Model Averaging Prediction by $K$-Fold Cross-Validation

Xinyu Zhang† and Chu-An Liu‡

March 10, 2021

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

---

*We thank the seminar participants of University of California, Riverside and conference participants of Macroeconometric Modelling Workshop for their discussions and suggestions. All errors remain the authors'.

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

‡Institute of Economics, Academia Sinica. 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 allow for the parameter uncertainty. It is well known that the conditional expectation of a dependent variable given covariates is the best predictor. We adopt the 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 the 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 could 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.

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 generalized linear mixed-effects model (Zhang et al., 2016), the high-dimensional generalized linear model (Ando and Li, 2017), the time-varying parameter regression model (Sun et al., 2020), and the vector autoregressive model (Liao and Tsay, 2020). 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.

In simulations, we examine the finite sample performance of the proposed model averaging prediction and compare it with other existing model selection and model averaging methods. Monte Carlo simulations show that the model averaging prediction by $K$-fold cross-validation generally produces lower empirical risk than alternative 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.
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 observations $\{(Y_i, X_i); i = 1, \ldots, n\}$, where $Y_i$ is a scalar dependent variable and $X_i$ is a vector of predictors. Let the likelihood function be

$$L(\theta) = \prod_{i=1}^{n} f(Y_i|X_i, \theta),$$

where $f$ is an unknown conditional probability density function and $\theta$ 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 $(Y_i, X_i)$ 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 $m$th candidate model uses the following quasi-likelihood function

$$\prod_{i=1}^{n} f_{(m)}(Y_i|X_i, \theta_{(m)}),$$

where the function $f_{(m)}$ is known, but it could be misspecified, and $\theta_{(m)}$ 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 $\theta$ could be zeros, we could also restrict some elements of $\theta$ 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 \( \hat{\theta}_{(m)} \) denote the estimator of \( \theta_{(m)} \) in the \( m \)th candidate model. Thus, the prediction of \( Y_{n+1} \) associated with the new observation \( X_{n+1} \) from this \( m \)th model is

\[
\hat{Y}_{(m),n+1} = E_{(m)}(Y_{n+1}|X_{n+1}, \hat{\theta}_{(m)}) = \int y f_{(m)}(y|X_{n+1}, \hat{\theta}_{(m)}) dy,
\]

where \( E_{(m)} \) is the expectation taken under the \( m \)th candidate model.

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

\[
\hat{Y}_{n+1}(w) = \sum_{m=1}^{M} w_m \hat{Y}_{(m),n+1}.
\]

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 \( E(e_i|X_i) = 0 \) and \( E(e_i^2|X_i) = \sigma^2 \). The data \( (Y_i, X_i) \) are i.i.d. and the unknown parameters are \( \theta = (\beta, \sigma^2) \). Suppose that we have a sequence of nested candidate models \( m = 1, \ldots, M \), where the \( m \)th candidate model uses the first \( m \) predictors in \( X_i \) and the standard Gaussian likelihood. Note that under normality, the maximum likelihood estimator of \( \theta \) is equivalent to the ordinary least squares estimator. Thus, the prediction of \( Y_{n+1} \) from the \( m \)th model is \( \hat{Y}_{(m),n+1} = X_{(m),n+1}^\top \hat{\beta}_{(m)} \), where \( \hat{\beta}_{(m)} = (X_{(m)}^\top X_{(m)})^{-1} X_{(m)}^\top Y \) and \( X_{(m)} \) includes the first \( m \) predictors in \( X_i \). The averaging prediction is \( \hat{Y}_{n+1}(w) = \sum_{m=1}^{M} w_m \hat{Y}_{(m),n+1} = \sum_{m=1}^{M} w_m X_{(m),n+1}^\top \hat{\beta}_{(m)} ; \) this is the least squares averaging estimator studied in Hansen (2007) and Hansen (2008).

**Example 2.** Suppose that we observe a binary dependent variable \( Y_i \in \{ 0,1 \} \) and a vector of predictors \( X_i \). Assume that \( Y_i \) is conditionally Bernoulli with \( Pr(Y_i = 1|X_i) = F(X_i^\top \beta) \), where \( F(\cdot) \) is an unknown cumulative distribution function. Our goal is to estimate \( Pr(Y_{n+1} = 1|X_{n+1}) \) 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 \( \Phi \) is a standard normal distribution function, while the second candidate model is a logit model setting \( F(u) = (1 + e^{-u})^{-1} \). Let \( \hat{\beta}_P \) and \( \hat{\beta}_L \) be the maximum likelihood estimator for the probit and logit model, respectively. The predictions of \( Y_{n+1} \) based on these two models are
\( \hat{Y}_{(1),n+1} = \Phi(\mathbf{X}_{n+1}^\top \hat{\beta}_p) \) and \( \hat{Y}_{(2),n+1} = (1 + e^{-\mathbf{X}_{n+1}^\top \hat{\beta}_L})^{-1} \). Instead of choosing between these two non-nested models, we could take the model misspecification into account and construct an averaging prediction as \( \hat{Y}_{n+1}(w) = w_1 \hat{Y}_{(1),n+1} + w_2 \hat{Y}_{(2),n+1} \).

### 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 \( k \)th group from the data set and use the remaining \( n - J \) observations to calculate the estimator \( \hat{\theta}^{[-k]}_m \) for each model. That is, \( \hat{\theta}^{[-k]}_m \) is the estimator of \( \theta_m \) in the \( m \)th model without using the observations from the \( k \)th group.

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

\[
\hat{Y}^{[-k]}_{(m),j} = \int y f_m(y|\mathbf{X}_{(k-1)\times J+j}, \hat{\theta}^{[-k]}_m) dy, \tag{3.1}
\]

for \( j = 1, \ldots, J \) and \( m = 1, \ldots, M \), where the subscript \( (k-1) \times J + j \) denotes the observations in the \( k \)th group.
**Step 3:** Compute the predictions for all observations for each model as follows

\[
\tilde{Y}_m = (\tilde{Y}_{1-1}^{(m)}, \ldots, \tilde{Y}_{J-1}^{(m)}, \ldots, \tilde{Y}_{J-K}^{(m)}, \ldots, \tilde{Y}_{J-K}^{(m)}, \ldots, \tilde{Y}_{J-K}^{(m)})^\top, \quad (3.2)
\]

and construct the \(K\)-fold cross-validation criterion

\[
CV_K(w) = \frac{1}{n} \|Y - \tilde{Y}(w)\|_2^2, \quad (3.3)
\]

where \(Y = (Y_1, \ldots, Y_n)^\top\) and \(\tilde{Y}(w) = \sum_{m=1}^M w_m \tilde{Y}_m = (\tilde{Y}_1^{[-1]}(w), \ldots, \tilde{Y}_J^{[-K]}(w))^\top\), where \(\tilde{Y}_j^{[-k]}(w) = \sum_{m=1}^M w_m \tilde{Y}_{j}^{[-k]}(w)\) 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

\[
\hat{w} = \arg\min_{w \in \mathcal{W}} CV_K(w), \quad (3.4)
\]

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

\[
\hat{Y}_{n+1}(\hat{w}) = \sum_{m=1}^M \hat{w}_m \hat{Y}_{(m),n+1}. \quad (3.5)
\]

Notice that the proposed \(K\)-fold cross-validation criterion is a quadratic function of the weight vector. Let \(\tilde{e} = (\tilde{e}_1, \ldots, \tilde{e}_M)\), where \(\tilde{e}_m = Y - \tilde{Y}_m\) is the \(K\)-fold cross-validation prediction error for the \(m\)th model. Then, the proposed criterion (3.3) can be written as a quadratic function of \(w\) as follows

\[
CV_K(w) = \frac{1}{n} w^\top \tilde{e}^\top \tilde{e} w. \quad (3.6)
\]

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, 10, and 20, 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 small, but the computational cost could be quite heavy when \(n\) and \(K\) are large. Our simulation and empirical results show that 5-fold, 10-fold, and 20-fold provide sufficiently accurate results with computational efficiency.
4 Theoretical properties

In this section, we present the asymptotic properties of the proposed averaging prediction. Define the risk function as

\[ R(w) \equiv \mathbb{E}\left[ \left\{ \hat{Y}_{n+1}(w) - \mathbb{E}(Y_{n+1}|X_{n+1}) \right\}^2 \right], \]

where \( \mathbb{E}(Y_{n+1}|X_{n+1}) = \int y f(y|X_{n+1}, \theta) dy \) is the true conditional expectation function. Ideally, one would aim to select the model weights \( w \) to minimize the risk function \( R(w) \) with respect to \( w \) in the set \( \mathcal{W} \). Unfortunately, this is infeasible because this minimization depends on the unknown conditional probability density function \( f \). Instead of minimizing \( R(w) \) 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.** There exists a limiting value \( \theta^*_m \) for \( \hat{\theta}_m \) such that \( \hat{\theta}_m - \theta^*_m = O_p(M^{1/2}n^{-1/2}) \) uniformly for \( m = 1, \ldots, M \).

Assumption 1 is a high-level condition, which ensures that the estimator of \( \theta_m \) in each candidate model has a limit \( \theta^*_m \). Here, \( \theta^*_m \) 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), and it holds under appropriate primitive assumptions; see conditions in Theorem 3.2 of White (1982). Note that Assumption 1 implies that the dimension of \( \theta_m \) is fixed. An extension to the case with a diverging number of parameters is left for future research.

We now introduce some notation associated with the limiting value \( \theta^*_m \). Let \( \theta^* = (\theta^*_1 \top, \ldots, \theta^*_M \top) \top \) be a vector that collects limiting values from all candidate models. Similarly, let \( \hat{\theta} = (\hat{\theta}_1 \top, \ldots, \hat{\theta}_M \top) \top \) and \( \theta^{[-k]} = (\hat{\theta}_1^{[-k]} \top, \ldots, \hat{\theta}_M^{[-k]} \top) \top \). Notice that \( \hat{\theta} \) and \( \theta^{[-k]} \) have the same limiting values \( \theta^* \) because \( n \) and \( n - n/K \) have the same order for any \( K \in \{2, \ldots, n\} \).

The prediction of \( Y_{n+1} \) calculated based on the limiting value \( \theta^*_m \) in the \( m \)th model is

\[ Y^*_m, n+1 = \int y f_{(m)}(y|X_{n+1}, \theta^*_m) dy, \quad (4.1) \]
and the averaging prediction calculated based on the limiting value $\theta^*_m$ is

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

Notice that Equations (4.1) and (4.2) correspond to Equations (2.3) and (2.4), except $\hat{\theta}^{(m)}$ is replaced by $\theta^*_m$. Similarly, we use $R^*(w)$ to denote the risk function calculated based on the limiting parameter value instead of the estimated parameter value, that is, $R^*(w) \equiv E[\{Y^*_{n+1}(w) - E(Y_{n+1} | X_{n+1})\}^2]$. Let $\xi_n \equiv \inf_{w \in \mathcal{W}} R^*(w)$ denote the minimum risk in the class of averaging estimators associated with the limiting value $\theta^*_m$.

**Assumption 2.** Let $\mathcal{O}(\theta^*, \varrho)$ denote a neighborhood of $\theta^*$ for some constant $\varrho$ such that $\|\theta^* - \theta\| \leq \varrho$ for any $\theta \in \mathcal{O}(\theta^*, \varrho)$. (i) $\tilde{Y}^*_m(n+1), \tilde{\theta}^*_m(n+1), Y_i$, and $E(Y_{n+1} | X_{n+1})$ are $O_p(1)$ uniformly for $m = 1, \ldots, M$ and $i = 1, \ldots, n$. (ii) $\tilde{Y}^*_m(n+1)$ and $\tilde{\theta}^*_{[-k]}$ are differentiable with respect to $\tilde{\theta}$ and $\tilde{\theta}^*_{[-k]}$, respectively. (iii) There exists a constant $\varrho$ such that

$$\sup_{\theta^* \in \mathcal{O}(\theta^*, \varrho)} \frac{\partial \tilde{Y}^*_m(n+1)}{\partial \theta} |_{\theta^*} = O_p(1) \quad \text{and} \quad \sup_{\theta^* \in \mathcal{O}(\theta^*, \varrho)} \frac{\partial \tilde{\theta}^*_{[-k]}}{\partial \theta} |_{\theta^*} = O_p(1),$$

uniformly for $m = 1, \ldots, M$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$.

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

$$\sup_{\theta^* \in \mathcal{O}(\theta^*, \varrho)} \frac{\partial \tilde{Y}^*_m(n+1)}{\partial \theta} |_{\theta^*} = (0^\top, \ldots, X_{(m),n+1}^\top, \ldots, 0^\top)^\top,$$

and thus Assumption 2 (iii) holds if $X_{(m),i} = O_p(1)$ holds uniformly for $i = 1, \ldots, n+1$ and $m = 1, \ldots, M$. Therefore, Assumption 2 (iii) is quite mild, since the covariate $X_j$ is $O_p(1)$ for any random variable with well-defined probability distribution.

In Example 2, we can show that

$$\sup_{\theta^* \in \mathcal{O}(\theta^*, \varrho)} \frac{\partial \tilde{Y}^*_1(n+1)}{\partial \theta} |_{\theta^*} = \phi(\mathbf{X}_{n+1}^\top \beta^*_P) \times (\mathbf{X}_{n+1}^\top, 0^\top)^\top,$$
Assumption 3. \( e(1 + X_{\infty}) \) is the limiting value of \( \hat{\phi} \), where \( E(\epsilon_{m+1}^{(j)}) \) is the expectation of the above equation is bounded.

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

Let \( \varepsilon_i = Y_i - E(Y_i|X_i) \) denote the error term in the prediction problem. Similarly, let \( \varepsilon_{(m),i} = Y_{(m),i}^* - E(Y_{(m),i}|X_i) \) denote the prediction error based on the limiting value \( \theta_{(m)}^* \) in the \( m \)th model. The following assumption imposes moment conditions of the error term and prediction error.

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

We now verify Assumption 4 in Example 1. Notice that \( Y_{(m),i}^* = X_{(m),i}^T \beta_{(m)}^* \), where \( \beta_{(m)}^* \) is the limiting value of \( \hat{\beta}_{(m)} \). Therefore, a simple sufficient condition for Assumption 4 is that \( E(X_{(m),i}^T \beta_{(m)}^*)^4 \), \( E(\sum_{j=1}^{\infty} \beta_j X_{ji})^4 \), and \( E\epsilon_i^4 \) are all bounded by a constant.
Assumption 5. $n^{-1/2}M^2\xi_n^{-1} = o(1)$.

Assumption 5 puts a bound on the number of models relative to the sample size, and it specifies that $M^2$ 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). This condition requires that all candidate models be misspecified. To better understand this condition, suppose that the $m^o$th model is correctly specified. Thus, we have $f(m^o)(\cdot) = f(\cdot)$ and $\theta^*(m^o) = \theta$, where $\theta$ is the true value defined in (2.1). Then it follows that

$$\xi_n = \inf_{w \in W} E \left[ \{Y_{n+1}(w) - E(Y_{n+1}|X_{n+1})\}^2 \right] \leq E \left[ \{Y_{m^o,n+1} - E(Y_{n+1}|X_{n+1})\}^2 \right] = 0,$$

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 that

$$\frac{R(\hat{w})}{\inf_{w \in W} R(w)} \to 1 \quad (4.3)$$

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 could 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 $\hat{\tau} = \sum_{m \in \mathcal{D}} \hat{w}_m$ be the sum of $K$-fold cross-validation weights given to the correctly specified models. We aim to show that $\hat{\tau} \to 1$ under some regularity conditions.

We first discuss the alternative condition for Assumption 5. Let $\mathcal{W}_S = \{w \in \mathcal{W} : \sum_{m \notin \mathcal{D}} w_m = 1\}$ 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.** $n^{-1/2} M^2 \left\{ \inf_{w \in \mathcal{W}_S} R^*(w) \right\}^{-1} = o(1)$.

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.

**Theorem 2.** Under Assumptions 1, 2, 4 and 6, if $\mathcal{D}$ is not empty, then we have that $\hat{\tau} \to 1$ 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 two simulation designs. The first design is the binary choice model, and we consider non-nested candidate models. The second 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 $\alpha$ 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 $E(Y_{n+1}|X_{n+1}) = Pr(Y_{n+1} = 1|X_{n+1}) = P_{n+1}$ and $\hat{Y}_{n+1} = \hat{E}(Y_{n+1}|X_{n+1}) = \hat{P}_{n+1}$. 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 5-fold cross-validation weights (labeled $K = 5$).

5. Model averaging estimator with 10-fold cross-validation weights (labeled $K = 10$).


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 estimator with $K$-fold cross-validation is described in Section 3.

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 $s$th replication. The empirical risk function is calculated as follows:

$$
\frac{1}{S} \sum_{s=1}^{S} \left\{ \hat{Y}_{n+1}^{(s)}(\hat{w}^{(s)}) - \mathbb{E}(Y_{n+1}^{(s)}|X_{n+1}^{(s)}) \right\}^2 = \frac{1}{S} \sum_{s=1}^{S} \left\{ \hat{P}_{n+1}^{(s)}(\hat{w}^{(s)}) - P_{n+1}^{(s)} \right\}^2,
$$

where $\hat{P}_{n+1}^{(s)}(\hat{w}^{(s)})$ is the prediction based on Probit, Logit, Equal, and $K$-fold cross-validation weights in the $s$th 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 $K = 5, 10, \text{ and } 20$. We then compute $\hat{P}_{n+1}^{(s)}(\hat{w}^{(s)})$ associated with the new observation $x_{n+1}$ for each method. The risk function is calculated by averaging across 2500 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 $n = 100, 200, 500, \text{ and } 1,000$ in four panels, and in each panel, the relative risk is displayed for $\alpha$ between 0 and 2. We first compare the finite sample performance between Probit and Logit. Probit has smaller relative risk than Logit for small $\alpha$, but larger relative risk than Logit for large $\alpha$. Recall that both models are misspecified, and neither Probit nor Logit uniformly dominates the other. For a fixed value of $\alpha$, 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 $K$-fold cross-validation weights. The relative risk of $K$-fold cross-validation is quite similar for $K = 5$, 10, and 20, and the $K$-fold cross-validation dominates the Equal estimator for most situations. When the sample size is small, the $K$-fold cross-validation performs better than Logit for small $\alpha$, but the relative risk of the $K$-fold cross-validation is slightly larger than that of Logit for large $\alpha$. When the sample size increases, the relative risk of Logit to $K$-fold cross-validation is getting larger for small $\alpha$, and the relative risk of the $K$-fold cross-validation is quite similar to that of Logit for large $\alpha$. 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 1,000, respectively. The model weight is calculated by averaging the $K$-fold cross-validation weights placed on the probit model across 2500 simulation replications. As we expected, the $K$-fold cross-validation weights are quite similar for $K = 5$, 10, and 20, which is consistent with the similar finite sample performance among these choices of $K$. 

Figure 1: Relative risk
Figure 2: Model weight placed on the probit model

\( K \) displayed in Figure 1. Notice that the \( K \)-fold cross-validation assigns more weights on Probit when \( \alpha \) is small, and put less weights on Probit when \( \alpha \) 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. An interesting observation is that the \( K \)-fold cross-validation weight on Probit is not monotonically decreasing in \( \alpha \).

### 5.2 Nonlinear regression model

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

\[
Y_i = \mu(X_i) + e_i = \exp \left( \sum_{j=1}^{p} \beta_j X_{ji} \right) + e_i, \quad \mathbb{E}(e_i|X_i) = 0,
\]

where \( X_{ji} \sim iid \) Uniform\((-1, 1)\). The error term is generated by \( e_i = \sigma_i \epsilon_i \), where \( \epsilon_i \) is generated from a log-normal distribution with mean zero and variance one. For the homoskedastic simulation, we set \( \sigma_i = 1 \), and for the heteroskedastic simulation, we set \( \sigma_i^2 = 0.5 + 0.5\sigma_{\mu i}^2 \). 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 $m$th model uses the first $m$ predictors in $X_i$. That is, the $m$th 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:

Case 1: $p = 10$, $M = 8$, $\beta = (1^{\delta}, 2^{\delta}, \ldots, p^{\delta})'$,

Case 2: $p = 10$, $M = 10$, $\beta = (1^{\delta}, 2^{\delta}, \ldots, p^{\delta})'$,

Case 3: $p = 10$, $M = 10$, $\beta = (1^{\delta}, 2^{\delta}, \ldots, 6^{\delta}, 0, 0, 0, 0)'$.

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 the correctly specified model. 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).
7. Model averaging estimator with 5-fold cross-validation weights (labeled $K = 5$).
8. Model averaging estimator with 10-fold cross-validation weights (labeled $K = 10$).

We briefly discuss each estimator. The AIC criterion for the $m$th model is $AIC_m = n \log(\hat{\sigma}_m^2) + 2p_m$, where $\hat{\sigma}_m^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_{mi}^2$. Here, $\hat{e}_{mi}$ 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 \( m \)th model is \( \text{BIC}_m = n \log(\hat{\sigma}_m^2) + \log(n)p_m \). For AIC and BIC, we select the model with the smallest \( \text{AIC}_m \) and \( \text{AIC}_m \), 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 \( \hat{\omega}_m = \exp(-\frac{1}{2}\text{AIC}_m) / \sum_{j=1}^{M} \exp(-\frac{1}{2}\text{AIC}_j) \). The SBIC estimator is a simplified form of Bayesian model averaging with diffuse priors, and the SBIC weight is \( \hat{\omega}_m = \exp(-\frac{1}{2}\text{BIC}_m) / \sum_{j=1}^{M} \exp(-\frac{1}{2}\text{BIC}_j) \). 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 \( C(w) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\mu}_i(w))^2 + 2\hat{\sigma}^2 \sum_{m=1}^{M} w_m \pi_m \), where \( \hat{\mu}_i(w) \) is the model averaging estimator of \( \mu(X_i) \) and \( \pi_m \) is a bias-adjusting term based on the first derivatives of \( \mu(X_i) \) and the Hessian matrix. The Equal estimator assigns \( \frac{1}{M} \) weight to each candidate model, and the \( K \)-fold cross-validation weights are described in Section 3.

We evaluate the finite sample behavior of each method based on the following empirical risk function \( \frac{1}{S} \sum_{s=1}^{S} \left( \hat{Y}_{n+1}^{(s)}(\hat{w}^{(s)}) - E(Y_{n+1}^{(s)}|X_{n+1}^{(s)}) \right)^2 \), where \( E(Y_{n+1}^{(s)}|X_{n+1}^{(s)}) = \mu(X_{n+1}^{(s)}) \equiv \mu_{n+1}^{(s)} \) and \( \hat{Y}_{n+1}^{(s)}(\hat{w}^{(s)}) = \hat{\mu}_{n+1}^{(s)}(\hat{w}^{(s)}) \) is the prediction based on each method in the \( s \)th replication. Like the first simulation design, the risk function is calculated by averaging across 2500 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, AIC performs better than BIC for \( n = 100 \), but BIC achieves lower relative risk for larger sample sizes in Case 3. For SAIC and SBIC, the pattern of relative performance between SAIC and SBIC is quite similar to that of AIC and BIC. One interesting observation is that SAIC and SBIC have better performance than AIC and BIC, 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-
Table 1: Relative risk in the homoskedastic setup

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>AIC</th>
<th>BIC</th>
<th>SAIC</th>
<th>SBIC</th>
<th>NMA</th>
<th>Equal</th>
<th>K=5</th>
<th>K=10</th>
<th>K=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>100</td>
<td>2.658</td>
<td>2.465</td>
<td>2.432</td>
<td>2.244</td>
<td>1.730</td>
<td>1.141</td>
<td>0.984</td>
<td>1.000</td>
<td>0.985</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>2.167</td>
<td>2.155</td>
<td>2.102</td>
<td>2.054</td>
<td>1.398</td>
<td>0.978</td>
<td>0.966</td>
<td>1.000</td>
<td>1.013</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>1.432</td>
<td>1.448</td>
<td>1.432</td>
<td>1.435</td>
<td>1.068</td>
<td>1.253</td>
<td>0.982</td>
<td>1.000</td>
<td>1.001</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>1.399</td>
<td>1.403</td>
<td>1.244</td>
<td>1.280</td>
<td>1.072</td>
<td>1.332</td>
<td>1.017</td>
<td>1.000</td>
<td>1.005</td>
</tr>
<tr>
<td>Case 2</td>
<td>100</td>
<td>3.230</td>
<td>2.999</td>
<td>2.917</td>
<td>2.479</td>
<td>2.030</td>
<td>0.878</td>
<td>1.000</td>
<td>1.000</td>
<td>0.986</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>2.206</td>
<td>2.474</td>
<td>2.111</td>
<td>2.249</td>
<td>1.370</td>
<td>0.938</td>
<td>0.970</td>
<td>1.000</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>1.682</td>
<td>1.787</td>
<td>1.676</td>
<td>1.759</td>
<td>1.096</td>
<td>1.699</td>
<td>1.027</td>
<td>1.000</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>1.848</td>
<td>1.861</td>
<td>1.837</td>
<td>1.831</td>
<td>1.110</td>
<td>1.948</td>
<td>0.995</td>
<td>1.000</td>
<td>1.003</td>
</tr>
<tr>
<td>Case 3</td>
<td>100</td>
<td>3.505</td>
<td>3.836</td>
<td>3.599</td>
<td>2.937</td>
<td>2.911</td>
<td>1.258</td>
<td>1.021</td>
<td>1.000</td>
<td>1.004</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>5.380</td>
<td>4.701</td>
<td>4.914</td>
<td>4.477</td>
<td>3.318</td>
<td>1.187</td>
<td>0.963</td>
<td>1.000</td>
<td>0.944</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>2.142</td>
<td>1.857</td>
<td>2.021</td>
<td>1.823</td>
<td>1.511</td>
<td>1.039</td>
<td>1.004</td>
<td>1.000</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>1.424</td>
<td>1.275</td>
<td>1.176</td>
<td>1.144</td>
<td>1.076</td>
<td>1.169</td>
<td>1.002</td>
<td>1.000</td>
<td>0.992</td>
</tr>
</tbody>
</table>

Table 2: Relative risk in the heteroskedastic setup

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>AIC</th>
<th>BIC</th>
<th>SAIC</th>
<th>SBIC</th>
<th>NMA</th>
<th>Equal</th>
<th>K=5</th>
<th>K=10</th>
<th>K=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>100</td>
<td>3.948</td>
<td>4.192</td>
<td>3.775</td>
<td>3.637</td>
<td>2.421</td>
<td>1.150</td>
<td>1.002</td>
<td>1.000</td>
<td>0.986</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>1.417</td>
<td>1.650</td>
<td>1.442</td>
<td>1.582</td>
<td>1.068</td>
<td>0.998</td>
<td>0.990</td>
<td>1.000</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>1.292</td>
<td>1.241</td>
<td>1.285</td>
<td>1.252</td>
<td>1.027</td>
<td>1.548</td>
<td>1.009</td>
<td>1.000</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>1.132</td>
<td>1.127</td>
<td>1.132</td>
<td>1.128</td>
<td>0.998</td>
<td>1.676</td>
<td>0.989</td>
<td>1.000</td>
<td>1.002</td>
</tr>
<tr>
<td>Case 2</td>
<td>100</td>
<td>2.549</td>
<td>2.353</td>
<td>2.303</td>
<td>2.087</td>
<td>1.601</td>
<td>0.995</td>
<td>1.102</td>
<td>1.000</td>
<td>1.035</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>2.064</td>
<td>2.342</td>
<td>1.977</td>
<td>2.140</td>
<td>1.321</td>
<td>1.348</td>
<td>1.011</td>
<td>1.000</td>
<td>0.964</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>1.747</td>
<td>1.808</td>
<td>1.750</td>
<td>1.825</td>
<td>1.156</td>
<td>2.006</td>
<td>1.006</td>
<td>1.000</td>
<td>0.983</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>1.759</td>
<td>1.750</td>
<td>1.846</td>
<td>1.918</td>
<td>1.094</td>
<td>3.089</td>
<td>1.024</td>
<td>1.000</td>
<td>0.990</td>
</tr>
<tr>
<td>Case 3</td>
<td>100</td>
<td>3.479</td>
<td>2.800</td>
<td>3.096</td>
<td>2.284</td>
<td>2.391</td>
<td>0.899</td>
<td>0.887</td>
<td>1.000</td>
<td>1.069</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>5.282</td>
<td>4.210</td>
<td>4.947</td>
<td>4.090</td>
<td>3.324</td>
<td>1.089</td>
<td>1.037</td>
<td>1.000</td>
<td>1.054</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>1.839</td>
<td>1.753</td>
<td>1.770</td>
<td>1.704</td>
<td>1.331</td>
<td>1.110</td>
<td>1.037</td>
<td>1.000</td>
<td>1.006</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>1.753</td>
<td>1.548</td>
<td>1.713</td>
<td>1.546</td>
<td>1.168</td>
<td>1.048</td>
<td>0.975</td>
<td>1.000</td>
<td>0.987</td>
</tr>
</tbody>
</table>
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. Equal has better finite sample performance than NMA for \( n = 100 \) and 200 in Cases 1-3, but NMA has smaller relative risk than Equal for \( n = 800 \) in Cases 1-3. The proposed averaging prediction using \( K \)-fold cross-validation weights performs quite well and achieves lower relative risk than other methods in most situations. Similar to the finding in the first simulation design, the relative risk of 5-fold, 10-fold, and 20-fold cross-validation is quite similar.

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 still achieves lower relative risk than other methods in most situations.

Figure 3 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 2500 simulation replications. That is, we calculate the sum of weights by $\frac{1}{2500} \sum_{s=1}^{2500} \hat{\tau}_s$, where $\hat{\tau}_s$ is the sum of $K$-fold cross-validation weights given to the correctly specified model in the $s$th simulation replication. As shown in Figure 3, the sum of model weights is monotonically increasing and generally converges to one as the sample size increases, which is consistent with Theorem 2. Similar to results in Figure 2, the sums of $K$-fold cross-validation weights are quite similar for $K = 5, 10,$ and $20$.

Figure 4 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. It is clear to see that $K$-fold cross-validation assigns more weights to the just-fitted model than the over-fitted models. The individual weights on the 6th to 10th models all increase with the sample size, but the weight on the just-fitted model increases much
faster than those on the over-fitted models as the sample size increases.

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 30,000 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 of 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), amount of bill statement in NT dollars in the past six months (from April to September 2005), and amount of previous payment 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 $m$th 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. The sample sizes $n_1$ and $n_2$ vary between 100, 200, 500, and 1000. 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 second 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

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

where $\sigma^2$ is estimated by the sample analogue $\hat{\sigma}^2$ and $\hat{Y}_{n_1+j}(\hat{\mathbf{w}})$ is the probability prediction based on each method. As pointed out in Hansen (2008), the error variance $\sigma^2$ is the common leading term of the MSPE across all candidate models, and the scaling $\frac{1}{\hat{\sigma}^2}$ is used to ensure that results are scale-free. Note that we use the same estimator $\hat{\sigma}^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{e}_{Mi}^2$ for all methods, and $\hat{e}_{Mi}$ 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 and median of the MSPEs. For easy comparison, we divide the mean and median 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.

Tables 3 and 4 present the relative mean and median of the MSPEs in Case (i). It is clear that $K$-fold cross-validation achieves a lower mean and median of the MSPEs than other model selection and model averaging methods in all scenarios, and the prediction performance of $K$-fold cross-validation is quite similar for $K = 5$, 10, and 20. For a fixed value of the training sample size $n_1$, AIC, BIC, SAIC, SBIC, NMA, and Equal have similar prediction performance across different evaluation sample sizes $n_2$. As the training sample
Table 3: Relative mean of the MSPEs in Case (i) of the empirical example

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$n_2$</th>
<th>AIC</th>
<th>BIC</th>
<th>SAIC</th>
<th>SBIC</th>
<th>NMA</th>
<th>Equal</th>
<th>K=5</th>
<th>K=10</th>
<th>K=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>100</td>
<td>1.280</td>
<td>1.118</td>
<td>1.206</td>
<td>1.081</td>
<td>1.481</td>
<td>1.025</td>
<td>0.998</td>
<td>1.000</td>
<td>1.001</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>1.251</td>
<td>1.102</td>
<td>1.185</td>
<td>1.068</td>
<td>1.464</td>
<td>1.027</td>
<td>1.001</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>1.286</td>
<td>1.106</td>
<td>1.216</td>
<td>1.069</td>
<td>1.495</td>
<td>1.029</td>
<td>1.002</td>
<td>1.000</td>
<td>0.999</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>1.255</td>
<td>1.119</td>
<td>1.182</td>
<td>1.080</td>
<td>1.470</td>
<td>1.024</td>
<td>1.002</td>
<td>1.000</td>
<td>0.999</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>1.145</td>
<td>1.022</td>
<td>1.087</td>
<td>1.009</td>
<td>1.256</td>
<td>1.061</td>
<td>1.001</td>
<td>1.000</td>
<td>0.997</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>1.166</td>
<td>1.029</td>
<td>1.095</td>
<td>1.008</td>
<td>1.303</td>
<td>1.072</td>
<td>1.001</td>
<td>1.000</td>
<td>0.999</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>1.155</td>
<td>1.034</td>
<td>1.095</td>
<td>1.015</td>
<td>1.284</td>
<td>1.073</td>
<td>1.005</td>
<td>1.000</td>
<td>0.999</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>1.154</td>
<td>1.026</td>
<td>1.099</td>
<td>1.007</td>
<td>1.304</td>
<td>1.077</td>
<td>1.002</td>
<td>1.000</td>
<td>0.998</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>1.171</td>
<td>1.091</td>
<td>1.179</td>
<td>1.179</td>
<td>1.192</td>
<td>1.179</td>
<td>1.008</td>
<td>1.000</td>
<td>1.003</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>1.162</td>
<td>1.088</td>
<td>1.190</td>
<td>1.190</td>
<td>1.202</td>
<td>1.190</td>
<td>1.004</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>1.153</td>
<td>1.075</td>
<td>1.163</td>
<td>1.163</td>
<td>1.189</td>
<td>1.163</td>
<td>1.003</td>
<td>1.000</td>
<td>1.002</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>1.166</td>
<td>1.079</td>
<td>1.190</td>
<td>1.190</td>
<td>1.209</td>
<td>1.190</td>
<td>1.000</td>
<td>1.000</td>
<td>1.001</td>
</tr>
</tbody>
</table>

Table 4: Relative median of the MSPEs in Case (i) of the empirical example

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$n_2$</th>
<th>AIC</th>
<th>BIC</th>
<th>SAIC</th>
<th>SBIC</th>
<th>NMA</th>
<th>Equal</th>
<th>K=5</th>
<th>K=10</th>
<th>K=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>100</td>
<td>1.284</td>
<td>1.140</td>
<td>1.163</td>
<td>1.107</td>
<td>1.511</td>
<td>1.012</td>
<td>0.997</td>
<td>1.000</td>
<td>1.002</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>1.247</td>
<td>1.176</td>
<td>1.163</td>
<td>1.128</td>
<td>1.565</td>
<td>1.056</td>
<td>1.008</td>
<td>1.000</td>
<td>1.005</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>1.241</td>
<td>1.129</td>
<td>1.171</td>
<td>1.093</td>
<td>1.515</td>
<td>1.045</td>
<td>0.999</td>
<td>1.000</td>
<td>0.999</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>1.194</td>
<td>1.123</td>
<td>1.108</td>
<td>1.078</td>
<td>1.480</td>
<td>1.028</td>
<td>1.010</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>1.189</td>
<td>1.052</td>
<td>1.086</td>
<td>1.019</td>
<td>1.344</td>
<td>1.106</td>
<td>1.020</td>
<td>1.000</td>
<td>1.007</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>1.204</td>
<td>1.042</td>
<td>1.106</td>
<td>1.030</td>
<td>1.381</td>
<td>1.126</td>
<td>1.016</td>
<td>1.000</td>
<td>1.022</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>1.135</td>
<td>1.016</td>
<td>1.075</td>
<td>1.003</td>
<td>1.268</td>
<td>1.055</td>
<td>0.991</td>
<td>1.000</td>
<td>0.988</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>1.144</td>
<td>1.034</td>
<td>1.092</td>
<td>0.992</td>
<td>1.325</td>
<td>1.084</td>
<td>1.006</td>
<td>1.000</td>
<td>0.999</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>1.233</td>
<td>1.095</td>
<td>1.175</td>
<td>1.175</td>
<td>1.222</td>
<td>1.175</td>
<td>1.059</td>
<td>1.000</td>
<td>1.022</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>1.185</td>
<td>1.100</td>
<td>1.271</td>
<td>1.271</td>
<td>1.287</td>
<td>1.271</td>
<td>0.967</td>
<td>1.000</td>
<td>0.988</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>1.190</td>
<td>1.087</td>
<td>1.227</td>
<td>1.227</td>
<td>1.216</td>
<td>1.227</td>
<td>1.026</td>
<td>1.000</td>
<td>1.001</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>1.209</td>
<td>1.084</td>
<td>1.256</td>
<td>1.256</td>
<td>1.286</td>
<td>1.256</td>
<td>0.998</td>
<td>1.000</td>
<td>0.997</td>
</tr>
</tbody>
</table>
size increases, the mean of the MSPEs of these estimators are not getting close to one, which shows that the relative prediction performance of these estimators does not improve with sample size.

Tables 5 and 6 present the relative mean and median of the MSPEs in Case (ii). Similar to the results in Tables 3 and 4, the mean and median of the MSPEs of $K$-fold cross-validation are smaller than those of other model selection and model averaging methods in most scenarios. Furthermore, the prediction performance of 5-fold, 10-fold, and 20-fold cross-validation is quite similar for $n_1 = 200, 500$ and 1000. Comparing the results between Case (i) and Case (ii), we find that the relative mean and median of the MSPEs of AIC, BIC, and NMA in Case (ii) are slightly larger than those of AIC, BIC, and NMA in Case (i).

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 generally achieves lower empirical risk than other existing methods, and the finite sample performance of the proposed method is not sensitive to the choice of $K$. However, we do not provide any guideline for the number of groups in $K$-fold cross-validation, and it would be greatly desirable to study a data-driven approach for the choice of $K$. Another possible extension would be to extend the proposed method to the model setting with a diverging number of parameters.
Table 5: Relative mean of the MSPEs in Case (ii) of the empirical example

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$n_2$</th>
<th>AIC</th>
<th>BIC</th>
<th>SAIC</th>
<th>SBIC</th>
<th>NMA</th>
<th>Equal</th>
<th>K=5</th>
<th>K=10</th>
<th>K=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>100</td>
<td>1.313</td>
<td>1.064</td>
<td>1.211</td>
<td>1.012</td>
<td>1.465</td>
<td>0.992</td>
<td>0.997</td>
<td>1.000</td>
<td>1.001</td>
</tr>
<tr>
<td>200</td>
<td>1.330</td>
<td>1.069</td>
<td>1.219</td>
<td>1.013</td>
<td>1.525</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>1.307</td>
<td>1.058</td>
<td>1.203</td>
<td>1.007</td>
<td>1.495</td>
<td>0.999</td>
<td>1.003</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>1.317</td>
<td>1.064</td>
<td>1.217</td>
<td>1.009</td>
<td>1.488</td>
<td>0.994</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>1.202</td>
<td>1.059</td>
<td>1.112</td>
<td>1.017</td>
<td>1.339</td>
<td>1.016</td>
<td>0.997</td>
<td>1.000</td>
<td>1.003</td>
</tr>
<tr>
<td>200</td>
<td>1.218</td>
<td>1.066</td>
<td>1.120</td>
<td>1.025</td>
<td>1.363</td>
<td>1.008</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>1.203</td>
<td>1.071</td>
<td>1.114</td>
<td>1.021</td>
<td>1.343</td>
<td>1.009</td>
<td>0.999</td>
<td>1.000</td>
<td>1.002</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>1.213</td>
<td>1.072</td>
<td>1.122</td>
<td>1.026</td>
<td>1.350</td>
<td>1.007</td>
<td>1.003</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>1.194</td>
<td>1.176</td>
<td>1.067</td>
<td>1.067</td>
<td>1.279</td>
<td>1.067</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>200</td>
<td>1.181</td>
<td>1.152</td>
<td>1.050</td>
<td>1.050</td>
<td>1.240</td>
<td>1.050</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>1.189</td>
<td>1.207</td>
<td>1.060</td>
<td>1.060</td>
<td>1.273</td>
<td>1.060</td>
<td>1.005</td>
<td>1.000</td>
<td>1.002</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>1.175</td>
<td>1.195</td>
<td>1.054</td>
<td>1.054</td>
<td>1.249</td>
<td>1.054</td>
<td>1.001</td>
<td>1.000</td>
<td>1.001</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Relative median of the MSPEs in Case (ii) of the empirical example

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$n_2$</th>
<th>AIC</th>
<th>BIC</th>
<th>SAIC</th>
<th>SBIC</th>
<th>NMA</th>
<th>Equal</th>
<th>K=5</th>
<th>K=10</th>
<th>K=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>100</td>
<td>1.335</td>
<td>1.075</td>
<td>1.213</td>
<td>1.013</td>
<td>1.553</td>
<td>1.024</td>
<td>1.026</td>
<td>1.000</td>
<td>1.012</td>
</tr>
<tr>
<td>200</td>
<td>1.278</td>
<td>1.064</td>
<td>1.139</td>
<td>1.026</td>
<td>1.552</td>
<td>0.987</td>
<td>1.001</td>
<td>1.000</td>
<td>0.996</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>1.253</td>
<td>1.050</td>
<td>1.147</td>
<td>0.996</td>
<td>1.525</td>
<td>0.994</td>
<td>1.002</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>1.299</td>
<td>1.079</td>
<td>1.166</td>
<td>1.000</td>
<td>1.524</td>
<td>1.004</td>
<td>0.995</td>
<td>1.000</td>
<td>0.994</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>1.223</td>
<td>1.004</td>
<td>1.124</td>
<td>0.996</td>
<td>1.303</td>
<td>0.950</td>
<td>0.995</td>
<td>1.000</td>
<td>1.006</td>
</tr>
<tr>
<td>200</td>
<td>1.218</td>
<td>1.057</td>
<td>1.095</td>
<td>1.018</td>
<td>1.377</td>
<td>1.026</td>
<td>0.977</td>
<td>1.000</td>
<td>0.956</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>1.207</td>
<td>1.102</td>
<td>1.129</td>
<td>1.026</td>
<td>1.356</td>
<td>1.004</td>
<td>1.012</td>
<td>1.000</td>
<td>1.012</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>1.220</td>
<td>1.086</td>
<td>1.118</td>
<td>1.033</td>
<td>1.377</td>
<td>1.031</td>
<td>0.992</td>
<td>1.000</td>
<td>0.995</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>1.359</td>
<td>1.246</td>
<td>1.201</td>
<td>1.201</td>
<td>1.569</td>
<td>1.201</td>
<td>1.020</td>
<td>1.000</td>
<td>0.961</td>
</tr>
<tr>
<td>200</td>
<td>1.291</td>
<td>1.188</td>
<td>1.122</td>
<td>1.122</td>
<td>1.428</td>
<td>1.122</td>
<td>0.963</td>
<td>1.000</td>
<td>1.032</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>1.224</td>
<td>1.175</td>
<td>1.129</td>
<td>1.129</td>
<td>1.326</td>
<td>1.129</td>
<td>0.995</td>
<td>1.000</td>
<td>0.979</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>1.183</td>
<td>1.190</td>
<td>1.071</td>
<td>1.071</td>
<td>1.278</td>
<td>1.071</td>
<td>1.007</td>
<td>1.000</td>
<td>0.995</td>
<td></td>
</tr>
</tbody>
</table>
Appendix

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

\[ \tilde{w} = \arg\min_{w \in \mathcal{W}} \{ R(w) + a_n(w) + b_n \} , \]

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

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

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

A.1 Proof of Theorem 1

Let \( CV^*_K(w) = CV_K(w) - \{ Y - E(Y|X) \} \top \{ Y + E(Y|X) \} \), where the second term is unrelated to \( w \). Therefore,

\[ \tilde{w} = \arg\min_{w \in \mathcal{W}} CV_K(w) = \arg\min_{w \in \mathcal{W}} CV^*_K(w). \]

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

\[ \sup_{w \in \mathcal{W}} |R(w) - R^*(w)|/R^*(w) = o(1) \]  \hspace{1cm} (A.1)

and

\[ \sup_{w \in \mathcal{W}} |CV^*_K(w)/n - R^*(w)|/R^*(w) = o_p(1). \]  \hspace{1cm} (A.2)

We first consider (A.1). Observe that

\[ \xi_n^{-1} \sup_{w \in \mathcal{W}} \left| \left( \hat{Y}_{n+1}(w) - E(Y_{n+1}|X_{n+1}) \right)^2 - \left( Y^*_{n+1}(w) - E(Y_{n+1}|X_{n+1}) \right)^2 \right| \]

\[ = \xi_n^{-1} \sup_{w \in \mathcal{W}} \left| \left( \hat{\theta} - \theta^* \right) \top \left( \frac{\partial \hat{Y}_{n+1}(w) - E(Y_{n+1}|X_{n+1})}{\partial \theta} \right) \right|_{\theta = \bar{\theta}(w)} \]

\[ = \xi_n^{-1} O_p(n^{-1/2}M^{3/2}) \]
where the first equality uses Assumption 2, the second equality uses Assumptions 1 and 2, and the third equality uses Assumption 5. Hence, we have

\[
\frac{\sup_{w \in W} |R(w) - R^*(w)|}{R^*(w)} \leq \xi_n^{-1} \sup_{w \in W} |R(w) - R^*(w)|
\]

\[
= \xi_n^{-1} \sup_{w \in W} \left| \mathbb{E}\left[\left\{\tilde{Y}_{n+1}(w) - \mathbb{E}(Y_{n+1}|X_{n+1})\right\}^2 - \left\{Y^*_n(w) - \mathbb{E}(Y_{n+1}|X_{n+1})\right\}^2\right]\right| 
\]

\[
\leq \mathbb{E}\left(\xi_n^{-1} \sup_{w \in W} \left| \mathbb{E}\left[\left\{\tilde{Y}_{n+1}(w) - \mathbb{E}(Y_{n+1}|X_{n+1})\right\}^2 - \left\{Y^*_n(w) - \mathbb{E}(Y_{n+1}|X_{n+1})\right\}^2\right]\right| \right)
\]

\[
= o(1),
\]

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

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

\[
|CV^*_K(w)/n - R^*(w)|
\]

\[
= \left| \left\{\|\tilde{Y}(w) - Y\|^2 - \{Y - \mathbb{E}(Y|X)\}^T \{Y + \mathbb{E}(Y|X)\}\right\} /n - \mathbb{E}\left[\left\{Y^*_n(w) - \mathbb{E}(Y_{n+1}|X_{n+1})\right\}^2\right]\right| 
\]

\[
\leq \left[ \left\|Y^*_m(w) - \mathbb{E}(Y|X)\right\|^2/n - \mathbb{E}\left[\left\{Y^*_n(w) - \mathbb{E}(Y_{n+1}|X_{n+1})\right\}^2\right]\right] + \left| \left\|\tilde{Y}(w) - Y\|^2 - \|Y^*_m(w) - Y\|^2\right\| /n 
\]

\[
\leq \left\|Y^*_m(w) - \mathbb{E}(Y|X)\right\|^2/n - \mathbb{E}\left[\left\{Y^*_n(w) - \mathbb{E}(Y_{n+1}|X_{n+1})\right\}^2\right] + 2 \left\|Y^*_m(w)\right\|^T \{Y - \mathbb{E}(Y|X)\} /n + \left| \left\|\tilde{Y}(w) - Y\|^2 - \|Y^*_m(w) - Y\|^2\right\| /n. \right.
\]

(A.5)

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

\[
\frac{\sup_{w \in W} \left| \tilde{Y}(w) - Y\right|^2}{\left\|Y^*_m(w) - Y\right\|^2}
\]

\[
= \frac{\sup_{w \in W} \left| \sum_{k=1}^J \sum_{j=1}^J \left[ \tilde{Y}_j^{-[k]}(w) - Y_{(k-1) \times J+j}\right]^2 - \left\{Y^*_{(k-1) \times J+j}(w) - Y_{(k-1) \times J+j}\right\}^2\right|}{\left\|Y^*_m(w) - Y\right\|^2}
\]

\[
= O_p(n^{1/2}M^{3/2}). \right.
\]

(A.6)
It is seen that for any $\delta > 0$,
\[
\begin{align*}
\Pr \left\{ \xi_n^{-1} \sup_{w \in \mathcal{W}} \left\| Y^*(w) - E(Y|X) \right\|^2 / n - E \left\{ \{Y^*_{n+1}(w) - E(Y_{n+1}|X_{n+1})\}^2 \right\} > \delta \right\} \\
= \Pr \left\{ \xi_n^{-1} \sup_{w \in \mathcal{W}} \sum_{m=1}^{M} \sum_{m'=1}^{M} w_m w_{m'} \left\{ \frac{1}{n} \sum_{i=1}^{n} \{Y^*_{(m),i} - E(Y_i|X_i)\} \{Y^*_{(m'),i} - E(Y_i|X_i)\} \right. \\
- \left. E \left\{ \{Y^*_{(m),i} - E(Y_i|X_i)\} \{Y^*_{(m'),i} - E(Y_i|X_i)\} \right\} > \delta \right\} \\
\leq \sum_{m=1}^{M} \sum_{m'=1}^{M} \Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \{Y^*_{(m),i} - E(Y_i|X_i)\} \{Y^*_{(m'),i} - E(Y_i|X_i)\} \right| > \xi_n \delta \right\} \\
\leq \xi_n^{-1} \delta^{-1} n^{-1/2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \text{var} \left\{ \{Y^*_{(m),i} - E(Y_i|X_i)\} \{Y^*_{(m'),i} - E(Y_i|X_i)\} \right\},
\end{align*}
\] (A.7)

where the first inequality uses Boole’s inequality and the second inequality uses Chebyshev’s Inequality. Similar to (A.7), it is seen that
\[
\begin{align*}
\Pr \left\{ \xi_n^{-1} \sup_{w \in \mathcal{W}} \left| \frac{1}{n} \sum_{i=1}^{n} [Y^*_{i}(w) Y_i - E(Y_i|X_i)] \right| > \delta \right\} \\
= \Pr \left\{ \sup_{w \in \mathcal{W}} \sum_{m=1}^{M} w_m \left[ \frac{1}{n} \sum_{i=1}^{n} Y^*_{(m),i} Y_i - E(Y_i|X_i) \right] > \xi_n \delta \right\} \\
\leq \sum_{m=1}^{M} \Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} Y^*_{(m),i} Y_i - E(Y_i|X_i) \right| > \xi_n \delta \right\} \\
\leq \xi_n^{-1} \delta^{-1} n^{-1/2} \sum_{m=1}^{M} \text{var} \left\{ Y^*_{(m),i} Y_i - Y^*_{(m),i} E(Y_i|X_i) \right\}.
\end{align*}
\] (A.8)

By (A.5)-(A.8) and Assumptions 4 and 5, we obtain (A.2). This completes the proof. \qed

### A.2 Proof of Theorem 2

Following a similar argument in (A.7), we have
\[
\begin{align*}
\sup_{w \in \mathcal{W}} \left\| Y^*(w) - E(Y|X) \right\|^2 / n - E \left\{ \{Y^*_{n+1}(w) - E(Y_{n+1}|X_{n+1})\}^2 \right\} \\
= \sup_{w \in \mathcal{W}} \sum_{m=1}^{M} \sum_{m'=1}^{M} w_m w_{m'} \left[ \frac{1}{n} \sum_{i=1}^{n} \{Y^*_{(m),i} - E(Y_i|X_i)\} \{Y^*_{(m'),i} - E(Y_i|X_i)\} \right. \\
- \left. E \left\{ \{Y^*_{(m),i} - E(Y_i|X_i)\} \{Y^*_{(m'),i} - E(Y_i|X_i)\} \right\} \right] \\
\leq \sum_{m=1}^{M} \sum_{m'=1}^{M} \left| \frac{1}{n} \sum_{i=1}^{n} \{Y^*_{(m),i} - E(Y_i|X_i)\} \{Y^*_{(m'),i} - E(Y_i|X_i)\} \right|
\end{align*}
\]
\[-E \left[ \{ Y_{(m),i}^* - E(Y_i|X_i) \} \{ Y_{(m'),i}^* - E(Y_i|X_i) \} \right] \]
\[= O_p(n^{-1/2}M^2), \quad (A.9)\]

where the last step uses Assumption 4. Similar to (A.9), by Assumption 4 and (A.8), we have
\[\sup_{w \in W} \left| \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i^*(w) \{ Y_i - E(Y_i|X_i) \} \right] \right| = O_p(n^{-1/2}M). \quad (A.10)\]

Note that for any risk function of $w$ such as $R^*(\hat{w})$, we take the expectation in $R^*(\cdot)$ first, then plug in $\hat{w}$. Therefore, by (A.5), (A.6), (A.9), (A.10), we obtain
\[CV^*_K(\hat{w})/n = R^*(\hat{w}) + O_p \left( n^{-1/2}M^2 \right). \quad (A.11)\]

Let $\tau \equiv \sum_{m \in \mathcal{D}} w_m$ and let $\lambda$ be a weight vector with $\lambda_m = 0$ for $m \notin \mathcal{D}$ and $\lambda_m = w_m/(1-\tau)$ for $m \in \mathcal{D}$. For any correctly specified model $m \in \mathcal{D}$, it is easy to see that
\[Y_{(m),n+1}^* - E(Y_{n+1}|X_{n+1}) = 0. \quad (A.12)\]

Then, we have
\[R^*(w) = E \left[ \{ Y_{n+1}^*(w) - E(Y_{n+1}|X_{n+1}) \}^2 \right] \]
\[= E \left( \left[ \sum_{m=1}^{M} w_m \{ Y_{(m),n+1}^* - E(Y_{n+1}|X_{n+1}) \} \right]^2 \right) \]
\[= E \left( \left[ \sum_{m \notin \mathcal{D}} w_m \{ Y_{(m),n+1}^* - E(Y_{n+1}|X_{n+1}) \} \right]^2 \right) \]
\[= (1-\tau) E \left( \left[ \sum_{m \in \mathcal{D}} (1-\tau)^{-1} w_m \{ Y_{(m),n+1}^* - E(Y_{n+1}|X_{n+1}) \} \right]^2 \right) \]
\[= (1-\tau) E \left( \left[ \sum_{m=1}^{M} \lambda_m \{ Y_{(m),n+1}^* - E(Y_{n+1}|X_{n+1}) \} \right]^2 \right) \]
\[\equiv (1-\tau)R^*(\lambda). \quad (A.13)\]

Note that the above result holds for $\hat{\tau}$ and $\hat{\lambda}$ by replacing $w$ with $\hat{w}$ in all equations. Therefore, combining (A.11) and (A.13), we have
\[CV^*_K(\hat{w})/n = (1-\hat{\tau})R^*(\hat{\lambda}) + O_p(n^{-1/2}M^2). \quad (A.14)\]
Let \( \tilde{w} \) be a weight vector with \( \sum_{m \in D} \tilde{w}_m = 1 \). Then, we have \( R^*(\tilde{w}) = 0 \) by (A.12). Hence, by (A.11), it follows that

\[
CV_K^*(\tilde{w})/n = O_p(n^{-1/2}M^2). \quad (A.15)
\]

Next, by (A.14), (A.15), and the fact that \( \hat{w} \) minimizes \( CV_K^*(w) \), we have

\[
(1 - \hat{\tau})R^*(\hat{\lambda}) + O_p(n^{-1/2}M^2) \leq CV_K^*(\tilde{w})/n = O_p(n^{-1/2}M^2). \quad (A.16)
\]

Hence, it follows that

\[
(1 - \hat{\tau}) \inf_{w \in W_S} R^*(w) + O_p(n^{-1/2}M^2) \leq O_p(n^{-1/2}M^2). \quad (A.17)
\]

By (A.17) and Assumption 6, we obtain \( \hat{\tau} \to 1 \) in probability. This completes the proof. \( \square \)

References


