Statistica Sinica **35** (2025), 1025-1049
doi:https://doi.org/10.5705/ss.202022.0321

OPTIMAL MODEL AVERAGING FOR SINGLE-INDEX MODELS WITH DIVERGENT DIMENSIONS

Jiahui Zou, Wendun Wang, Xinyu Zhang* and Guohua Zou

Capital University of Economics and Business,
Erasmus University Rotterdam and Tinbergen Institute,
Chinese Academy of Sciences and Capital Normal University

Abstract: This paper offers a new approach to address the model uncertainty in (potentially) divergent-dimensional single-index models (SIMs). We propose a model-averaging estimator based on cross-validation, which allows the dimension of covariates and the number of candidate models to increase with the sample size. We show that when all candidate models are misspecified, our model-averaging estimator is asymptotically optimal with its squared loss asymptotically identical to that of the infeasible best possible averaging estimator. In a different situation where correct models are available in the model set, the proposed method assigns all weights to the correct models asymptotically. We also propose averaging regularized estimators and prescreening methods to deal with high-dimensional covariates. We illustrate the method via simulations and two empirical applications.

Key words and phrases: Asymptotic optimality, cross-validation, model averaging, single-index model, model screening.

1. Introduction

A linear regression model is a common tool to analyze the relationship between a response variable of interest y and a vector of covariates $\mathbf{x}$ in diversified fields. However, in many applications, such a relationship is nonlinear (Naik and Tsai, 2001; Liang, Wang and Carroll, 2007). A natural extension to relax linearity is to consider a single-index model (SIM) that enables y to depend on $\mathbf{x}$ via an unknown and possibly nonlinear link function g, i.e., $y = g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) + \epsilon$, where $\boldsymbol{\beta}$ is a vector of unknown parameters, and ϵ is the disturbance term. With an unknown link function, this model is more flexible than linear regression models, while maintaining relative ease of interpretation (Horowitz, 1998). It also avoids the curse of dimensionality in many nonparametric models. Various approaches have been proposed to estimate the SIM, e.g., average derivative estimation (Powell, Stock and Stoker, 1989), nonlinear least squares (Ichimura, 1993), and profile least squares (Liang et al., 2010). All of these methods require correct specification of the covariates. However, this knowledge is often unavailable in practice, especially when there are many covariates, so researchers are exposed to a potentially large

*Corresponding author. E-mail: E-mail: xinyu@amss.ac.cn

degree of model uncertainty with respect to which covariates should be included in the model.

A popular method to address model uncertainty is model selection, which picks the "best" model based on certain data-driven criteria, e.g., information criteria. Traditional model selection methods have been extended to SIMs, such as AIC (Naik and Tsai, 2001) and cross-validation (Kong and Xia, 2007). More recently, Cheng, Zeng and Zhu (2017) studied a shrinkage-type estimator to select and estimate covariates in SIMs. As an alternative to model selection, model averaging (MA) addresses model uncertainty by combining estimators from all candidate models with certain weights based on the model performance, and it often leads to a lower risk than model selection (Hansen, 2014a). The past decade has witnessed a burgeoning literature pertaining to model averaging. There are two main streams of averaging techniques: Bayesian model averaging (BMA) and frequentist model averaging (FMA). Although BMA is flexible and can be applied in many models, the choice of prior probabilities is often challenging and experiential (see Hoeting et al. (1999) for an excellent overview). There are various FMA methods, and a partial list includes smoothed information criteria (Buckland, Burnham and Augustin, 1997), optimal averaging (Hansen, 2007), and plug-in methods (Liu, 2015), among many others. Despite the increasing popularity of the averaging techniques, MA estimators for SIMs are hardly studied.

This paper proposes a new model-averaging estimator to address model uncertainty in SIMs. We focus on the optimal averaging method, which aims to achieve an averaged prediction that outperforms any single model. The proposed SIM averaging offers a flexible method to predict the response variable, explicitly considering the model uncertainty. To appropriately choose the averaging weights, we adopt a cross-validation criterion. It is easy to implement and does not require an unbiased estimator of risk, which is difficult to obtain for SIMs. An important merit of our approach is that it allows the dimension of covariates and the number of candidate models to diverge as the sample size increases. Moreover, we study model averaging of regularized estimators with an L_1 penalty as well as model screening to deal with the high-dimensional situations, where the number of covariates is overly large and may even exceed the number of observations. We establish the asymptotic theory of the regularization-based averaging estimator.

We contribute to the model averaging literature in three main respects. First, we propose a new model-averaging estimator for SIMs and establish its asymptotic optimality in terms of minimum squared loss. Since we consider the case of an unknown link function, we rely on a semiparametric approach to estimate each candidate model. Despite a wide range of FMA applications in various models, relatively fewer studies have considered nonlinear or semi-/non-parametric optimal model averaging. Zhang et al. (2016) studied averaging generalized linear (mixed-effects) models. Feng et al. (2022) proposed

an optimal averaging estimator for general nonlinear models. While allowing for a nonlinear relation between covariates and the response variable, these two studies considered parametric models with a given link function. We differ from them by studying a semiparametric model with an unknown link function. The unknown link function requires semiparametric estimation for candidate models, and further motivates a distinct criterion to determine the weights. Semi-/non-parametric averaging was first proposed by Liu (2018) in a non-optimal framework. Li et al. (2018) and Zhang and Wang (2019) then studied optimal model averaging in varying-coefficient and partially linear models, respectively. Recently developments on optimal semi-/non-parametric model averaging include Zhu et al. (2019), Racine et al. (2023) and Zhu et al. (2023) in different contexts. The current paper provides a comprehensive study on the properties of optimal model averaging for SIMs. To the best of our knowledge, this is the first study on semi-/non-parametric model averaging that allows the number of candidate models and the dimension of candidate models to diverge when the sample size increases, which is particularly useful in a high-dimensional setting. Our study also complements Hansen (2014b) by providing theory of the cross-validation-based averaging estimator for nonparametric models.

Second, we study the asymptotic properties of the SIM averaging estimator when the candidate models include correct models. In the framework of linear models, Zhang et al. (2020) showed the consistency of averaging coefficient estimators when there is at least one correct model in the candidate model set. However, no asymptotic results have been established for semi-/non-parametric model averaging when correct models are available. We fill in this gap by providing the asymptotic behavior of our weight estimators for SIMs. We show that our averaging method can consistently choose the correct models by asymptotically assigning all weights to the correct models, no matter whether the candidate models are of finite or diverging dimension. This result complements the asymptotic optimality when all candidate models are misspecified and demonstrates the validity of our method when there are correct models in the model space. Generalizing the weight convergence of parametric model averaging to SIM averaging is by no means trivial, because the estimators to be averaged in parametric models typically have a simple (linear) analytical form (Zhang et al., 2020). Moreover, compared with Zhang et al. (2020), our theoretical analysis is further complicated by allowing a nonlinear relationship between the response variable and covariates.

Last but not least, this paper deals with high-dimensional model averaging, and offers the first study on the properties of averaging *regularized* estimators. Zhang et al. (2020) advocated the use of regularized estimators in a preliminary model screening procedure to deal with high-dimensional covariates, but did not consider combining regularized estimators. In contrast, we propose an innovative regularization-based averaging approach and provide its theoretical justifications.

This approach allows us to deal with the cases in which there are more parameters to estimate than the available observations. Based on the regularized estimation, we also propose a preliminary model screening procedure to shrink the candidate model space. Our methods to deal with high-dimensionality differ from that of Ando and Li (2014), which reduced the dimension by grouping the covariates and only averaging estimators associated with preselected groups. The theoretical analysis on regularization-based averaging and screening can also be applied or extended to parametric and other semi-/non-parametric models.

We verify the theories via an extensive set of simulation experiments. We also apply the proposed method to two real datasets. The first revisits the relationship between financial development and income distribution using country level data, and the second predicts US firm sales growth.

The remainder of this paper is organized as follows. Section 2 introduces the averaging method for SIMs. Section 3 studies its theoretical properties. Section 4 considers regularization-based averaging and model screening. Section 5 presents the simulation study. Section 6 provides the empirical applications. Section 7 concludes. The online Supplementary Material contains more detailed theoretical discussions and simulation studies. Appendix provides additional auxiliary conditions needed for the theory.

2. Model Setup and Estimation

2.1. Single-index model averaging

Assume the following data generating process (DGP), which is also referred to as the true model:

$$y_i = \mu_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where y_i is the response variable of interest with mean μ_i, and the random disturbances $\epsilon_1, \ldots, \epsilon_n$ are independent and (possibly) heteroscedastic with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{E}(\epsilon_i^2) = \sigma_i^2$. Our purpose is to estimate μ_i and thus predict the response variable with p-dimensional covariates $\mathbf{x}_i = (x_{1,i}, \ldots, x_{p,i})^{\mathrm{T}}$ independent with ϵ_j for any $i, j = 1, \ldots, n$, where p is allowed to be finite or divergent when the sample size n increases. To this end, one may employ an SIM that enables us to flexibly model the dependence of μ_i on $\mathbf{x}_i$. However, it is unclear in practice which covariates in $\mathbf{x}_i$ should be used for the prediction. Let $\mathbf{x}_{(s),i}$ be the p_s-dimensional covariate vector whose elements are a subset of $\mathbf{x}_i$. Then the s^{th} candidate SIM model using covariates $\mathbf{x}_{(s),i}$ can be written as

$$y_i = g_{(s)}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}) + \epsilon_{(s),i}, \quad i = 1, \ldots, n; s = 1, \ldots, S_n,$$

where $\boldsymbol{\beta}_{(s)}$ is the associated parameter vector, $g_{(s)}(\cdot)$ is an unknown link function allowed to vary across candidate models, $\epsilon_{(s),i} = y_i - g_{(s)}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)})$, and S_n is the number of candidate models. As we allow p to be potentially divergent, p_s

may also diverge as $n \to \infty$ for some $s \in \{1, \ldots, S_n\}$, leading to an increasing dimension of the parameter vector $\boldsymbol{\beta}_{(s)}$ in some models. Clearly, if a candidate model omits useful covariates, the resulting estimator could be biased, while an overly large model containing many unnecessary covariates leads to an efficiency loss and imposes heavier computational burden, especially in the nonparametric context. We call a specification a misspecified model if it omits certain covariates in the DGP. In contrast, the s^{th} candidate specification is defined as a correct model if there exists a vector $\boldsymbol{\beta}_{(s)}$ such that $\mu_i = g_{(s)}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)})$. Note that a correct model does not necessarily coincide with the DGP, because it may include redundant covariates. Thus, the correct model is not unique.

We propose to tackle such model uncertainty by averaging the estimators obtained from various candidate models, each of which includes a distinct subset of covariates and (likely) a different link function. To estimate each candidate SIM, we follow Ichimura (1993) to achieve the identification of $\boldsymbol{\beta}_{(s)}$ by normalizing its first element to 1 and employ the nonlinear least squares (NLS). One of the advantages of NLS is its light computation burden, which is crucial in our case since our averaging technique is based on a cross-validation criterion and the number of candidate models is typically substantial. To define the NLS estimator, let $k(\cdot)$ be a kernel function. For the s^{th} candidate model, denote h_s as the bandwidth, $k_{h_s}(\cdot) = k(\cdot/h_s)/h_s$, and $\mathbf{K}_{(s)}(\boldsymbol{\beta}_{(s)}) = \{K_{(s),ij}(\boldsymbol{\beta}_{(s)})\}_{n\times n}$ as an $n \times n$ smoothing matrix with (i, j)-element

$$K_{(s),ij}(\boldsymbol{\beta}_{(s)}) = \frac{k_{h_s}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)} - \mathbf{x}_{(s),j}^{\mathrm{T}}\boldsymbol{\beta}_{(s)})}{\sum_{j^*=1}^{n} k_{h_s}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)} - \mathbf{x}_{(s),j^*}^{\mathrm{T}}\boldsymbol{\beta}_{(s)})}.$$

Further define $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ and $\mathbf{X}_{(s)} = (\mathbf{x}_{(s),1}, \ldots, \mathbf{x}_{(s),n})^{\mathrm{T}}$. The NLS estimator $\widehat{\boldsymbol{\beta}}_{(s)}$ for the s^{th} candidate model can then be obtained by minimizing the following objective function:

$$H_{(s),n}(\boldsymbol{\beta}_{(s)}) = n^{-1}\left\|\mathbf{y} - \mathbf{K}_{(s)}(\boldsymbol{\beta}_{(s)})\mathbf{y}\right\|^2. \tag{2.1}$$

The resulting estimator of $\boldsymbol{\mu}$ from the s^{th} candidate model is $\widehat{\boldsymbol{\mu}}_{(s)} = \mathbf{K}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})\mathbf{y}$. With the estimator of each candidate model, we can obtain the model averaging estimator of $\boldsymbol{\mu}$ as

$$\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{s=1}^{S_n} w_s \widehat{\boldsymbol{\mu}}_{(s)} = \mathbf{K}(\mathbf{w}, \widehat{\boldsymbol{\beta}})\mathbf{y},$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{(1)}^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\beta}}_{(S_n)}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{K}(\mathbf{w}, \widehat{\boldsymbol{\beta}}) = \sum_{s=1}^{S_n} w_s \mathbf{K}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})$, and the weight vector $\mathbf{w} = (w_1, \ldots, w_{S_n})^{\mathrm{T}}$ belongs to the set $\mathcal{W} = \{\mathbf{w} \in [0, 1]^{S_n} : \sum_{s=1}^{S_n} w_s = 1\}$. The averaging estimator $\widehat{\boldsymbol{\mu}}(\mathbf{w})$ offers an appealing method to predict the response variable.

2.2. Choosing the averaging weights

Given our main goal of prediction, our weight choice aims at minimizing the squared loss $L_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2$. We propose to choose the averaging weights by minimizing a J-fold cross-validation (CV) criterion in a similar manner to jackknife model averaging (Hansen and Racine, 2012). Our approach is a numerical model-averaging method that is easy to implement and hardly relies on the structure of the model, except the dependence features of data. Unlike the Mallows criterion, this method is more flexible, since it does not require an unbiased estimator of risk, which is often difficult to obtain for complex models such as the single-index model considered here.

To implement the J-fold CV, we divide the dataset into J_n blocks so that there are $M_n = \lfloor n/J_n \rfloor$ observations in each block, where $\lfloor \cdot \rfloor$ denotes the integer part of a number. For the s^{th} candidate model, let $\widehat{\boldsymbol{\beta}}_{(s)}^{[-j]}$ be the NLS estimator of $\boldsymbol{\beta}_{(s)}$ without using the observations from the j^{th} block for $j = 1, \ldots, J_n$. Then, the corresponding leave-block-out kernel estimator is $\widetilde{\boldsymbol{\mu}}_{(s)} = (\widetilde{\mu}_{(s),1}, \ldots, \widetilde{\mu}_{(s),n})^{\mathrm{T}}$ with

$$\begin{aligned}
\widetilde{\mu}_{(s),1} &= \left(\mathbf{0}_{M_n}^{\mathrm{T}}, \frac{k_{h_s}(\mathbf{x}_{(s),M_n+1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-1]} - \mathbf{x}_{(s),1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-1]})}{\sum_{M_n < i \le n} k_{h_s}(\mathbf{x}_{(s),i}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-1]} - \mathbf{x}_{(s),1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-1]})}, \right. \\
&\qquad \left. \ldots, \frac{k_{h_s}(\mathbf{x}_{(s),n}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-1]} - \mathbf{x}_{(s),1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-1]})}{\sum_{M_n < i \le n} k_{h_s}(\mathbf{x}_{(s),i}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-1]} - \mathbf{x}_{(s),1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-1]})} \right)^{\mathrm{T}} \mathbf{y}, \\
&\vdots \\
\widetilde{\mu}_{(s),n} &= \left(\frac{k_{h_s}(\mathbf{x}_{(s),1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-J_n]} - \mathbf{x}_{(s),n}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-J_n]})}{\sum_{1 \le i \le n - M_n} k_{h_s}(\mathbf{x}_{(s),i}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-J_n]} - \mathbf{x}_{(s),n}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-J_n]})}, \right. \\
&\qquad \left. \ldots, \frac{k_{h_s}(\mathbf{x}_{(s),n-M_n}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-J_n]} - \mathbf{x}_{(s),n}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-J_n]})}{\sum_{1 \le i \le n - M_n} k_{h_s}(\mathbf{x}_{(s),i}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-J_n]} - \mathbf{x}_{(s),n}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}^{[-J_n]})}, \mathbf{0}_{M_n}^{\mathrm{T}} \right)^{\mathrm{T}} \mathbf{y},
\end{aligned}$$

where $\mathbf{0}_{M_n}$ is an M_n-dimensional vector of zeros. The above equations suggest that there is a matrix $\widetilde{\mathbf{K}}_{(s)}(\widetilde{\boldsymbol{\beta}}_{(s)})$ with $\widetilde{\boldsymbol{\beta}}_{(s)} = (\widehat{\boldsymbol{\beta}}_{(s)}^{[-1]\mathrm{T}}, \ldots, \widehat{\boldsymbol{\beta}}_{(s)}^{[-J_n]\mathrm{T}})^{\mathrm{T}}$, such that the leave-block-out estimator of $\boldsymbol{\mu}$ under the s^{th} candidate model can be written as $\widetilde{\boldsymbol{\mu}}_{(s)} = \widetilde{\mathbf{K}}_{(s)}(\widetilde{\boldsymbol{\beta}}_{(s)})\mathbf{y}$. Let $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_{(1)}^{\mathrm{T}}, \ldots, \widetilde{\boldsymbol{\beta}}_{(S_n)}^{\mathrm{T}})^{\mathrm{T}}$ and $\widetilde{\mathbf{K}}(\mathbf{w}, \widetilde{\boldsymbol{\beta}}) = \sum_{s=1}^{S_n} w_s \widetilde{\mathbf{K}}_{(s)}(\widetilde{\boldsymbol{\beta}}_{(s)})$. The averaging leave-block-out estimator of $\boldsymbol{\mu}$ is then given by

$$\widetilde{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{s=1}^{S_n} w_s \widetilde{\boldsymbol{\mu}}_{(s)} = \widetilde{\mathbf{K}}(\mathbf{w}, \widetilde{\boldsymbol{\beta}})\mathbf{y}.$$

The J-fold CV criterion can be obtained by $\mathrm{CV}_{J_n}(\mathbf{w}) = \|\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \mathbf{y}\|^2$, and the weight vector is chosen by minimizing $\mathrm{CV}_{J_n}(\mathbf{w})$ over $\mathbf{w} \in \mathcal{W}$, i.e.,

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{W}}{\operatorname{argmin}}\, \mathrm{CV}_{J_n}(\mathbf{w}). \tag{2.2}$$

The resulting averaging estimator of $\boldsymbol{\mu}$ is

$$\widehat{\boldsymbol{\mu}}(\widehat{\mathbf{w}}) = \sum_{s=1}^{S_n} \widehat{w}_s \widehat{\boldsymbol{\mu}}_{(s)} = \mathbf{K}(\widehat{\mathbf{w}}, \widehat{\boldsymbol{\beta}})\mathbf{y},$$

which we refer to as the J-fold CV model-averaging (JCVMA) estimator.

Since the CV objective function can be rewritten as $\mathrm{CV}_{J_n}(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}$, where the element of $\mathbf{A}$ is $A_{s,m} = \{\mathbf{K}_{(s)}(\widetilde{\boldsymbol{\beta}}_{(s)})\mathbf{y} - \mathbf{y}\}^{\mathrm{T}}\{\mathbf{K}_{(m)}(\widetilde{\boldsymbol{\beta}}_{(m)})\mathbf{y} - \mathbf{y}\}$ for $s, m = 1, \ldots, S_n$, the weight calculation in (2.2) is a quadratic programming problem which is easy to solve. The computational cost of this optimization mainly lies in the CV estimator of $\{\widetilde{\boldsymbol{\mu}}_{(s)}\}_{s=1}^{S_n}$, and will be substantial if S_n and J_n are large. We shall discuss how to determine S_n and which models to combine when the entire model space is huge in Section 4.

3. Asymptotic Properties

This section studies the asymptotic properties of the proposed averaging estimator. We first examine the squared loss of our averaging estimator when all candidate models are misspecified. Then we consider the case where the set of candidate models includes the correct (but not necessarily true) models.

3.1. Asymptotic optimality

This section studies the property of JCVMA when none of the candidate models is correct. The following regularity conditions are required for the asymptotic optimality of the JCVMA estimator, most of which are standard in the literature. All limiting processes below correspond to $n \to \infty$ unless stated otherwise.

Condition 1. *For $s = 1, \ldots, S_n$ and $r = 1, \ldots, p_s$, the r^{th} element of $\widehat{\boldsymbol{\beta}}_{(s)}$ obtained from* (2.1)*, $\widehat{\beta}_{(s),r}$, has a limiting value $\beta^*_{(s),r}$.*

This condition ensures the existence of the limit of $\widehat{\boldsymbol{\beta}}_{(s)}$, which is often referred to as the "quasi-true" parameter. Similar conditions are imposed in many model averaging studies, such as Zhang et al. (2016), Ando and Li (2017). We further denote $\boldsymbol{\beta}^*_{(s)} = (\beta^*_{(s),1}, ..., \beta^*_{(s),p_s})^{\mathrm{T}}$, $\boldsymbol{\mu}^*_{(s)} = \mathbf{K}_{(s)}(\boldsymbol{\beta}^*_{(s)})\mathbf{y}$ and $\widetilde{\boldsymbol{\mu}}^*_{(s)} = \widetilde{\mathbf{K}}_{(s)}(\mathbf{1}_{J_n} \otimes \boldsymbol{\beta}^*_{(s)})\mathbf{y}$, where $\otimes$ represents the Kronecker product and $\mathbf{1}_{J_n}$ is a $J_n \times 1$ vector of 1.

Condition 2.

(i) $\sigma_{\max} = O(1)$, *where* $\sigma_{\max} = \max_{1\le i\le n} \sigma_i$.

(ii) $\max_{1\le i\le n} |\mu_i| = O(1)$.

Condition 2 restricts the magnitude of the variance and the mean of y_i. It is satisfied if the response variable has a finite support. Similar conditions are imposed by (Ando and Li, 2017, Asms. A1 and A4) and (Zhu et al., 2019, Conds. C.1 and C.7).

Condition 3. *There exists a positive sequence* $\{d_n\}$ *such that*

$$\max_{1\le s\le S_n}\max_{1\le i\le n}\max_{1\le j\le n} K_{(s),ij}(\boldsymbol{\beta}^*_{(s)}) = O_P(d_n), \tag{3.1}$$

$$n\max_{1\le s\le S_n}\max_{1\le j\le n}\sum_{i\ne j}^{n} K^2_{(s),ij}(\boldsymbol{\beta}^*_{(s)}) = O_P(1), \tag{3.2}$$

where $K_{(s),ij}(\boldsymbol{\beta}^*_{(s)}) = k_{h_s}(\mathbf{x}^{\mathrm{T}}_{(s),i}\boldsymbol{\beta}^*_{(s)} - \mathbf{x}^{\mathrm{T}}_{(s),j}\boldsymbol{\beta}^*_{(s)})/\sum_{j^*=1}^{n} k_{h_s}(\mathbf{x}^{\mathrm{T}}_{(s),i}\boldsymbol{\beta}^*_{(s)} - \mathbf{x}^{\mathrm{T}}_{(s),j^*}\boldsymbol{\beta}^*_{(s)})$.

This condition controls the magnitude of the kernel weight $K_{(s),ij}(\boldsymbol{\beta}^*_{(s)})$ assigned to observation i's neighbors j for each set of covariates $\mathbf{x}_{(s)}$. Such weights measure how much an observation contributes to the estimator $\widehat{\boldsymbol{\mu}}_{(s)}$ for the s^{th} candidate model. Combining with the fact that the row sum of the smoothing matrix $\mathbf{K}_{(s)}(\boldsymbol{\beta}_{(s)})$ is 1, this condition implies that all observations receive nonzero kernel weights, such that they all contribute to estimating $\boldsymbol{\mu}_{(s)}$ to different extents. If $\{\mathbf{x}_i\}$ is independent and identically distributed, (3.1) and (3.2) hold with probability one.

Condition 4. *The smoothing matrix satisfies*

$$\max_{1\le s\le S_n}\max_{1\le j\le n}\sum_{i=1}^{n} K_{(s),ij}(\boldsymbol{\beta}^*_{(s)}) = O_P(1).$$

Condition 4 concerns the L_∞ norms of $\mathbf{K}_{(s)}$ and is widely used in nonparametric models, such as Assumption 1.3.3(i) of Härdle, Liang and Gao (2007).

In addition, we also need to guarantee that the NLS estimator of $\boldsymbol{\beta}$-parameters and its cross-validation version for each candidate model are consistent in the sense that they converge to their corresponding quasi-true values. The conditions involved are standard in the literature (Ichimura, 1993), which are provided in the Appendix A.1.

Lemma 1. *Under Conditions* 1–4 *and* A.1–A.5 *in the Appendix, we have that*

$$\max_{1\le s\le S_n}\sqrt{\frac{n}{S_n p_s}}\left\|\widehat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}^*_{(s)}\right\| = O_P(1),$$

and

$$\max_{1\le j\le J_n}\max_{1\le s\le S_n}\sqrt{\frac{n-M_n}{S_n p_s}}\left\|\widehat{\boldsymbol{\beta}}^{[-j]}_{(s)} - \boldsymbol{\beta}^*_{(s)}\right\| = O_P(1).$$

This lemma states that the NLS estimator $\widehat{\boldsymbol{\beta}}_{(s)}$ obtained from minimizing (2.1) and its CV version $\widehat{\boldsymbol{\beta}}_{(s)}^{[-j]}$ (leaving observations of the j^{th} block out) both converge to the quasi-true value $\boldsymbol{\beta}_{(s)}^{*}$ of the s^{th} model at the uniform speed of $\sqrt{n/S_n p_s}$ and $\sqrt{(n-M_n)/S_n p_s}$, respectively. Importantly, these convergence results hold in both finite- and divergent-dimensional cases, which is vital for establishing the asymptotic optimality.

To state the next condition regarding the second-order derivatives of the link function, let $\lceil\cdot\rceil$ be the ceiling of a number and we denote

$$\widehat{g}_{(s)}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}) = \frac{\sum_{j=1}^{n} y_j k_{h_s}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)} - \mathbf{x}_{(s),j}^{\mathrm{T}}\boldsymbol{\beta}_{(s)})}{\sum_{j^*=1}^{n} k_{h_s}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)} - \mathbf{x}_{(s),j^*}^{\mathrm{T}}\boldsymbol{\beta}_{(s)})},$$
$$\widehat{g}_{(s)}^{[-\mathcal{B}(i)]}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}) = \frac{\sum_{j=1}^{n} y_j k_{h_s}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)} - \mathbf{x}_{(s),j}^{\mathrm{T}}\boldsymbol{\beta}_{(s)})}{\sum_{j^*\in\mathcal{A}(i)} k_{h_s}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)} - \mathbf{x}_{(s),j^*}^{\mathrm{T}}\boldsymbol{\beta}_{(s)})},$$

where the superscript $[-\mathcal{B}(i)]$ denotes the estimator without using the entire block that contains the i^{th} observation, i.e., $\mathcal{B}(i) = \{\lceil i/M_n\rceil M_n - M_n + 1, \ldots, \lceil i/M_n\rceil M_n\}$, and $\mathcal{A}(i) = \{1,\ldots,n\}\backslash\mathcal{B}(i)$. We further denote $\mathcal{O}(\boldsymbol{\beta}_{(s)}^{*},\rho)$ as a neighborhood of $\boldsymbol{\beta}_{(s)}^{*}$ for some positive constant ρ, i.e., $\{\boldsymbol{\beta}_{(s)} \in \mathcal{R}^{p_s} : \|\boldsymbol{\beta}_{(s)} - \boldsymbol{\beta}_{(s)}^{*}\| \le \rho\}$, and $\lambda_{\max}(\cdot)$ as the maximum eigenvalue.

Condition 5. *There exists a $\rho > 0$ such that*

$$\max_{1\le s\le S_n} \sup_{\substack{\boldsymbol{\beta}_{(s)}^{(1)},\ldots,\boldsymbol{\beta}_{(s)}^{(n)} \\ \in\mathcal{O}(\boldsymbol{\beta}_{(s)}^{*},\rho)}} \lambda_{\max}\left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial\widehat{g}_{(s)}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^{(i)})}{\partial\boldsymbol{\beta}_{(s)}} \frac{\partial\widehat{g}_{(s)}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^{(i)})}{\partial\boldsymbol{\beta}_{(s)}^{\mathrm{T}}}^{\mathrm{T}}\right\} = O_P(p_{\max}), \tag{3.3}$$

and

$$\max_{1\le s\le S_n} \sup_{\substack{\boldsymbol{\beta}_{(s)}^{(1)},\ldots,\boldsymbol{\beta}_{(s)}^{(n)} \\ \in\mathcal{O}(\boldsymbol{\beta}_{(s)}^{*},\rho)}} \lambda_{\max}\left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial\widehat{g}_{(s)}^{[-\mathcal{B}(i)]}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^{(i)})}{\partial\boldsymbol{\beta}_{(s)}} \frac{\partial\widehat{g}_{(s)}^{[-\mathcal{B}(i)]}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^{(i)})}{\partial\boldsymbol{\beta}_{(s)}^{\mathrm{T}}}^{\mathrm{T}}\right\} = O_P(p_{\max}), \tag{3.4}$$

where $p_{\max} = \max_{1\le s\le S_n} p_s$ denotes the maximum dimension of candidate models.

In this condition, (3.3) essentially controls the magnitude of $\|\widehat{\boldsymbol{\mu}}_{(s)} - \boldsymbol{\mu}_{(s)}^{*}\|^2$ through the differential mean value theorem, and (3.4) is a CV version of (3.3) due to the J-fold CV estimator, which controls the magnitude of $\|\widetilde{\boldsymbol{\mu}}_{(s)} - \widetilde{\boldsymbol{\mu}}_{(s)}^{*}\|^2$. This condition is satisfied if $\{g_{(s)}(\cdot)\}_{s=1}^{S_n}$ is sufficiently smooth, such as $\max_{1\le s\le S_n} \|\partial g_{(s)}(\mathbf{x}_{(s)}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^{*})/\partial\boldsymbol{\beta}_{(s)}^{*}\|^2 = O(p_{\max})$ for each $s = 1,\ldots,S_n$. Similar conditions are often used when studying model averaging for parametric models, e.g., Condition C.4 of Zhang et al. (2016) and Condition C.2 of Zhang et al. (2020).

Denote $\boldsymbol{\mu}^*(\mathbf{w}) = \sum_{s=1}^{S_n} w_s \mathbf{K}_{(s)}(\boldsymbol{\beta}^*_{(s)})\mathbf{y}$ as the averaging estimator based on the quasi-true parameters $\boldsymbol{\beta}^*_{(s)}$ for $s = 1, \ldots, S_n$. Denote $L_n^*(\mathbf{w}) = \|\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu}\|^2$ as the corresponding squared loss, and $\xi_n = \inf_{\mathbf{w}\in\mathcal{W}} L_n^*(\mathbf{w})$ as the minimum squared loss over all averaging estimators. We assume the following condition.

Condition 6.

(i) $\xi_n^{-1} S_n^{1/2} n p_{\max} (n - M_n)^{-1/2} = o_P(1)$.

(ii) $\xi_n^{-1} d_n M_n n = o_P(1)$.

This set of conditions resembles (8) in Theorem 1 of Wan, Zhang and Zou (2010) and Condition C.2 of Zhu et al. (2019), which essentially requires that all candidate models are misspecified to a non-trivial extent, such that their mean squared errors are not too small. Thus, it precludes any scenario in which the correct models are included in the set of candidate models. Note that this condition does not conflict with Condition 5 because they concern different distance measures. In particular, Condition 5 controls the distance between the estimators of $\boldsymbol{\mu}_{(s)}$ (i.e., $\widehat{\boldsymbol{\mu}}_{(s)}$ and $\widetilde{\boldsymbol{\mu}}_{(s)}$) and their corresponding quasi-true values (i.e., $\boldsymbol{\mu}^*_{(s)}$ and $\widetilde{\boldsymbol{\mu}}^*_{(s)}$), while Condition 6 concerns the degree of misspecification which is the distance between the quasi-true value $\boldsymbol{\mu}^*_{(s)}$ and the true value $\boldsymbol{\mu}$.

Moreover, Condition 6 also provides restrictions on the relative divergent rates of $p_{\max}, S_n, M_n$ and ξ_n, namely ξ_n is required to grow at a rate no slower than $S_n^{1/2} n p_{\max}(n - M_n)^{-1/2}$ and $d_n M_n n$. For example, if ξ_n explodes at a rate of $n^{1-\alpha}$ for some $\alpha > 0$, then $S_n^{1/2} p_{\max}(n - M_n)^{-1/2} n^{\alpha}$ and $d_n M_n n^{\alpha}$ are both required to converge to 0, which further implies that α needs to be small. If ξ_n explodes at a rate of n, $S_n = O(1)$ and $M_n = O(1)$, then we can allow $p_{\max}$ to grow at a rate of $n^{1/2-c}$ for some positive constant $c < 1/2$. Overall, Condition 6 is more likely to be satisfied when ξ_n approaches infinity at a faster rate, or in other words, when all candidate models are misspecified to a larger extent such that the squared loss of the best possible averaging estimator is large.

Theorem 1. *Under the conditions of Lemma* 1 *and Conditions* 5 *and* 6*, we have that*

$$\frac{L_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})} \to 1 \quad \textit{in probability.}$$

Theorem 1 shows that the JCVMA estimator of $\boldsymbol{\mu}$ is asymptotically optimal in the sense that it leads to a squared loss that is asymptotically identical to that of the infeasible best possible model-averaging estimator.

3.2. Weight convergence

This section studies the limiting behavior of averaging weights when the set of candidate models includes at least one correct model. Without loss of generality, we assume that the first S_0 (≥ 1) models are correct. We denote $\widehat{w}_\Delta = \sum_{s=1}^{S_0} \widehat{w}_s$ as

the sum of weights given to the S_0 correct models, where $\widehat{w}_s$ is the s^{th} element of the JCVMA weight vector $\widehat{\mathbf{w}}$. Denote $\mathcal{W}_F = \{\mathbf{w} \in \mathcal{W} : w_s = 0 \text{ for } s = 1, \ldots, S_0\}$ as the set of weight vectors that assign zero weights to the correct models. Let $\xi_F = \inf_{\mathbf{w}\in\mathcal{W}_F} L_n^*(\mathbf{w})$ be the squared loss of only averaging misspecified models.

Condition 7.

(i) $\xi_F^{-1} S_n^{1/2} n p_{\max} (n - M_n)^{-1/2} = o_P(1)$.

(ii) $\xi_F^{-1} d_n M_n n = o_P(1)$.

Condition 7 replaces ξ_n in Condition 6 by ξ_F, and can be regarded as a counterpart of Condition 6 for the cases in which correct models are present in the model space. It requires that the squared loss of the best possible averaging of misspecified models has a sufficiently large divergent rate, in order to distinguish between the misspecified and correct models. A similar set of conditions is discussed for linear regressions in Zhang et al. (2020).

Theorem 2. *If the conditions of Lemma* 1 *and Conditions* 5 *and* 7 *hold, then* $\widehat{w}_\Delta \to 1$ *in probability.*

Theorem 2 shows that JCVMA tends to assign all weights to the correct models if they exist in the candidate model set. Consistent selection of the correct models enables us to examine the (nonlinear) relation between covariates and the response variable.

To conclude the prediction performance when candidate models include the correct models, we need the following extra condition for ξ_F, S_n and $p_{\max}$.

Condition 8.

(i) $\xi_F^{-1} S_n p_{\max}^2 = o_P(1)$.

(ii) $\xi_F^{-3} n^3 \{S_n^{1/2} p_{\max} (n - M_n)^{-1/2} + d_n M_n\} = o_P(1)$.

Condition 8(i) imposes a stronger restriction on the speed that S_n and $p_{\max}$ diverge. Condition 8(ii) adds the two equalities in Condition 7, and multiplies the left-hand side by $\xi_F^{-2} n^2$. Note that the limit of $\xi_F^{-2} n^2$ is usually not zero, so $\xi_F^{-1} S_n^{1/2} n p_{\max} (n - M_n)^{-1/2}$ and $\xi_F^{-1} d_n M_n n$ must converge to zero faster than those in Condition 7. More specifically, combining Condition 8(ii) with the fact that $\xi_F = O_P(n)$, a result implied by the expression of $L_n^*(\mathbf{w})$, we have that $\xi_F^{-1}\{S_n^{1/2} n p_{\max} (n - M_n)^{-1/2} + d_n M_n n\} = \xi_F^2 n^{-2} \xi_F^{-3} \{S_n^{1/2} n^3 p_{\max} (n - M_n)^{-1/2} + d_n M_n n^3\} = o_P(1)$, which further implies Condition 7.

Corollary 1. *If the conditions of Lemma* 1 *and Conditions* 5 *and* 8 *hold, then*

$$\frac{L_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}_F} L_n(\mathbf{w})} \to 0 \quad \textit{in probability.}$$

This corollary establishes the asymptotic optimality when correct models are contained in the candidate model set, which complements the asymptotic optimality in Theorem 1. It shows that when correct models are available in the candidate set, the squared loss of the JCVMA estimator of $\boldsymbol{\mu}$, namely $L_n(\widehat{\mathbf{w}})$, is asymptotically negligible compared to that of any averaging estimator that assigns zero weights to the correct models. In conjunction with Theorem 2, this corollary suggests that JCVMA also provides good prediction when correct models are available, since it asymptotically assigns all weights to the correct models and thus outperforms those averaging estimators that lack weight convergence and assign zero weights to correct models.

4. Averaging Regularized Estimators and Prescreening

Thus far, we have studied SIM averaging when $p_{\max} < n$ and S_n is not too large, even though both of them are allowed to diverge as n increases. In some applications, there may exist a huge number of potential covariates such that some candidate models have more parameters to estimate than the sample size and the number of all possible models is overly large. Hence, in this section, we study how to perform JCVMA in such situations. We first consider averaging regularized estimators for candidate models in the presence of many covariates, and then study how to choose S_n and the set of candidate models when the entire model space is too large to be completely considered.

4.1. Averaging regularized estimators

When there exist a large number of covariates, NLS estimators obtained from solving (2.1) can be rather inefficient and sometimes even infeasible for some candidate models due to (too) many parameters. Hence, we consider an alternative method to estimate the s^{th} candidate SIM using NLS with an L_1 penalty. Particularly, the estimator of $\boldsymbol{\beta}_{(s)}$ for the s^{th} candidate model can be obtained as

$$\widehat{\boldsymbol{\beta}}^R_{(s)} = \underset{\boldsymbol{\beta}_{(s)}}{\operatorname{argmin}} \left\{ H_{(s),n}(\boldsymbol{\beta}_{(s)}) + \lambda_s \|\boldsymbol{\beta}_{(s)}\|_1 \right\}, \tag{4.1}$$

where $H_{(s),n}(\boldsymbol{\beta}_{(s)})$ is the NLS objective function of the s^{th} model defined in (2.1), $\|\boldsymbol{\beta}_{(s)}\|_1 = \sum_{i=1}^{p_s} |\beta_i|$ is the penalty, and λ_s is the model-specific tuning parameter. The above optimization problem can be solved, e.g., by the coordinate descent algorithm (Friedman, Hastie and Tibshirani, 2010). Furthermore, we denote $\widehat{\boldsymbol{\mu}}^R_{(s)} = \mathbf{K}_{(s)}(\widehat{\boldsymbol{\beta}}^R_{(s)})\mathbf{y}$, $\widehat{\boldsymbol{\mu}}^R(\mathbf{w}) = \sum_{s=1}^{S_n} w_s \widehat{\boldsymbol{\mu}}^R_{(s)}$ and $L^R_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}^R(\mathbf{w}) - \boldsymbol{\mu}\|^2$.

To study the property of the regularization-based JCVMA estimator, we need conditions to ensure the consistency of regularized candidate estimators and to control the speed of S_n and d_n, parallel to Lemma 1 and Condition 6 for the unregularized cases. To save space, we relegate the regularization version of similar conditions to the online Supplementary Material.

Corollary 2. *If Conditions* 2–4 *and* S.6–S.8 *in the online Supplementary Material hold, then*

$$\frac{L_n^R(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}} L_n^R(\mathbf{w})} \to 1 \quad \textit{in probability.}$$

Corollary 2 shows that the asymptotic optimality of JCVMA continues to hold when the candidate SIMs are estimated by NLS with an L_1 penalty. In conjunction with model screening discussed in the following subsection, the regularization technique offers a way to implement model averaging when the number of covariates exceeds the sample size.

4.2. Model averaging based on prescreening

When p is particularly large or even exceeds the sample size, not only some candidate models are difficult to estimate, but the model space is also huge, rendering estimation and combination of all possible models infeasible. In this case, we can implement a model-screening step prior to averaging, which we refer to as prescreening. Pre-screening can be used when $p < n$ but all possible combinations of covariates still lead to excessively numerous candidate models, i.e., 2^p is large, and it is also useful in the high-dimensional cases in which $p > n$. We propose two approaches to pre-screen the models and construct the set of candidate models for averaging, depending on the relation between p and n.

First, when $p < n$ and estimating the full model is feasible, we can order the covariates based on their marginal correlations with the response variable, and construct the set of candidate models by including one extra covariate at each time based on the ordering. The idea of model screening based on bivariate correlation is in a similar spirit as that of the "sure independence screening" proposed by Fan and Lv (2008). Similar screening procedures have been used in other model-averaging studies, such as Claeskens, Croux and van Kerckhoven (2006) and Ando and Li (2014).

Second, when $p > n$, it is impossible to estimate the full model using the standard NLS as in (2.1), and we propose to pre-screen the models based on regularized estimation of the full model as in (4.1). Particularly, we can solve the following optimization problem: $\min_{\boldsymbol{\beta}} \{H_n(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\}$, where $H_n(\boldsymbol{\beta})$ is the same objective function as (2.1) but using all the covariates, and λ is the tuning parameter. With a feasible amount of different values of λ, we can obtain a set of corresponding candidate estimators, which can then be conveniently averaged. The idea of using regularized estimation for screening is advocated by Zhang et al. (2016), but they only consider *parametric* models and provide no theory.

To justify the SIM averaging estimator obtained after a preliminary model screening step, we can show that it remains asymptotically optimal, i.e., the squared loss of the postscreening JCVMA is asymptotically identical to that of the infeasible best possible model-averaging estimator obtained from the original model set $\mathcal{W}$ (without prescreening). We provide a summary of assumptions and

a rough illustration of the idea to prove this result as follows. Let $\mathcal{D}$ be a (random) subset of $\{1,\ldots,S_n\}$ and $\mathcal{W}^{\mathcal{D}} = \{\mathbf{w} \in [0,1]^{S_n} : \sum_{s\in\mathcal{D}} w_s = 1 \text{ and } \sum_{s\notin\mathcal{D}} w_s = 0\}$ be a subset of $\mathcal{W}$. Note that $\mathcal{W}^{\mathcal{D}}$ is also random due to the randomness of $\mathcal{D}$. The postscreening model-averaging estimator based on the subset $\mathcal{D}$ is obtained by using the weight vector $\widehat{\mathbf{w}}^s = \arg\min_{\mathbf{w}\in\mathcal{W}^{\mathcal{D}}} \mathrm{CV}_{J_n}(\mathbf{w})$. We make an additional assumption that there exists a non-negative series of $\{\nu_n\}$ and a weight series of $\{\mathbf{w}_n\} \in \mathcal{W}$, such that $\xi_n^{-1}\nu_n = o_P(1)$, $\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{CV}_{J_n}(\mathbf{w}) = \mathrm{CV}_{J_n}(\mathbf{w}_n) - \nu_n$, and $\Pr(\mathbf{w}_n \in \mathcal{W}^{\mathcal{D}}) \to 1$. This assumption enures that there exists a weight in $\mathcal{W}^{\mathcal{D}}$ to achieve the minimal CV loss asymptotically. This is the same as Assumption 1 in Zhang et al. (2016), in which more explanations are provided. Under this additional condition as well as the conditions of Theorem 1, we can then use the same arguments as Theorem 3 of Zhang et al. (2016) to show that the postscreening model-averaging estimator based on the candidate model set $\mathcal{W}_n^{\mathcal{D}}$ still achieves the asymptotic optimality, namely $L_n(\widehat{\mathbf{w}}^s)/\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w}) \to 1$ in probability.

5. Simulation Study

We examine the finite-sample performance of JCVMA and compares it with the popular model selection and averaging methods. A brief presentation of the DGP and results are provided here, and more details are in the online Supplementary Material.

5.1. Simulation setup

We consider two nonlinear link functions, the sine function and Tobit model. For each nonlinear function, we study two cases that differ in the dimension of covariates. First, we fix the dimension of covariates to be finite. Second, we allow the dimension of covariates and the number of candidate models to be divergent. Furthermore, for each of the cases, we consider whether the correct models are included in the set of candidate models. We consider the sample sizes for estimation as $n \in \{100, 200, 300, 400, 500\}$, and set the testing size as 1,000; all results are based on $D = 1000$ replications.

We compare JCVMA with three information criteria: AIC, BIC and a variant of AIC, denoted as AICC, which is designed especially for SIMs proposed by Naik and Tsai (2001). We also compare the smoothed versions of the three information criteria, which use the values of the criteria for each candidate model as weights to construct the averaging estimators, namely, SAIC, SBIC and SAICC.

5.2. Simulation results

We evaluate the performance of the methods from three perspectives. First, to verify the asymptotic optimality of the JCVMA in Theorem 1, we report the relative squared loss of each method with respect to the best possible averaging

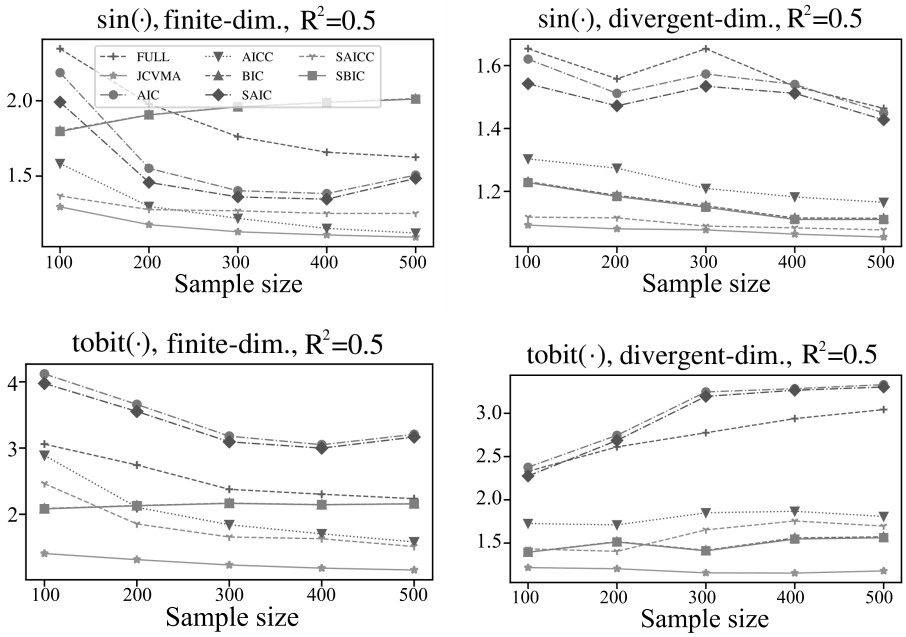


Figure 1. Relative squared loss when all candidate models are misspecified.

estimator in Figure 1, where the best possible averaging weight is calculated by minimizing $\|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2$ over $\mathcal{W}$ given the true value $\boldsymbol{\mu}$ in the testing set. To save space, we only report the results of $R^2 = 0.5$, which is closest to our empirical datasets. Increasing R^2 improves the performance of all methods, but the conclusion regarding the relative performance of all methods remains the same. It shows that the proposed JCVMA produces the lowest relative squared loss for both cases of fixed and divergent dimensions and for all sample sizes. Moreover, the relative squared loss of JCVMA generally decreases and tends to one when the sample size increases. The convergence of JCVMA confirms its asymptotic optimality as stated in Theorem 1. In contrast, the curves of other averaging estimators do not show clear convergence to one. We also consider the normalized mean squared prediction error (NMSPE) as an alternative measure of prediction performance, and the results are available in the online Supplementary Material.

Next, to verify the convergence of weights when correct models exist in the candidate model set as shown in Theorem 2, we plot the weights assigned to the correct models when n increases in Figure 2. Generally, we find that the sum of these weights is monotonically increasing and converges to one as n enlarges. When R^2 increases, the sum of weights on correct models and the convergence rate of this sum both improve. These results confirm the validity of Theorem 2.

An important implication of the weight convergence is that JCVMA has a smaller squared loss than other averaging estimators that lack weight convergence and fail to assign positive weights to correct models, as shown in Corollary 1. Figure 3 confirms this corollary, showing that the relative squared loss of JCVMA with respect to the best possible averaging estimators that only use misspecified models is indeed less than one and generally decreases as n increases. This relative squared loss also decreases when R^2 increases.

We also consider the cases when $p > n$ and employ model screening. The results are qualitatively similar and provided in the online Supplementary Material.

6. Empirical Applications

In this section, we apply our method to two empirical applications. The first studies the relationship between financial development and inequality using cross-country data, and the second examines US firm sales growth. While the use of aggregated data largely averages out micro-level noise, the cross-country study may be subject to substantial heterogeneity and omitted variables. In contrast, the corporate analysis using firm-level data enjoys more abundant and homogeneous data, but it also unavoidably suffers from potential outliers. Thus, by analyzing two types of datasets, we can explore how our method performs in different environments.

6.1. Financial development and income distribution

Given the substantial cross-country difference in inequality and the level of financial development, it is of particular interest for both academics and policy makers to understand whether and how financial development affects the income distribution. In this section, we revisit the relationship between financial development and the distribution of income, first studied by Beck, Demirgüç-Kunt and Levine (2007). Our response variable is the growth rate of Gini coefficient (G). We measure financial development by private credit (P), which is a logarithm of credit by financial intermediaries to the private sector divided by GDP. Other explanatory variables include the logarithm of the initial Gini coefficient (G_{init}), initial human capital stock (H_{init}) measured by the logarithm of secondary school attainment in the initial year, international openness (O) measured by the sum of exports and imports divided by GDP, and inflation (I). See Beck, Demirgüç-Kunt and Levine (2007) for more details on the variable definitions and constructions. We employ the same dataset as Beck, Demirgüç-Kunt and Levine (2007), which covers 78 countries over the period from 1958 to 1997. After deleting missing values, we obtain a sample containing $n = 256$ observations.

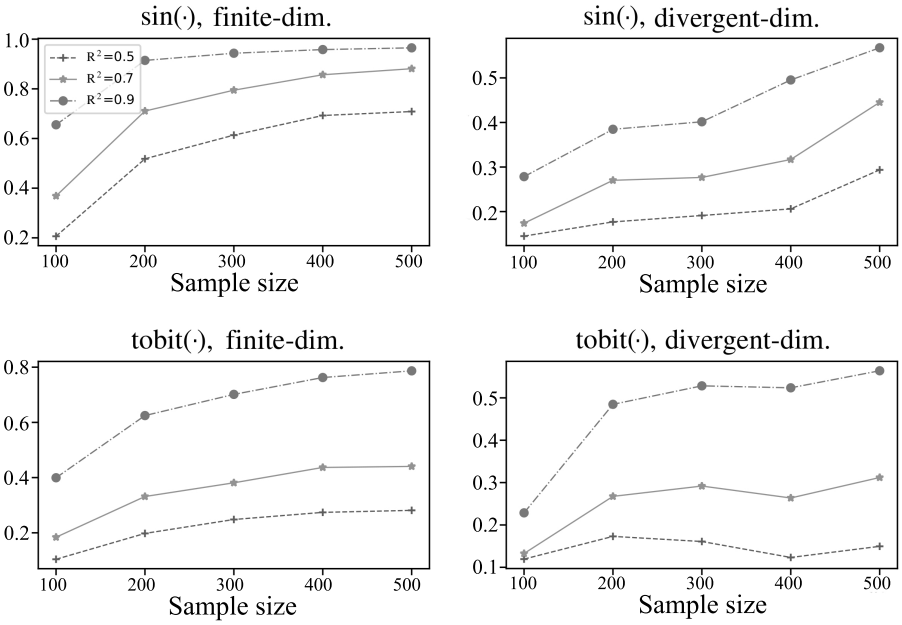


Figure 2. Sum of weights assigned to correct models.

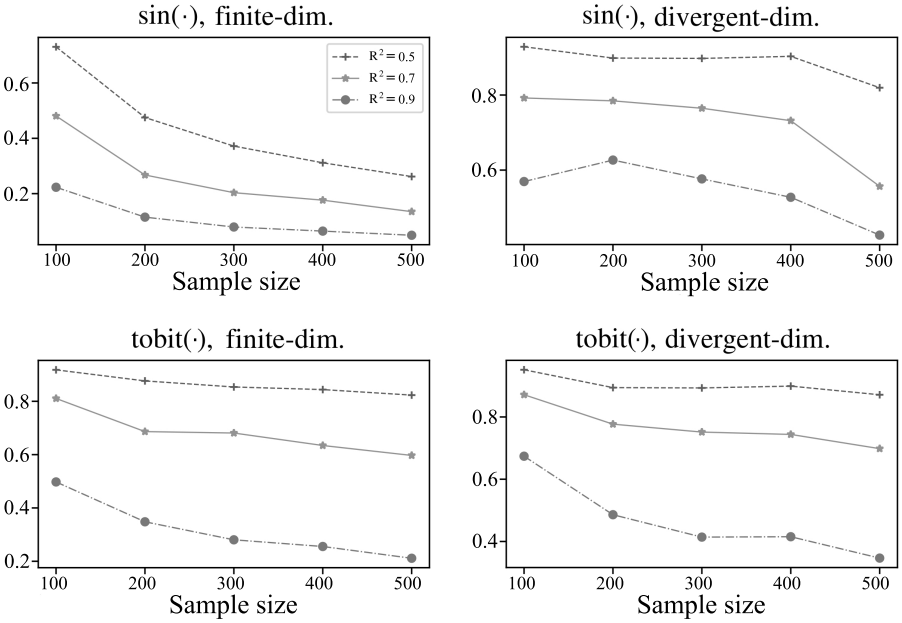
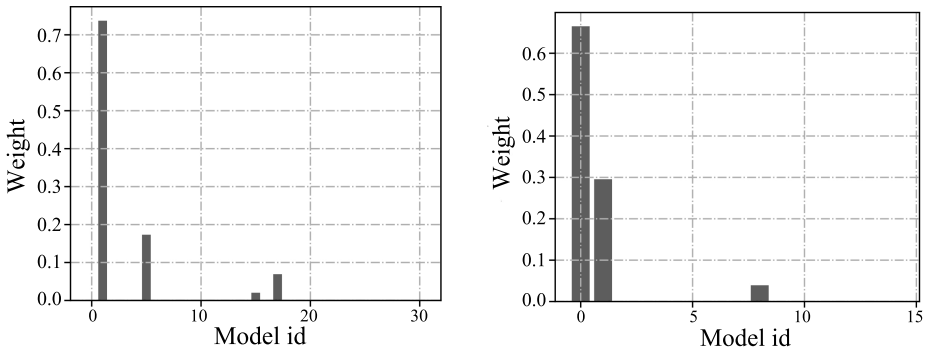


Figure 3. Convergence of $L_n(\widehat{\mathbf{w}})/\inf_{\mathbf{w}\in\mathcal{W}_F} L_n(\mathbf{w})$.



(a) The case of 5 covariates (b) The case of 15 covariates

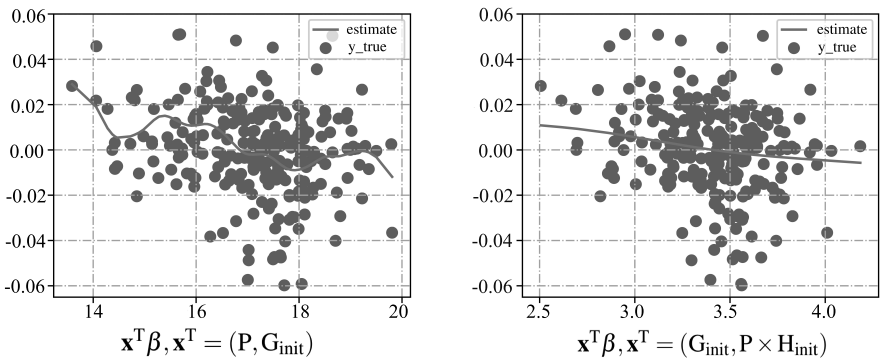
Figure 4. The bar diagrams of model-averaging weights

Economic theory suggests that the impact of financial development on income distribution may be two-fold, because, on the one hand, improvement in the financial system may help reduce inequality by relaxing constraints for the poor who lack collateral and credit histories (Beck, Levine and Levkov, 2010), but on the other hand, the poor mostly rely on informal financial sources and thus may benefit less from such an improvement than the rich. Therefore, the (net) impact of financial development on the income distribution is likely to be nonlinear, as suggested by Greenwood and Jovanovic (1990).

To model the potentially nonlinear relation between financial development and income distribution and account for the model uncertainty, we apply the proposed JCVMA to the SIM with two sets of covariates. The first includes the five covariates $(P, G_{init}, H_{init}, O, I)$ in Beck, Demirgüç-Kunt and Levine (2007), leading to $2^5 - 1 = 31$ candidate models. The second set additionally includes the multiplicative terms of every two covariates to control for potential interaction effects, thus containing 15 regressors in total and leading to $2^{15} - 1 = 32767$ models if one considers all possible combinations of 15 regressors. In the second case, estimating and averaging all possible models is computationally formidable, and thus we employ the ordering-based prescreening discussed in Section 4.

We first examine how the JCVMA weights are distributed across candidate models. Figure 4 presents the histogram of weighs for each model in the two cases with different sets of covariates. Clearly, in both cases the weights concentrate on only a few models. In the case of 5 covariates (left sub-figure), roughly 90% of the weights concentrate on two candidate models, and in the case of 15 covariates (right sub-figure) the two most heavily weighted models account for 97% of the weights. It happens that financial development (P), the variable of interest, is always included in one of the top two models in both cases.

Given such weight distributions, we then investigate the relation between financial development and the income distribution focusing on the most heavily



(a) The case of 5 covariates (b) The case of 15 covariates

Figure 5. Growth of Gini coefficient: True vs. estimated values

weighted model that contains this covariate. Generally, the total effect of a covariate in SIMs should be jointly inferred by the coefficient estimates and the estimated link function. We find that for the model that includes financial development, the estimated coefficient of financial development is significantly positive at the 5% level, where the confidence interval based on JCVMA is obtained by bootstrapping with 500 resamplings. Figure 5 plots the true and predicted values of the Gini coefficient growth against the linear function $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ for the most heavily weighted candidate SIMs, which include financial development. It is revealed that the estimated link function is positive for small values of $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ but negative when $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ is moderate or large. These estimation results jointly imply that for a portion of observations, the effect of financial development on the growth of Gini coefficient is significantly negative, which explains the negative overall effect of OLS as reported by Beck, Demirgüç-Kunt and Levine (2007). However, this effect is significantly positive for observations with relatively small values of $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. The variability of the link function implies that financial development does help alleviate income inequality when the degree of inequality is stable with little inflation, but in some countries, e.g., Korea, Indonesia, and several European countries in the 1960s-1970s with particularly high inflation, financial development further accelerates the growth of inequality. Our results are consistent with the economic theory that financial development exerts two-fold effects depending on the economic and social status (Greenwood and Jovanovic, 1990; Beck, Levine and Levkov, 2010).

Next, we examine the performance of JCVMA in predicting the growth of Gini coefficient. We consider the pseudo out-of-sample prediction over time by dividing the entire time period into a training and a testing subsample. We estimate the parameters, link functions and weights of each candidate model from the training set, and use these estimates to predict the response variable in

Table 1. MSPE of growth of Gini coefficient.

Training sample size	$\approx 0.6n$ 154	$\approx 0.65n$ 165	$\approx 0.7n$ 173	$\approx 0.75n$ 190	$\approx 0.8n$ 197	$\approx 0.85n$ 216
			5 covariates			
JCVMA	0.860	**0.643**	**0.904**	**0.893**	**0.907**	0.901
AIC	0.849	0.646	1.435	1.179	0.924	**0.794**
BIC	**0.840**	0.646	0.911	0.909	0.912	0.995
AICC	**0.840**	0.646	0.911	0.904	0.924	**0.794**
SAIC	0.865	0.644	1.263	1.010	0.918	**0.794**
SBIC	**0.840**	0.646	0.911	0.909	0.921	0.993
SAICC	0.871	0.692	0.951	0.953	0.982	1.066
Full	1.000	1.000	1.000	1.000	1.000	1.000
			15 covariates			
JCVMA	**0.983**	**0.748**	0.892	**0.864**	**0.616**	**0.559**
AIC	1.000	0.772	0.840	1.045	0.888	0.915
BIC	0.997	0.752	**0.763**	0.901	0.644	0.560
AICC	1.676	1.106	0.840	1.745	0.888	0.560
SAIC	1.000	0.772	0.840	1.045	0.888	0.915
SBIC	0.997	0.752	**0.763**	0.901	0.644	0.560
SAICC	1.676	1.106	0.840	1.745	0.781	0.594
Full	1.000	1.000	1.000	1.000	1.000	1.000

Notes: All numbers are normalized by dividing the MSPE of the full SIM, so a value smaller than 1 suggests a better prediction than the full SIM. The upper panel considers 5 covariates, and the bottom panel considers 15 covariates including interaction terms.

the testing set. To reduce randomness, we vary the division point and set it at 1987, 1988, 1989, 1990, 1991 and 1993, so that the first subsample, which is used as the training set, consists of approximately 60%, 65%, 70%, 75%, 80% and 85% of the entire sample, respectively. Accordingly, the second subsample is used as the testing set. This setup also allows us to examine how the competing methods perform across different training sample sizes. We follow Hansen (2008) to evaluate the competing methods according to the mean squared prediction error (MSPE) defined as $\text{MSPE} = n_{\text{test}}^{-1} \left\| \widehat{\mathbf{y}} - \mathbf{y}_{\text{test}} \right\|^2 - \widehat{\sigma}^2$, where $\mathbf{y}_{\text{test}}$ is the response variable of the testing set, $\widehat{\mathbf{y}}$ is its predicted value, and $\widehat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the estimated variance of y_i based on the entire sample with $\bar{y}$ being the sample mean of y_i. The results are presented in Table 1. All numbers are divided by the MSPE of the full SIM, so that a value smaller than 1 suggests a better prediction than the full SIM. Table 1 shows that JCVMA performs the best in 9 of 12 cases. Even when JCVMA does not produce the lowest MSPE, it is close to the best method, suggesting its robustness, while the performances of other methods greatly vary across cases. Moreover, we find that the superiority of JCVMA becomes more obvious when the number of covariates grows from 5 to 15, suggesting that the method may be particularly useful in the presence of a

Table 2. MSPE of corporate sales growth.

Training sample size	$\approx 0.50n$ 575	$\approx 0.64n$ 728	$\approx 0.78n$ 886	$\approx 0.90n$ 1,025
JCVMA	**0.597**	**0.305**	**0.176**	0.057
AIC	1.983	0.915	1.000	0.335
BIC	0.738	0.420	0.204	0.175
AICC	2.929	0.816	1.383	0.335
SAIC	1.975	0.646	1.000	0.287
SBIC	0.738	0.420	0.204	0.175
SAICC	0.646	0.336	0.239	**0.052**
Full	1.000	1.000	1.000	1.000

Notes: All numbers are normalized by dividing the MSPE of the full SIM, so a value smaller than 1 suggests a better prediction than the full SIM.

large degree of uncertainty.

6.2. US firm sales growth

Our second application focuses on predicting the sales growth of US manufacturing firms using a wide range of potential covariates. Our prediction is based on the first-order lagged values of 12 potential covariates, namely, Tobin's q, cash flow, property, plant, and equipment (PPE), logarithm of total assets, level of sales, capital expenditure, leverage, earnings before interest and taxes, total liabilities, price-to-book (P/B) ratio, net income, and Z-score. All variables are collected from Compustat, and we use the sample from 2000 to 2006 to avoid the severe economic crisis broke out since 2007, during which the functional relation among financial variables may change remarkably from that of other years. After removing the missing data and firms with less than 3 observations, we obtain a sample of $n = 1141$ observations.

As in the first application, we employ ordering-based pre-screening to reduce the number of candidate models, and use the same method to choose the optimal bandwidth. To evaluate the pseudo out-of-sample prediction performances of JCVMA and competing methods, we divide the entire time period into two subsamples at 2002, 2003, 2004 and 2005, such that the first subsample used as the training set consists of approximately 50%, 64%, 78% and 90% of the entire observations, respectively. We evaluate the prediction performance using the MSPE, and the results are presented in Table 2 with all numbers normalized by dividing the MSPE of the full SIM as above.

Table 2 shows that JCVMA produces the most accurate prediction in most of the cases, except when the training size is approximately $0.9n$. Particularly, when the training sample is $0.5n$, the MSPE of JCVMA is more than 40% lower than that of the full model and is approximately 8% lower than that of the second-best method, SAICC. When the training size increases to $0.64n$ and further to $0.78n$,

JCVMA improves over the full model even more remarkably, and outperforms the second-best methods SAICC and SBIC by almost 9% and 14%, respectively. Even when JCVMA is not the best method for the training size of $0.9n$, it produces the second-lowest MSPE, which is very close to that of the best method, SAICC. Further examination reveals that JCVMA tends to assign relatively large weights for small models, and this fact partly explains the large discrepancy between JCVMA and the full model and suggests that many covariates may have weak predictability for sales growth.

From the two empirical examples, we can see that JCVMA performs robustly well in prediction for different types of data, while the performance of the competing methods varies remarkably across datasets. By accounting for nonlinearity and model uncertainty, the proposed JCVMA also provides new economic insights.

7. Concluding Remarks

This paper proposes a model-averaging method to address the model uncertainty in single-index models, and our averaging method allows the numbers of covariates and candidate models to diverge when the sample size increases. We also propose model averaging based on regularized estimation and prescreening to deal with many covariates and candidate models. We demonstrate the superior properties of the proposed method when all candidate models are misspecified and when correct models are available in the candidate model set. Some of our theories and techniques, including the weight convergence, the treatment of diverging dimension, and regularization-based averaging, can be applied or extended to other semi-/non-parametric models.

Supplementary Material

The online Supplementary Material file contains some explanations of conditions, technical proofs, detailed simulation studies and another empirical example.

Acknowledgments

Jiahui Zou's work was supported by the National Natural Science Foundation of China (Grant No. 12201431). Xinyu Zhang's work was partially supported by the National Natural Science Foundation of China (Grant Nos. 71925007, 72091212 and 72495124) and the CAS Project for Young Scientists in Basic Research (YSBR-008). Guohua Zou's work was partially supported by the National Natural Science Foundation of China (Grant Nos. 12426308 and 12031016) and Beijing Natural Science Foundation (Grant No. Z210003).

Appendix

This appendix provides additional conditions for Lemma 1 and Corollary 2. Some detailed explanations are provided in the online Supplementary Material.

A.1. Conditions for Lemma 1

The following regularity conditions are required for the consistency of the NLS estimator and its cross-validation version for each candidate model.

Condition A.1. (*i*) *The kernel function* $k(s)$ *is a bounded symmetric density with a compact support;* (*ii*) *The following quantities are finite:* $\int |\tau k'(\tau)|\,\mathrm{d}\tau$, $\int \tau^2|k'(\tau)|\,\mathrm{d}\tau$, $\int k'^2(\tau)\,\mathrm{d}\tau$, $\int |\tau| k'^2(\tau)\,\mathrm{d}\tau$ *and* $\int \tau^2 k'^2(\tau)\,\mathrm{d}\tau$, *where* $k'(s)$ *is the first-order derivative of* $k(s)$.

Condition A.2. (*i*) $\max_{1\le s\le S_n} h_s = o(1)$. (*ii*) $\sum_{s=1}^{S_n} n^{-1}h_s^{-3}p_s = O_P(1)$. (*iii*) $\max_{1\le s\le S_n}(nh_s^4 + h_s^{-1})/M_n^2 n d_n^2 = O(1)$.

Condition A.3. (*i*) *There exists a universal constant* $\bar{C} > 0$ *such that* $\max_{1\le s\le S_n}\max_{1\le i\le n}\|\mathbf{x}_{(s),i}\| \le \sqrt{p_s}\bar{C}$. (*ii*) $\max_{1\le s\le S_n}\max_{1\le i\le n}|\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^*|$ *and* $\max_{1\le s\le S_n}\max_{1\le i\le n}|x_{(s),ir}\beta_r^*|$ *are bounded.* (*iii*) $\max_{1\le s\le S_n}\max_{1\le i\le n}|g_{(s)}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^*)| = O_P(1)$. (*iv*) *There exists a constant* $\underline{c}$ *such that* $\min_{1\le s\le S_n}\min_{r:1\le r\le p_s,\beta_{(s),r}^*\neq 0}|\beta_{(s),r}^*| > \underline{c} > 0$. (*v*) $\max_{1\le s\le S_n}(nS_np_s)^{-1/2}\|\partial\sum_{i=1}^n\{\mu_i - \sum_{j\neq i}^n K_{(s),ij}(\boldsymbol{\beta}_{(s)}^*)\mu_j\}^2/\partial\boldsymbol{\beta}_{(s)}\| = O_P(1)$.

Let $\rho_{(s)}(v_1,\ldots,v_{p_s})$ denote the joint density function of $x_{(s),1}\beta_1^*,\ldots,x_{(s),p_s}\beta_{p_s}^*$ for the s^{th} candidate model, where $\mathbf{x}_{(s)} = (x_{(s),1},\ldots,x_{(s),p_s})^{\mathrm{T}}$ and $\boldsymbol{\beta}_{(s)}^* = (\beta_1^*,\ldots,\beta_{p_s}^*)^{\mathrm{T}}$. Let $f_{(s)}(t)$ denote the density of $\mathbf{x}_{(s)}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^*$, and denote $f'_{(s)}(t)$ and $f''_{(s)}(t)$ as the first and second-order derivatives of $f_{(s)}(t)$, respectively. Further let $\phi_{(s)}(t) = g_{(s)}(t)f_{(s)}(t)$ and $\varphi_{(s)}(t) = g_{(s)}^2(t)f_{(s)}(t)$.

Condition A.4. (*i*) *There exists a constant* $\bar{C}$ *such that*

$$\int \rho_{(s)}\left(v_1,\ldots,v_{k-1},t-\sum_{l\neq k}^{p_s}v_l,v_{k+1},\ldots,v_{p_s}\right)\mathrm{d}v_1\ldots\mathrm{d}v_{k-1}\,\mathrm{d}v_{k+1}\ldots\mathrm{d}v_{p_s} < \bar{C}$$

uniformly for s *and* t. (*ii*) *There exist some constants* $\underline{c}$ *and* $\bar{C}$ *such that* $\underline{c} < f_{(s)}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^*) < \bar{C}$ *almost surely for* $s = 1,\ldots,S_n$; $i = 1,\ldots,n$. (*iii*) *There exists a universal constant* $\bar{C}$ *such that* $|f'_{(s)}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^*)| < \bar{C}$, $|f''_{(s)}(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^*)| < \bar{C}$ *almost surely for* $s = 1,\ldots,S_n$; $i = 1,\ldots,n$. (*iv*) *There exists a constant* $G > 0$ *and* $\omega_{(s)}(v_1,\ldots,v_{k-1},v_{k+1},\ldots,v_{p_s}) > 0$ *such that*

$$\begin{aligned}&\left|\rho_{(s)}(v_1,\ldots,v_{k-1},t_1,v_{k+1},\ldots,v_{p_s}) - \rho_{(s)}(v_1,\ldots,v_{k-1},t_2,v_{k+1},\ldots,v_{p_s})\right|\\ &\le G\omega_{(s)}(v_1,\ldots,v_{k-1},v_{k+1},\ldots,v_{p_s})|t_1-t_2|,\end{aligned}$$

for any s *and* k, *where* $\int\omega_{(s)}(v_1,\ldots,v_{k-1},v_{k+1},\ldots,v_{p_s})\,\mathrm{d}v_1\cdots\mathrm{d}v_{k-1}\,\mathrm{d}v_{k+1}\cdots\mathrm{d}v_{p_s} < \infty$ *and* $\sum_{r=1,r\neq k}^{p_s}\int|v_l|\omega_{(s)}(v_1,\ldots,v_{k-1},v_{k+1},\ldots,v_{p_s})\,\mathrm{d}v_1\cdots\mathrm{d}v_{k-1}\,\mathrm{d}v_{k+1}\cdots\mathrm{d}v_{p_s}$

$< \infty$ *uniformly for any* s. (v) $f'_{(s)}(t)$ *and* $f''_{(s)}(t)$ *satisfy the Lipschitz condition, i.e., there exist two constants* c_1 *and* c_2 *such that* $|f'_{(s)}(t_1) - f'_{(s)}(t_2)| \leq c_1|t_1 - t_2|$ *and* $|f''_{(s)}(t_1) - f''_{(s)}(t_2)| \leq c_2|t_1 - t_2|$; (vi) $\phi'_{(s)}(t)$ *and* $\varphi'_{(s)}(t)$ *satisfy the Lipschitz condition.*

Condition A.5. (i) *For any* $s = 1, \ldots, S_n$, *the objective function* $H_{(s),n}(\boldsymbol{\beta}_{(s)})$ *defined in* (2.1) *is twice continuously differentiable.* (ii) *There exists a constant* $c_0 > 0$ *such that*

$$\min\left[\min_{1\leq s\leq S_n}\lambda_{\min}\left\{\frac{\partial^2 H_{(s),n}(\boldsymbol{\beta}^*_{(s)})}{\partial\boldsymbol{\beta}_{(s)}\partial\boldsymbol{\beta}^{\mathrm{T}}_{(s)}}\right\}, \min_{1\leq s\leq S_n}\min_{1\leq j\leq J_n}\lambda_{\min}\left\{\frac{\partial^2 H^{[-j]}_{(s),n}(\boldsymbol{\beta}^*_{(s)})}{\partial\boldsymbol{\beta}_{(s)}\partial\boldsymbol{\beta}^{\mathrm{T}}_{(s)}}\right\}\right] \geq c_0 > 0.$$

References

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

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

Beck, T., Demirgüç-Kunt, A. and Levine, R. (2007). Finance, inequality and the poor. *Journal of Economic Growth* **12**, 27–49.

Beck, T., Levine, R. and Levkov, A. (2010). Big bad banks? The winners and losers from bank deregulation in the United States. *The Journal of Finance* **65**, 1637–1667.

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

Cheng, L., Zeng, P. and Zhu, Y. (2017). BS-SIM: An effective variable selection method for high-dimensional single index model. *Electronic Journal of Statistics* **11**, 3522–3548.

Claeskens, G., Croux, C. and van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* **62**, 972–979.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **70**, 849–911.

Feng, Y., Liu, Q., Yao, Q. and Zhao, G. (2022). Model averaging for nonlinear regression models. *Journal of Business & Economic Statistics* **40**, 785–798.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Greenwood, J. and Jovanovic, B. (1990). Financial development, growth and the distribution of income. *Journal of Political Economy* **98**, 1076–1107.

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

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

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

Hansen, B. E. (2014b). Nonparametric sieve regression: Least squares, averaging least squares and cross-validation. In *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics* (Edited by J. S. Racine, L. Su, A. Ullah and B. E. Hansen). Oxford University Press.

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

Härdle, W. K., Liang, H. and Gao, J. (2007). *Partially Linear Models*. Springer, Berlin, Heidelberg.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* **14**, 382–417.

Horowitz, J. L. (1998). *Semiparametric Methods in Econometrics*. Springer.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58**, 71–120.

Kong, E. and Xia, Y. (2007). Variable selection for the single-index model. *Biometrika* **94**, 217–229.

Li, C., Li, Q., Racine, J. S. and Zhang, D. (2018). Optimal model averaging of varying coefficient models. *Statistica Sinica* **28**, 2795–2809.

Liang, H., Liu, X., Li, R. and Tsai, C.-L. (2010). Estimation and testing for partially linear single-index models. *The Annals of Statistics* **38**, 3811–3836.

Liang, H., Wang, S. and Carroll, R. J. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika* **94**, 185–198.

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

Liu, C.-A. (2018). Averaging estimators for kernel regressions. *Economics Letters* **171**, 102–105.

Naik, P. A. and Tsai, C.-L. (2001). Single-index model selections. *Biometrika* **88**, 821–832.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403–1430.

Racine, J. S., Li, Q., Yu, D. and Zheng, L. (2023). Optimal model averaging of mixed-data kernel-weighted spline regressions. *Journal of Business & Economic Statistics* **41**, 1251–1261.

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

Zhang, X. and Wang, W. (2019). Optimal model averaging estimation for partially linear models. *Statistica Sinica* **29**, 693–718.

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

Zhang, X., Zou, G., Liang, H. and Carroll, R. J. (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association* **115**, 972–984.

Zhu, R., Wan, A. T. K., Zhang, X. and Zou, G. (2019). A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association* **114**, 882–892.

Zhu, R., Zhang, X., Wan, A. T. K. and Zou, G. (2023). Kernel averaging estimators. *Journal of Business & Economic Statistics* **41**, 157–169.

(Received September 2022; accepted June 2023)