbol{\theta}_{(m)}^\star} \right\} \right. \\
& \left. \times \left\{ \sum_{m=1}^{M} w_m \left(\widehat{\boldsymbol{\theta}}_{(m)} - \boldsymbol{\theta}_{(m)}^* \right)^\top \frac{\partial \widehat{Y}_{(m),n+1}}{\partial \widehat{\boldsymbol{\theta}}_{(m)}} \, |_{\widehat{\boldsymbol{\theta}}_{(m)} = \boldsymbol{\theta}_{(m)}^\star} + 2 \sum_{m=1}^{M} w_m Y_{(m),n+1}^* - 2\mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1}) \right\} \right|
\end{aligned}$$

$$
\begin{aligned}
&= \xi_n^{-1} O_p(n^{-1/2} M^{1/2}) \\
&= o_p(1), \qquad \text{(C.3)}
\end{aligned}
$$

where $\boldsymbol{\theta}^{\star}_{(m)}$ is in $\mathcal{O}(\boldsymbol{\theta}^{*}_{(m)}, \varrho)$, the second equality uses Assumption 2, the third equality uses Assumptions 1 and 2, and the fourth equality uses Assumption 5.

Hence, we have

$$
\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}} \frac{|R(\mathbf{w}) - R^*(\mathbf{w})|}{R^*(\mathbf{w})} \\
\leq\ & \xi_n^{-1} \sup_{\mathbf{w}\in\mathcal{W}} |R(\mathbf{w}) - R^*(\mathbf{w})| \\
=\ & \xi_n^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \mathrm{E}\left[\{\widehat{Y}_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2 - \{Y^*_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2 \right] \right| \\
\leq\ & \mathrm{E}\left(\xi_n^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \{\widehat{Y}_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2 - \{Y^*_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2 \right| \right) \\
=\ & o(1), \qquad \text{(C.4)}
\end{aligned}
$$

where the third step uses Assumption 3 and the last step is due to (C.3). Therefore, we obtain (C.1).

We next consider (C.2). Let $\mathbf{Y}^*(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{Y}^*_{(m)}$ and $\mathbf{Y}^*_{(m)} = (Y^*_{(m),1}, \ldots, Y^*_{(m),n})^\top$, where $Y^*_{(m),i}$ is the prediction of Y_i calculated based on the limiting value $\boldsymbol{\theta}^*_{(m)}$. Observe that

$$
\begin{aligned}
&|CV^*_K(\mathbf{w})/n - R^*(\mathbf{w})| \\
=\ & \left| \left\{ \|\widetilde{\mathbf{Y}}(\mathbf{w}) - \mathbf{Y}\|^2 - \{\mathbf{Y} - \mathrm{E}(\mathbf{Y}|\mathbf{X})\}^\top \{\mathbf{Y} + \mathrm{E}(\mathbf{Y}|\mathbf{X})\} \right\} /n - \mathrm{E}\left[\{Y^*_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2 \right] \right| \\
\leq\ & \left| \left\{ \|\mathbf{Y}^*(\mathbf{w}) - \mathbf{Y}\|^2 - \{\mathbf{Y} - \mathrm{E}(\mathbf{Y}|\mathbf{X})\}^\top \{\mathbf{Y} + \mathrm{E}(\mathbf{Y}|\mathbf{X})\} \right\} /n - \mathrm{E}\left[\{Y^*_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2 \right] \right| \\
&+ \left| \|\widetilde{\mathbf{Y}}(\mathbf{w}) - \mathbf{Y}\|^2 - \|\mathbf{Y}^*(\mathbf{w}) - \mathbf{Y}\|^2 \right| /n \\
=\ & \Big| \Big[\|\mathbf{Y}^*(\mathbf{w}) - \mathrm{E}(\mathbf{Y}|\mathbf{X}) - \mathbf{Y} + \mathrm{E}(\mathbf{Y}|\mathbf{X})\|^2 - \{\mathbf{Y} - \mathrm{E}(\mathbf{Y}|\mathbf{X})\}^\top \{\mathbf{Y} + \mathrm{E}(\mathbf{Y}|\mathbf{X})\} \Big] /n \\
&- \mathrm{E}\left[\{Y^*_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2 \right] \Big| + \left| \|\widetilde{\mathbf{Y}}(\mathbf{w}) - \mathbf{Y}\|^2 - \|\mathbf{Y}^*(\mathbf{w}) - \mathbf{Y}\|^2 \right| /n \\
\leq\ & \left| \|\mathbf{Y}^*(\mathbf{w}) - \mathrm{E}(\mathbf{Y}|\mathbf{X})\|^2 /n - \mathrm{E}\left[\{Y^*_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2 \right] \right| \\
&+ 2 \left| \mathbf{Y}^*(\mathbf{w})^\top \{\mathbf{Y} - \mathrm{E}(\mathbf{Y}|\mathbf{X})\} \right| /n + \left| \|\widetilde{\mathbf{Y}}(\mathbf{w}) - \mathbf{Y}\|^2 - \|\mathbf{Y}^*(\mathbf{w}) - \mathbf{Y}\|^2 \right| /n. \qquad \text{(C.5)}
\end{aligned}
$$

Similar to (C.3), by Assumptions 1 and 2, we have

$$
\sup_{\mathbf{w}\in\mathcal{W}} \left| \|\widetilde{\mathbf{Y}}(\mathbf{w}) - \mathbf{Y}\|^2 - \|\mathbf{Y}^*(\mathbf{w}) - \mathbf{Y}\|^2 \right|
$$

$$
\begin{aligned}
&= \sup_{\mathbf{w}\in\mathcal{W}}\left|\sum_{k=1}^{K}\sum_{j=1}^{J}\left[\{\widetilde{Y}_j^{[-k]}(\mathbf{w})-Y_{(k-1)\times J+j}\}^2-\{Y^*_{(k-1)\times J+j}(\mathbf{w})-Y_{(k-1)\times J+j}\}^2\right]\right| \\
&= \sup_{\mathbf{w}\in\mathcal{W}}\left|\sum_{k=1}^{K}\sum_{j=1}^{J}\{\widetilde{Y}_j^{[-k]}(\mathbf{w})-Y^*_{(k-1)\times J+j}(\mathbf{w})\}\{\widetilde{Y}_j^{[-k]}(\mathbf{w})+Y^*_{(k-1)\times J+j}(\mathbf{w})-2Y_{(k-1)\times J+j}\}\right| \\
&= \sup_{\mathbf{w}\in\mathcal{W}}\left|\sum_{k=1}^{K}\sum_{j=1}^{J}\left\{\sum_{m=1}^{M}w_m(\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}-\boldsymbol{\theta}^*_{(m)})^\top\frac{\partial\widetilde{Y}^{[-k]}_{j,(m)}}{\partial\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}}\Big|_{\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}=\boldsymbol{\theta}^\star_{(m),k,j}}\right\}\left\{\sum_{m=1}^{M}w_m(\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}-\boldsymbol{\theta}^*_{(m)})^\top\right.\right. \\
&\quad\left.\left.\times\frac{\partial\widetilde{Y}^{[-k]}_{j,(m)}}{\partial\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}}\Big|_{\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}=\boldsymbol{\theta}^\star_{(m),k,j}}+2\sum_{m=1}^{M}w_mY^*_{(m),(k-1)\times J+j}-2\mathrm{E}(Y_{(k-1)\times J+j}|\mathbf{X}_{(k-1)\times J+j})-2\varepsilon_{(k-1)\times J+j}\right\}\right| \\
&= O_p\left(\sup_{\mathbf{w}\in\mathcal{W}}\sum_{k=1}^{K}\sum_{j=1}^{J}\left\{\sum_{m=1}^{M}w_m\left\|\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}-\boldsymbol{\theta}^*_{(m)}\right\|\left\|\frac{\partial\widetilde{Y}^{[-k]}_{j,(m)}}{\partial\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}}\Big|_{\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}=\boldsymbol{\theta}^\star_{(m),k,j}}\right\|\right\}\right. \\
&\quad\times\left\{\sum_{m=1}^{M}w_m\left\|\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}-\boldsymbol{\theta}^*_{(m)}\right\|\left\|\frac{\partial\widetilde{Y}^{[-k]}_{j,(m)}}{\partial\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}}\Big|_{\widehat{\boldsymbol{\theta}}^{[-k]}_{(m)}=\boldsymbol{\theta}^\star_{(m),k,j}}\right\|+\sum_{m=1}^{M}w_m|Y^*_{(m),(k-1)\times J+j}|\right. \\
&\quad\left.\left.+|\mathrm{E}(Y_{(k-1)\times J+j}|\mathbf{X}_{(k-1)\times J+j})|+|\varepsilon_{(k-1)\times J+j}|\right\}\right) \\
&= O_p(M)+O_p(n^{1/2}M^{1/2}),
\end{aligned}
\tag{C.6}
$$

where $\boldsymbol{\theta}^\star_{(m),1,1},\ldots,\boldsymbol{\theta}^\star_{(m),K,J}$ are all in $\mathcal{O}(\boldsymbol{\theta}^*_{(m)},\varrho)$ defined in Assumption 2.

Then, it follows that for any $\delta>0$,

$$
\begin{aligned}
&\mathrm{pr}\left\{\xi_n^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|\|\mathbf{Y}^*(\mathbf{w})-\mathrm{E}(\mathbf{Y}|\mathbf{X})\|^2/n-\mathrm{E}\left[\{Y^*_{n+1}(\mathbf{w})-\mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2\right]\right|>\delta\right\} \\
\le\ &\mathrm{pr}\left\{\xi_n^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\sum_{m=1}^{M}\sum_{m'=1}^{M}w_mw_{m'}\left|\frac{1}{n}\sum_{i=1}^{n}\{Y^*_{(m),i}-\mathrm{E}(Y_i|\mathbf{X}_i)\}\{Y^*_{(m'),i}-\mathrm{E}(Y_i|\mathbf{X}_i)\}\right.\right. \\
&\qquad\left.\left.-\mathrm{E}\left[\{Y^*_{(m),i}-\mathrm{E}(Y_i|\mathbf{X}_i)\}\{Y^*_{(m'),i}-\mathrm{E}(Y_i|\mathbf{X}_i)\}\right]\right|>\delta\right\} \\
\le\ &\sum_{m=1}^{M}\sum_{m'=1}^{M}\mathrm{pr}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\{Y^*_{(m),i}-\mathrm{E}(Y_i|\mathbf{X}_i)\}\{Y^*_{(m'),i}-\mathrm{E}(Y_i|\mathbf{X}_i)\}\right.\right. \\
&\qquad\left.\left.-\mathrm{E}\left[\{Y^*_{(m),i}-\mathrm{E}(Y_i|\mathbf{X}_i)\}\{Y^*_{(m'),i}-\mathrm{E}(Y_i|\mathbf{X}_i)\}\right]\right|>\xi_n\delta\right\} \\
\le\ &\xi_n^{-2}\delta^{-2}n^{-1}\sum_{m=1}^{M}\sum_{m'=1}^{M}\mathrm{var}\left[\{Y^*_{(m),i}-\mathrm{E}(Y_i|\mathbf{X}_i)\}\{Y^*_{(m'),i}-\mathrm{E}(Y_i|\mathbf{X}_i)\}\right],
\end{aligned}
\tag{C.7}
$$

where the second inequality uses Boole's inequality and the third inequality uses Chebyshev's Inequality. Similar to (C.7), it follows that

$$
\mathrm{pr}\left\{\xi_n^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{1}{n}\sum_{i=1}^{n}\left[Y_i^*(\mathbf{w})\left\{Y_i-\mathrm{E}(Y_i|\mathbf{X}_i)\right\}\right]\right|>\delta\right\}
$$

$$
\begin{aligned}
&= \operatorname{pr}\left\{\sup_{\mathbf{w}\in\mathcal{W}}\left|\sum_{m=1}^{M} w_m \frac{1}{n}\sum_{i=1}^{n}\left[Y^*_{(m),i}\left\{Y_i - \mathrm{E}(Y_i|\mathbf{X}_i)\right\}\right]\right| > \xi_n\delta\right\} \\
&\leq \sum_{m=1}^{M}\operatorname{pr}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\left\{Y^*_{(m),i}Y_i - Y^*_{(m),i}\mathrm{E}(Y_i|\mathbf{X}_i)\right\}\right| > \xi_n\delta\right\} \\
&\leq \xi_n^{-2}\delta^{-2}n^{-1}\sum_{m=1}^{M}\operatorname{var}\left\{Y^*_{(m),i}Y_i - Y^*_{(m),i}\mathrm{E}(Y_i|\mathbf{X}_i)\right\}. \qquad \text{(C.8)}
\end{aligned}
$$

By (C.5)-(C.8) and Assumptions 4 and 5, we obtain (C.2). This completes the proof. □

Proof of Theorem 2: Following an argument similar to that in (C.7), for any $\delta > 0$, we have

$$
\begin{aligned}
&\operatorname{pr}\left\{M^{-1}n^{1/2}\sup_{\mathbf{w}\in\mathcal{W}}\left|\|\mathbf{Y}^*(\mathbf{w}) - \mathrm{E}(\mathbf{Y}|\mathbf{X})\|^2/n - \mathrm{E}\left[\{Y^*_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2\right]\right| > \delta\right\} \\
&= O(\delta^{-2}). \qquad \text{(C.9)}
\end{aligned}
$$

Hence, it follows that

$$
\sup_{\mathbf{w}\in\mathcal{W}}\left|\|\mathbf{Y}^*(\mathbf{w}) - \mathrm{E}(\mathbf{Y}|\mathbf{X})\|^2/n - \mathrm{E}\left[\{Y^*_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2\right]\right| = O_p(n^{-1/2}M). \quad \text{(C.10)}
$$

Similar to (C.10), we have

$$
\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{1}{n}\sum_{i=1}^{n}\left[Y_i^*(\mathbf{w})\left\{Y_i - \mathrm{E}(Y_i|\mathbf{X}_i)\right\}\right]\right| = O_p(n^{-1/2}M^{1/2}). \qquad \text{(C.11)}
$$

Note that for any risk function of $\widehat{\mathbf{w}}$ such as $R^*(\widehat{\mathbf{w}})$, we take the expectation in $R^*(\cdot)$ first, then plug in $\widehat{\mathbf{w}}$. Therefore, by (C.5), (C.6), (C.10), (C.11), we obtain

$$
CV_K^*(\widehat{\mathbf{w}})/n = R^*(\widehat{\mathbf{w}}) + O_p\left(n^{-1/2}M\right). \qquad \text{(C.12)}
$$

Let $\tau = \sum_{m\in\mathcal{D}} w_m$ and let $\boldsymbol{\lambda}$ be a weight vector with $\lambda_m = 0$ for $m \in \mathcal{D}$ and $\lambda_m = w_m/(1-\tau)$ for $m \notin \mathcal{D}$. For any correctly specified model $m \in \mathcal{D}$, it is easy to see that

$$
Y^*_{(m),n+1} - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1}) = 0. \qquad \text{(C.13)}
$$

Then, we have

$$
R^*(\mathbf{w}) = \mathrm{E}\left[\{Y^*_{n+1}(\mathbf{w}) - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}^2\right]
$$

$$
\begin{aligned}
&= \mathrm{E}\left(\left[\sum_{m=1}^{M} w_m\{Y^*_{(m),n+1} - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}\right]^2\right) \\
&= \mathrm{E}\left(\left[\sum_{m\notin\mathcal{D}} w_m\{Y^*_{(m),n+1} - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}\right]^2\right) \\
&= (1-\tau)^2\mathrm{E}\left(\left[\sum_{m\notin\mathcal{D}} (1-\tau)^{-1} w_m\{Y^*_{(m),n+1} - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}\right]^2\right) \\
&= (1-\tau)^2\mathrm{E}\left(\left[\sum_{m=1}^{M} \lambda_m\{Y^*_{(m),n+1} - \mathrm{E}(Y_{n+1}|\mathbf{X}_{n+1})\}\right]^2\right) \\
&\equiv (1-\tau)^2 R^*(\boldsymbol{\lambda}). \qquad \text{(C.14)}
\end{aligned}
$$

Note that the above result holds for $\widehat{\tau}$ and $\widehat{\boldsymbol{\lambda}}$ by replacing $\mathbf{w}$ with $\widehat{\mathbf{w}}$ in all equations. Therefore, combining (C.12) and (C.14), we have

$$CV^*_K(\widehat{\mathbf{w}})/n = (1-\widehat{\tau})^2 R^*(\widehat{\boldsymbol{\lambda}}) + O_p(n^{-1/2}M). \qquad \text{(C.15)}$$

Let $\widetilde{\mathbf{w}}$ be a weight vector with $\sum_{m\in\mathcal{D}} \widetilde{w}_m = 1$. Then, we have $R^*(\widetilde{\mathbf{w}}) = 0$ by (C.13). Hence, by (C.12), it follows that

$$CV^*_K(\widetilde{\mathbf{w}})/n = O_p(n^{-1/2}M). \qquad \text{(C.16)}$$

Next, by (C.15), (C.16), and the fact that $\widehat{\mathbf{w}}$ minimizes $CV^*_K(\mathbf{w})$, we have

$$(1-\widehat{\tau})^2 R^*(\widehat{\boldsymbol{\lambda}}) + O_p(n^{-1/2}M) \le CV^*_K(\widetilde{\mathbf{w}})/n = O_p(n^{-1/2}M). \qquad \text{(C.17)}$$

Hence, it follows that

$$(1-\widehat{\tau})^2 \inf_{\mathbf{w}\in\mathcal{W}_S} R^*(\mathbf{w}) + O_p(n^{-1/2}M) \le O_p(n^{-1/2}M). \qquad \text{(C.18)}$$

By (C.18) and Assumption 6, we obtain $\widehat{\tau} \to 1$ in probability. This completes the proof. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petroc and F. Csake (Eds.), *Second International Symposium on Information Theory*, pp. 267–281. Akademiai Kiado.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics 16*(1), 125–127.

Ando, T. and K.-C. Li (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association 109*(505), 254–265.

Ando, T. and K.-C. Li (2017). A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics 45*(6), 2654–2679.

Andrews, D. W. (1991). Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics 47*(2-3), 359–377.

Arlot, S. and M. Lerasle (2016). Choice of V for V-fold cross-validation in least-squares density estimation. *The Journal of Machine Learning Research 17*(1), 7256–7305.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*(2), 123–140.

Brownlee, J. (2018). *Better deep learning: Train faster, reduce overfitting, and make better predictions*. Machine Learning Mastery.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997). Model selection: An integral part of inference. *Biometrics 53*, 603–618.

Castillo, I., J. Schmidt-Hieber, and A. Van der Vaart (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics 43*(5), 1986–2018.

Charkhi, A., G. Claeskens, and B. E. Hansen (2016). Minimum mean squared error model averaging in likelihood models. *Statistica Sinica 26*(2), 809–840.

Cheng, X. and B. E. Hansen (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics 186*(2), 280–293.

Cheng, X., Z. Liao, and R. Shi (2019). On uniform asymptotic risk of averaging GMM estimators. *Quantitative Economics 10*(3), 931–979.

Claeskens, G. and N. L. Hjort (2008). *Model selection and model averaging*. Cambridge University Press.

Fang, F. and Y. Chen (2019). A new approach for credit scoring by directly maximizing the Kolmogorov–Smirnov statistic. *Computational Statistics & Data Analysis 133*, 180–194.

Feng, Y., Q. Liu, Q. Yao, and G. Zhao (2021). Model averaging for nonlinear regression models. *Journal of Business & Economic Statistics*, forthcoming.

Fernández, C., E. Ley, and M. F. Steel (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics 100*(2), 381–427.

Fernández-Villaverde, J. and J. F. Rubio-Ramírez (2004). Comparing dynamic equilibrium models to data: a Bayesian approach. *Journal of Econometrics 123*(1), 153–187.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*(1), 119–139.

Gao, Y., X. Zhang, S. Wang, T. T.-l. Chong, and G. Zou (2019). Frequentist model averaging for threshold models. *Annals of the Institute of Statistical Mathematics 71*(2), 275–306.

Gao, Y., X. Zhang, S. Wang, and G. Zou (2016). Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics 192*(1), 139–151.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association 70*(350), 320–328.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica 75*(4), 1175–1189.

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics 146*(2), 342–350.

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics 5*(3), 495–530.

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics 167*(1), 38–46.

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*(464), 879–899.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics 15*(3), 958–975.

Liao, J.-C. and W.-J. Tsay (2020). Optimal multi-step VAR forecasting averaging. *Econometric Theory 36*(6), 1099–1126.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics 186*(1), 142–159.

Liu, Q. and R. Okui (2013). Heteroscedasticity-robust C_p model averaging. *The Econometrics Journal 16*(3), 463–472.

Lu, X. and L. Su (2015). Jackknife model averaging for quantile regressions. *Journal of Econometrics 188*(1), 40–58.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics 15*(4), 661–675.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). Finite mixture models. *Annual Review of Statistics and Its Application 6*, 355–378.

Melnykov, V. and R. Maitra (2010). Finite mixture models and model-based clustering. *Statistics Surveys 4*, 80–116.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys 29*(1), 46–75.

Peng, J. and Y. Yang (2021). On improvability of model selection by model averaging. *Journal of Econometrics*, forthcoming.

Qiu, Y., T. Xie, J. Yu, and X. Zhang (2020). Mallows-type averaging machine learning techniques. Working Paper.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning 5*(2), 197–227.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association 88*(422), 486–494.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica 7*(2), 221–242.

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature 58*(3), 644–719.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological) 36*(2), 111–133.

Sun, Y., Y. Hong, T.-H. Lee, S. Wang, and X. Zhang (2021). Time-varying model averaging. *Journal of Econometrics 222*(2), 974–992.

Wan, A. T. K., X. Zhang, and G. Zou (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics 156*(2), 277–283.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks 5*(2), 241–259.

Yang, Y. (2000). Combining different procedures for adaptive regression. *Journal of Multivariate Analysis 74*(1), 135–161.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association 96*(454), 574–588.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika 92*(4), 937–950.

Yeh, I.-C. and C.-h. Lien (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications 36*(2), 2473–2480.

Yuan, Z. and Y. Yang (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association 100*(472), 1202–1214.

Zhang, C. and Y. Ma (2012). *Ensemble machine learning: Methods and applications*. Springer.

Zhang, X. (2010). *Model Averaging and Its Applications*. Ph. D. thesis, Academy of Mathematics and Systems Science, Chinese Academy of Sciences.

Zhang, X. (2015). Consistency of model averaging estimators. *Economics Letters 130*, 120–123.

Zhang, X. and H. Liang (2011). Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics 39*(1), 174–200.

Zhang, X., A. T. Wan, and G. Zou (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics 174*(2), 82–94.

Zhang, X., D. Yu, G. Zou, and H. Liang (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association 111*(516), 1775–1790.

Zhang, X., G. Zou, and H. Liang (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika 101*(1), 205–218.