JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2021, VOL. 00, NO. 0, 1–14
https://doi.org/10.1080/07350015.2020.1870477

Model Averaging for Nonlinear Regression Models

Yang Feng[a], Qingfeng Liu[b], Qingsong Yao[c], and Guoqing Zhao[c]

[a]Department of Biostatistics, School of Global Public Health, New York University, New York, NY; [b]Department of Economics, Otaru University of Commerce, Otaru City, Hokkaido, Japan; [c]School of Economics, Renmin University of China, Beijing, China

ABSTRACT

This article considers the problem of model averaging for regression models that can be nonlinear in their parameters and variables. We consider a nonlinear model averaging (NMA) framework and propose a weight-choosing criterion, the nonlinear information criterion (NIC). We show that up to a constant, NIC is an asymptotically unbiased estimator of the risk function under nonlinear settings with some mild assumptions. We also prove the optimality of NIC and show the convergence of the model averaging weights. Monte Carlo experiments reveal that NMA leads to relatively lower risks compared with alternative model selection and model averaging methods in most situations. Finally, we apply the NMA method to predicting the individual wage, where our approach leads to the lowest prediction errors in most cases.

ARTICLE HISTORY
Received April 2020
Accepted December 2020

KEYWORDS
Asymptotic optimality; Model averaging; Nonlinear regression model; Weight-choosing criterion

1. Introduction

Model selection has been an active research topic in statistics and econometrics for over 40 years, and various methods have been proposed based on distinct criteria, including Akaike information criterion (AIC; Akaike 1973), Mallows' C_p (Mallows 1973), Bayesian information criterion (BIC; Schwarz 1978), leave-one-out cross-validation (Stone 1974), risk inflation criterion (RIC; Foster and George 1994), and focused information criterion (FIC; Claeskens and Hjort 2003). Despite the great success of model selection, model averaging may be a better option in many situations (Claeskens and Hjort 2008).

Compared with model selection, model averaging has two crucial advantages. First, it reduces the uncertainty of model selection. Given a data set and a set of candidate models, a model selection method selects a particular model from the candidate set, which is not stable because it is not a continuous process in terms of the data. Indeed, even if we slightly perturb one observation in the dataset, a specific model selection method may give us a different model from the original chosen one. The model averaging method gives us a possible way to reduce the uncertainty of model selection by assigning each model a weight, which is continuous in the data and usually changes a little when a small part of the dataset changes. Second, compared with model selection, model averaging can reduce the risk of model misspecification and estimation. For example, to explain a specific economic phenomenon, there may be many plausible candidate models. In that case, using an averaged model instead of a particular model selected by some model selection method, the risk arising from misspecification can be reduced. Model averaging reduces to a model selection method if we assign weight one to a particular model and zeros to the rest, which implies that a model selection method is a special case of model averaging. From this perspective, the model averaging estimator has the potential to achieve a better estimate than a specific post-model-selection estimator. Hansen (2014) showed that the limiting risk of the least-squares estimator of the full linear model is larger than that of the Mallows model averaging (MMA) estimator, proposed by Hansen (2007); Zhang, Ullah, and Zhao (2016) further showed that under some conditions, the risk of the full model is strictly larger than that of the MMA estimator in finite sample cases.

In the existing literature, the Bayesian model averaging (BMA) and the frequentist model averaging (FMA) are two different approaches to model averaging. BMA has been studied by many researchers for decades (see reviews, e.g., Draper 1995; Hoeting et al. 1999; Clyde and George 2004). The idea of FMA dates back to as early as Bates and Granger (1969), which focuses on forecast combination. Under a nonlinear model averaging (NMA) framework, this article proposes a nonlinear information criterion (NIC) for choosing the weights, which is inspired by the MMA in Hansen (2007). Hansen (2007) showed that MMA is asymptotically optimal in the sense that the unknown loss function is asymptotically minimized. After that, to allow for heteroscedastic models, Hansen and Racine (2012) proposed the jackknife model averaging (JMA), and Liu and Okui (2013) adjusted the penalty term of the Mallows criterion based on White's estimator (White 1980) of the covariance matrix to accommodate the heteroscedasticity. Cheng, Ing, and Yu (2015), Dardanoni et al. (2015), Gao et al. (2016), and Liu, Okui, and Yoshimura (2016) proposed robust model averaging procedures when the model has general error terms with heteroscedasticity and autocorrelation.

CONTACT Qingsong Yao greatpine@ruc.edu.cn School of Economics, Renmin University of China, Beijing 100872, China.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.

Other work on FMA includes instrumental variable estimation (Kuersteiner and Okui 2010), high-dimensional regression models (Ando and Li 2014), factor-augmented regression models (Cheng and Hansen 2015), quantile regression models (Lu and Su 2015), generalized linear models and/or generalized linear mixed-effects models (Zhang et al. 2016; De Luca, Magnus, and Peracchi 2018), high-dimensional generalized linear models (Ando and Li 2017), semiparametric ultra-high-dimensional models (Chen et al. 2018), varying-coefficient partially linear models (Zhu et al. 2019), and GARCH models (Liu, Yao, and Zhao 2020). For more details, see reviews on model averaging methods in Claeskens and Hjort (2008) and Moral-Benito (2015).

The method proposed in this article is based on the estimation results of a set of nonlinear regression models. Nonlinear regression models are important tools for economic applications because there are many economic models that take nonlinear forms, for example, the Cobb–Douglas (CD) and the constant elasticity of substitution (CES) production functions, the translog utility function, and the Box–Cox transformation. As discussed in Bates and Watts (1988), compared with linear regression models, the main advantages of nonlinear models are parsimony and, sometimes, better interpretability. Theoretically, although most nonlinear models can be approximated with some degree of accuracy by a linear combination of a set of basis functions locally or globally, the number of basis functions needed may be very large compared with the limited sample size, causing the problem of the "curse of dimensionality." Moreover, whereas a linear framework makes statistical inference convenient, such simplification puts aside the prior or structural information delivered by the domain experts and theories, leading to an unnecessarily large number of basis functions that are hard to interpret. In comparison, a nonlinear model based on specific economic theory can provide a straightforward interpretation with economic implications.

In this article, we consider a NMA framework for regression models that may be nonlinear in both variables and parameters. Also, we propose the model averaging estimator of the unknown conditional mean and construct the corresponding weight-choosing criterion, the NIC. The main theoretical contributions of this article are 3-fold. First, we show that the NIC is a high-order asymptotically unbiased estimator of the unknown risk function. Second, we prove that the NIC asymptotically minimizes the unknown loss function when the number of candidate models is fixed or moderately increasing with sample size. Third, we show that the model averaging weights selected by minimizing the NIC converge to the optimal weights that minimize the unknown mean squared error.

We now discuss several related works. Zhang and Liang (2011) developed an FMA method for partially linear models. In their analysis, all the models in the candidate model set are limited to generalized additive partial linear models. Sueishi (2013) developed model selection and averaging methods for moment restriction models with a focused information criterion based on the generalized empirical likelihood estimator. That work focused on a local misspecification framework, which is different from our setting. Zhang, Zou, and Carroll (2015) proposed a model averaging method based on the Kullback–Leibler divergence for models with homoscedastic normally distributed error terms. Their method can also be applied to models that are nonlinear with respect to parameters and variables. However, they did not provide any theoretical results for nonlinear models. Zhang et al. (2016) studied the model averaging method for generalized linear models and generalized linear mixed-effects models. After transformation with a common link function, the expected responses of the candidate models are linear in canonical parameters. In contrast, the expected response of the candidate models in NMA are nonlinear in parameters, and may not be written in a linear form even after any transformation of the response.

We now describe a motivating example for the NMA framework. The empirical application in Section 5 of this article focuses on predicting the individual wage, using explanatory variables including education, experience, tenure, etc. Studying the nonlinear impacts of continuous variables including education and experience on individual wage is very important. For example, under preference heterogeneity, the log-wage can be concave in years of education (Lemieux 2006), indicating diminishing marginal returns of education; Mincer (1996) also showed that, under the heterogeneity of individual preference and earning opportunity, the average log-wage could either be concave or convex in years of education. These observations imply that, instead of being constant, education's marginal returns can fall or rise as the years of education increase. Some empirical researches also support the nonlinearity of the impacts of education (e.g., Heckman, Lochner, and Todd 2008). To deal with the nonlinear impacts of the explanatory factors, many researchers add quadratic terms of explanatory factors into the basic Mincer equation. However, such practice sometimes lacks flexibility (see, e.g., Murphy and Welch 1990). Quadratic specification implies that the marginal impact increases or decreases linearly with the change of the explanatory factors, which may be too restrictive in empirical applications. The same issue also applies to higher-order polynomial approximations. Moreover, using higher-order polynomials implies that as the explanatory variable tends to infinity, the marginal return will tend to positive or negative infinity, which may be unrealistic. To solve this issue, we introduce nonlinear factors to form our candidate models. The nonlinear form we consider is a power function of the corresponding predictor, with the exponent being a parameter to be estimated. This nonlinear form is more flexible and parsimonious than the Mincer equation model with some quadratic terms or models of a high-order polynomial. Moreover, as there is no consensual model from economic theory for predicting individual wages, the risk of misspecification of any model is high. Thus, we adopt the NMA framework to improve prediction performance. The empirical results show that the NMA outperforms model selection methods and other model averaging approaches.

The remainder of the article is organized as follows. In Section 2, we introduce the NMA framework for nonlinear regression models and construct the weight-choosing criterion NIC. Section 3 mainly focuses on the theoretical properties of NIC. In Section 4, Monte Carlo experiments are conducted to illustrate the finite sample properties. In Section 5, as an empirical application of our method, we apply the NMA with the NIC to predict individual wages. We provide concluding

remarks in Section 6. All technical details are relegated to the Supplementary Materials.

2. Nonlinear Model Averaging

2.1. The Setup

Suppose we observe n independent and identically distributed (iid) pairs $\{(X_1, y_1), \ldots, (X_n, y_n)\}$ from (X, y), where the data generation process of (X, y) is described as follows.

Assumption 1. The data-generation process is

$$y = \mu(X) + \varepsilon, \tag{1}$$

where the random vector $X = (x_1, x_2, \ldots)^T$ is countably infinite and distributed on the Euclidean space, Ω. μ is an $\Omega \to \mathbb{R}$ measurable function. The error ε satisfies $E[\varepsilon|X] = 0$ and $\text{var}[\varepsilon|X] = \sigma^2$.

Assumption 1 is standard and has been used in many existing works related to nonlinear regression models such as Jennrich (1969) and White (1981). Here, the goal is usually to estimate the unknown conditional mean function $\mu(\cdot)$. To estimate μ, suppose a total of S candidate models are given as $\mathcal{M} = \{M_1, \ldots, M_S\}$. For each $s = 1, \ldots, S$, we assume the following working parametric model:

$$y = f_s(X, \boldsymbol{\theta}_s) + \epsilon_s, \tag{2}$$

where the format of $f_s(\cdot, \cdot)$ is given, $\boldsymbol{\theta}_s = (\theta_{s,1}, \theta_{s,2}, \ldots, \theta_{s,K_s})^T$ is the unknown K_s-dimensional parameter vector, and ϵ_s is the random error. We impose the following assumption on the sth model $f_s(X, \boldsymbol{\theta}_s)$.

Assumption 2. $\boldsymbol{\theta}_s \in \Theta_s$, where Θ_s is a compact and convex parameter space. $f_s(X, \boldsymbol{\theta}_s)$ is an $\Omega \times \Theta_s \to \mathbb{R}$ function that is measurable for every $\boldsymbol{\theta}_s \in \Theta_s$ and twice differentiable for every $X \in \Omega$. Moreover, there exists $\boldsymbol{\theta}_{s,0}$ that uniquely minimizes $E\left[\mu(X) - f_s(X, \boldsymbol{\theta}_s)\right]^2$ on Θ_s.

Assumption 2 is imposed to ensure the parameter vector is identifiable for each candidate model. Moreover, although the function $f_s(X, \boldsymbol{\theta}_s)$ is a function of X, it can only depend on a finite number of elements of X. This is natural since, in the real world, the response may be influenced by uncountable factors; however, we usually choose to focus on a few specific ones implied by the candidate model. For example, when the sth model is specified as $f_s(X, \boldsymbol{\theta}_s) = \sum_{k=1}^{K_s} \theta_k x_{s_k}$, $\{s_1, \ldots, s_{K_s}\} \subseteq \mathbb{N}$, it is, in fact, a linear regression model containing a subset of all variables.

With sample $\left\{(X_t, y_t)\right\}_{t=1}^n$, the *nonlinear least-squares estimator* of $\boldsymbol{\theta}_s$ is defined as

$$\widehat{\boldsymbol{\theta}}_{s,n} = \arg\min_{\boldsymbol{\theta}_s \in \Theta_s} n^{-1} \sum_{t=1}^{n} \left[y_t - f_s(X_t, \boldsymbol{\theta}_s)\right]^2, \tag{3}$$

whose existence are guaranteed by Assumptions 1 and 2 combined with Lemma 2 in Jennrich (1969). Then, the approximation of $\mu(X)$ under the sth model is given by $f_s(X, \widehat{\boldsymbol{\theta}}_{s,n})$.

2.2. The NMA Estimator

Given the S candidate models, the idea of model selection is to select a single "best" model according to some criteria. The past two decades have witnessed an explosion of the model selection literature with many significant advances in methodology, theory, as well as algorithms. However, it is well documented in the simulation and empirical results in various works (e.g., Hansen 2007; Liu, Okui, and Yoshimura 2016; Zhang et al. 2016, among others) that when the sample size is relatively small and the signal-to-noise ratio is low, model-averaging methods could work better than model selection in terms of lower prediction error. In addition, model selection could be unstable in the sense that sometimes a small perturbation of the data could lead to a totally different selection result (Yang 2001; Zhang and Liang 2011). Consequently, selecting a single model and discarding others will inevitably increase the instability of the estimation and approximation.

To address the above issues, we consider the *nonlinear model averaging* (NMA) framework, with the following model averaging estimator for $\mu(X)$:

$$\widehat{\mu}(X, W) = \sum_{s=1}^{S} w_s f_s(X, \widehat{\boldsymbol{\theta}}_{s,n}), \tag{4}$$

where the NMA weight vector $W = (w_1, \ldots, w_S)^T \in \mathscr{H}_S = \{W | W \in [0,1]^S, \sum_{s=1}^S w_s = 1\}$ and $f_s(X, \widehat{\boldsymbol{\theta}}_{s,n})$ represents the estimate of $\mu(X)$ from model s. Note that the differences between any two candidate nonlinear models lie, not only in the included explanatory variables, but also in the functional forms. Each candidate model may characterize only some of the properties of the true data generating process. Consequently, a properly weighted average of all the candidate models has the potential to increase accuracy.

Remark 1. The NMA estimator given in (4) can be viewed as an extension of the linear model averaging studied in Hansen (2007). To see this, suppose the sth model is given by $f_s(X, \boldsymbol{\theta}_s) = \sum_{k=1}^{K_s} \theta_k x_k$ and the candidate models are nested ($K_1 < K_2 < \cdots < K_S$), then the NMA estimator is given by

$$\widehat{\mu}(X, W) = (x_1, \ldots, x_{K_S}) \sum_{s=1}^{S} w_s \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{s,n} \\ \mathbf{0}_s \end{pmatrix},$$

where $\mathbf{0}_s$ is a $(K_S - K_s)$-dimensional vector with all elements zero; this coincides with the MMA estimator in Hansen (2007).

2.3. Weight-Choosing Criterion

As is true for most model averaging methods, the key question here is how to choose the weights for each candidate model. Here, we introduce the NIC for the NMA as specified in (4). Define the *loss function* and the *risk function* under the NMA framework as $L_n(W) = n^{-1} \sum_{t=1}^n (\mu(X_t) - \widehat{\mu}(X_t, W))^2$ and $R_n(W) = E(L_n(W)|\mathbf{X}_n)$, respectively, where $\mathbf{X}_n = (X_1, \ldots, X_n)$. Our basic aim is to find a weight to minimize the risk $R_n(W)$ given the covariates $\mathbf{X}_n$. Since the risk function is unobservable, we propose to substitute it with an asymptotically

unbiased estimator. To construct such an estimator, we first define

$$\beta_{s,t}(\boldsymbol{\theta}_s) = \partial f_s(X_t, \boldsymbol{\theta}_s)/\partial \boldsymbol{\theta}_s,$$
$$\gamma_{s,t}(\boldsymbol{\theta}_s) = \partial^2 f_s(X_t, \boldsymbol{\theta}_s)/\partial \boldsymbol{\theta}_s \partial \boldsymbol{\theta}_s^T, \quad (5)$$

$$\begin{aligned}\Lambda_{s,n} &= \frac{1}{2}\partial^2 \sum_{t=1}^{n} \left(y_t - f_s\left(X_t, \widehat{\boldsymbol{\theta}}_{s,n}\right)\right)^2 / \partial \boldsymbol{\theta}_s \partial \boldsymbol{\theta}_s^T \\ &= \sum_{t=1}^{n} \left[\beta_{s,t}\left(\widehat{\boldsymbol{\theta}}_{s,n}\right)\beta_{s,t}^T\left(\widehat{\boldsymbol{\theta}}_{s,n}\right) - \left(y_t - f_s(X_t, \widehat{\boldsymbol{\theta}}_{s,n})\right)\gamma_{s,t}\left(\widehat{\boldsymbol{\theta}}_{s,n}\right)\right], \end{aligned} \quad (6)$$

and

$$\pi_{s,n} = n^{-1}\sum_{t=1}^{n} \beta_{s,t}^T\left(\widehat{\boldsymbol{\theta}}_{s,n}\right)\Lambda_{s,n}^{+}\beta_{s,t}\left(\widehat{\boldsymbol{\theta}}_{s,n}\right), \quad (7)$$

where $\Lambda_{s,n}^{+}$ is the generalized inverse of $\Lambda_{s,n}$. Note that when $\widehat{\boldsymbol{\theta}}_{s,n}$ is an inner point of Θ_s and $n^{-1}\sum_{t=1}^{n}\left(y_t - f_s(X_t, \boldsymbol{\theta}_s)\right)^2$ is uniquely minimized at $\widehat{\boldsymbol{\theta}}_{s,n}$ on Θ_s, $\Lambda_{s,n}$ is positive definite and nonsingular, and thus $\Lambda_{s,n}^{+} = \Lambda_{s,n}^{-1}$. This applies to a wide range of model families. For example, when the sth model is a linear regression model with covariate matrix, $\mathbf{X}_n$, where no perfect multicollinearity exists, we have $\gamma_{s,t}\left(\widehat{\boldsymbol{\theta}}_{s,n}\right) = \mathbf{0}$ and $\Lambda_{s,n} = \mathbf{X}_n^T\mathbf{X}_n$ is positive definite.

Given the above notations, the NIC for NMA is defined as

$$\mathrm{NIC}(W) = n^{-1}\sum_{t=1}^{n}\left[y_t - \widehat{\mu}(X_t, W)\right]^2 + 2\widehat{\sigma}_n^2\sum_{s=1}^{S} w_s\pi_{s,n}, \quad (8)$$

where $\widehat{\sigma}_n^2$ is an estimate of σ^2. In the NIC, the first term $n^{-1}\sum_{t=1}^{n}\left(y_t - \widehat{\mu}(X_t, W)\right)^2$ (the sum of the squared residuals) represents the goodness of fit, whereas the second term is a bias-adjusting term. When $\Lambda_{s,n}$ is positive definite, we have $\pi_{s,n} \geq 0$ for all $1 \leq s \leq S$, where the equality holds if and only if $\beta_{s,t}\left(\widehat{\boldsymbol{\theta}}_{s,n}\right) = \mathbf{0}$ for all $1 \leq t \leq n$. On this condition, the bias-adjusting term in the NIC is positive and can be interpreted as a penalty.

Remark 2. When the model averaging weight set is restricted to $\widetilde{\mathscr{H}}_S = \{e_1, \ldots, e_S\}$, where e_s is a unit vector whose sth element is 1 and other elements are all 0, then choosing a weight from $\widetilde{\mathscr{H}}_S$ is equivalent to selecting a model from the candidate set $\left\{f_s(X, \boldsymbol{\theta}_s)\right\}_{s=1}^{S}$. Then, the NIC reduces to

$$\mathrm{NIC}(e_s) = n^{-1}\sum_{t=1}^{n}\left[y_t - f_s(X_t, \widehat{\boldsymbol{\theta}}_{s,n})\right]^2 + 2\widehat{\sigma}_n^2\pi_{s,n}.$$

It is clear that, in this special scenario, $\mathrm{NIC}(e_s)$ is a model selection criterion, and can be regarded as a generalization of the Mallows' C_p (Mallows 1973) to nonlinear regression models. Indeed, when $f_s(X, \boldsymbol{\theta}_s) = \sum_{k=1}^{K_s}\theta_k x_{s_k}$, we have $\beta_{s,t}(\widehat{\boldsymbol{\theta}}_{s,n}) = \left(x_{s_1,t}, \ldots, x_{s_{K_s},t}\right)^T$ and $\gamma_{s,t}(\widehat{\boldsymbol{\theta}}_{s,n}) = \mathbf{0}$. This leads to

$$\begin{aligned}\pi_{s,n} &= n^{-1}\sum_{t=1}^{n}\beta_{s,t}(\widehat{\boldsymbol{\theta}}_{s,n})\left(\sum_{r=1}^{n}\beta_{s,t}^T(\widehat{\boldsymbol{\theta}}_{s,n})\beta_{s,t}(\widehat{\boldsymbol{\theta}}_{s,n})\right)^{-1}\beta_{s,t}^T(\widehat{\boldsymbol{\theta}}_{s,n}) \\ &= n^{-1}K_s.\end{aligned}$$

This shows that $\mathrm{NIC}(e_s)$ is equivalent to Mallows' C_p under linear regression settings. For nonlinear regression models, however, $\mathrm{NIC}(e_s)$ usually differs from Mallows' C_p by imposing a penalty to models not only on the number of parameters but also on the complexity of the functional form of the models.

3. Properties of NIC

In this section, we provide the statistical properties of the NMA with the weight-choosing criterion NIC. In particular, we first show that the NIC is an asymptotically unbiased estimator of the risk function, up to some constant. Then, we show that the NIC is asymptotically optimal in the cases of both fixed and diverging numbers of candidate models, where the unknown loss function is asymptotically minimized. Finally, we illustrate the properties of model averaging weights.

We first consider the case where σ^2 is known. More specifically, consider the following infeasible criterion:

$$\mathrm{NIC}^*(W) = n^{-1}\sum_{t=1}^{n}\left[y_t - \widehat{\mu}(X_t, W)\right]^2 + 2\sigma^2\sum_{s=1}^{S} w_s\pi_{s,n}. \quad (9)$$

We study the asymptotic properties of NIC*. First, we introduce three additional assumptions.

Assumption 3. Define $\Phi_{s,n} = \sum_{t=1}^{n}\left[\beta_{s,t}\beta_{s,t}^T - \left(\mu(X_t) - f_s(X_t, \boldsymbol{\theta}_{s,0})\right)\gamma_{s,t}\right]$ and $\Phi_s = En^{-1}\Phi_{s,n}$, where $\beta_{s,t} = \beta_{s,t}\left(\boldsymbol{\theta}_{s,0}\right)$ and $\gamma_{s,t} = \gamma_{s,t}\left(\boldsymbol{\theta}_{s,0}\right)$, Φ_s is nonsingular for all $1 \leq s \leq S$.

Assumption 4. (1) $E\mu^2(X) < \infty$; (2) for all $1 \leq s \leq S$ and $1 \leq p, q, r \leq K_s$, we have $\partial^3 f_s(X, \boldsymbol{\theta}_s)/\partial\theta_{s,p}\partial\theta_{s,q}\partial\theta_{s,r}$ exists for all $(X, \boldsymbol{\theta}_s) \in \Omega \times \Theta_s$; moreover, for all $0 \leq i + j + k \leq 3$,

$$\left|\partial^{i+j+k} f_s(X, \boldsymbol{\theta}_s)/\partial\theta_{s,p}^i\partial\theta_{s,q}^j\partial\theta_{s,r}^k\right| \leq m(X),$$

where $m(X)$ is measurable and $Em^2(X) < \infty$.

Assumption 5. $\boldsymbol{\theta}_{s,0}$ is an interior point of Θ_s for all $1 \leq s \leq S$.

Assumptions 3–5 are standard in the existing research on nonlinear least-squares estimation (Jennrich 1969; White 1981). Assumption 3 requires that Φ_s is nonsingular for all $1 \leq s \leq S$. Such a restriction is mainly used to obtain the asymptotic normality of $\widehat{\boldsymbol{\theta}}_{s,n}$. Assumption 4 poses some moment conditions on the true conditional mean and the candidate models. It requires that the true conditional mean is square-integrable. Moreover, it requires that each candidate model has a third derivative, and both the model form and its derivatives are bounded by a squared-integrable function independent of $\boldsymbol{\theta}_s$. Note that Assumption 4 is sufficient to obtain the conditions required in Jennrich (1969) and White (1981). In Assumption 5, we assume that the unknown $\boldsymbol{\theta}_{s,0}$ is interior to Θ_s, which is also used to guarantee the asymptotic normality of the nonlinear least-squares estimator.

Supposing all the assumptions hold, we have the following theorem.

Theorem 1. If Assumptions 1–5 hold, then

$$E\left(\mathrm{NIC}^*(W) - \alpha_n(W)\middle|\mathbf{X}_n\right) = R_n(W) + \sigma^2, \quad (10)$$

where $\alpha_n(W) = \sum_{s=1}^{S} w_s \alpha_n(s)$ and $\alpha_n(s) = O_p(n^{-\frac{3}{2}})$; moreover, when the sth model is a linear regression model, then $\alpha_n(s) = 0$.

In Theorem 1, taking $W = e_s$, $\alpha_n(s)$ denotes a higher-order bias term that makes the relationship $E(\mathrm{NIC}^*(e_s) - \alpha_n(s)|\mathbf{X}_n) = R_n(e_s) + \sigma^2$ hold. As shown in (10), when S is fixed the NIC* is an asymptotically unbiased estimator of the risk function $R_n(W)$ with a bias of order at most $O_p(n^{-\frac{3}{2}})$, which vanishes when n goes to infinity, and an extra constant term, σ^2, independent of W. When S increases with the sample size n, we can guarantee that $\alpha_n(W) = o_p(1)$ if $\sup_{1\le s\le S}|\alpha_n(s)| = o_p(1)$. When all the candidate models are linear regression models, we have $\alpha_n(W) = 0$, so the adjusting term can be omitted and the NIC* shares the same property as the $C_n(W)$ in Hansen (2007) since $E(\mathrm{NIC}^*(W)|\mathbf{X}_n) = R_n(W) + \sigma^2$.

Now, we are ready to study the asymptotic property of the NIC, where the error variance is estimated. Parallel to Theorem 1, we have the following theorem for the NIC.

Theorem 2. If Assumptions 1–5 hold, denote $\mathfrak{B}_s = E\left(\beta_{s,t}\beta_{s,t}^T\right)$, then

$$E(\mathrm{NIC}(W) - \widetilde{\alpha}_n(W)|\mathbf{X}_n) = R_n(W) + 2n^{-1}A \cdot \mathrm{tr}\left(\sum_{s=1}^{S} w_s \mathfrak{B}_s \Phi_s^{-1}\right) + \sigma^2, \quad (11)$$

where $\widetilde{\alpha}_n(W) = \sum_{s=1}^{S} w_s \widetilde{\alpha}_n(s)$, $\widetilde{\alpha}_n(s) = o_p(n^{-1})$, and $A = p\lim_{n\to\infty}(\widehat{\sigma}_n^2 - \sigma^2)$.

Theorem 2 indicates that when σ^2 is unknown and is substituted with its estimator, additional bias due to the estimation inaccuracy of σ^2 is introduced. Given the candidate models and the model averaging weight W, the bias increases as the estimate of σ^2 becomes less accurate. This implies that we should be cautious when we choose the $\widehat{\sigma}_n^2$ to make the bias as small as possible.

For the choice of $\widehat{\sigma}_n^2$, we can use the estimated variance based on the sth model, which is given by $\widehat{\sigma}_n^2(s) = n^{-1}\sum_{t=1}^{n}(y_t - f_s(X_t, \widehat{\boldsymbol{\theta}}_{s,n}))^2$. When the sth model is correctly specified, that is, $\mu(X) = f_s(X, \boldsymbol{\theta}_{s,0})$ a.s. holds, Jennrich (1969) showed that $\widehat{\sigma}_n^2(s) \to_p \sigma^2$ and consequently $A = 0$, the NIC is again an asymptotically unbiased estimator of the risk function. However, as we show in the following proposition, when μ is unknown and the sth model is misspecified, $\widehat{\sigma}_n^2(s)$ leads to an upward estimation bias. Define $R_0(s) = E\left(\mu(X) - f_s(X, \boldsymbol{\theta}_{s,0})\right)^2$, which is $O(1)$ under Assumption 2. We have the following proposition.

Proposition 1. If Assumptions 1–5 hold, then $\widehat{\sigma}_n^2(s) \to_p \sigma^2 + R_0(s)$.

When $\mu(X) = f_s(X, \boldsymbol{\theta}_{s,0})$ a.s. holds, we have $R_0(s) = 0$ and $\widehat{\sigma}_n^2(s) \to_p \sigma^2$. However, due to the possibility of the model being misspecified as well as the omission of variables, $R_0(s) > 0$ can hold for all $1 \le s \le S$ and we are not able to obtain an unbiased estimator of σ^2 given the candidate model set. On this condition, to make the bias term as small as possible, we can use $\min_{1\le s\le S}\widehat{\sigma}_n^2(s)$ as the estimated variance. Note that when the largest model nests all the remaining candidates, such a practice is equal to using the estimated variance from the largest model, which is advocated in Mallows (1973).

As an alternative, we can also use the model averaging method to construct the estimator of σ^2. Define $\widehat{\sigma}_n^2(W) = n^{-1}\sum_{t=1}^{n}(y_t - \sum_{s=1}^{S} w_s f_s(X_t, \widehat{\boldsymbol{\theta}}_{s,n}))^2$, then $\widehat{\sigma}_n^2(W)$ can be used as the variance estimator. Similarly to Proposition 1, $\widehat{\sigma}_n^2(W)$ provides an upward-biased estimator for the true σ^2 if the candidate model set is misspecified, that is, $\inf_{W\in\mathscr{H}_S} R_0(W) > 0$, where $R_0(W) = E(\mu(X) - \sum_{s=1}^{S} w_s f_s(X, \boldsymbol{\theta}_{s,0}))^2$. [1]

Remark 3. For fixed S, $\widehat{\sigma}_n^2(W)$ is a consistent estimator of σ^2 if $\mu(X) = \sum_{s=1}^{S} w_s f_s(X, \boldsymbol{\theta}_{s,0})$ holds almost surely. However, this may be impossible when all the candidate models considered are misspecified, implying that $\widehat{\sigma}_n^2(W)$ could be biased. One way to fix this problem is to allow the number of candidate models, S, to increase with sample size n, so that the largest model tends closer to the true conditional mean. For example, when both $n^{-1}\sum_{i=1}^{n}\left(\mu(X_t) - f_S\left(X_t, \boldsymbol{\theta}_{S,0}\right)\right)^2 \to_p 0$ and $n^{-1}\sum_{i=1}^{n}\left(f_S\left(X_t, \widehat{\boldsymbol{\theta}}_{S,n}\right) - f_S\left(X_t, \boldsymbol{\theta}_{S,0}\right)\right)^2 \to_p 0$ hold, we can show that $\widehat{\sigma}_n^2(S) = n^{-1}\sum_{t=1}^{n}\varepsilon_t^2 + o_p(1)$. This implies that $\widehat{\sigma}_n^2(S) - \sigma^2 = \widehat{\sigma}_n^2(S) - n^{-1}\sum_{t=1}^{n}\varepsilon_t^2 + n^{-1}\sum_{t=1}^{n}\varepsilon_t^2 - \sigma^2 = o_p(1)$. In Theorem 6, we will also show that as long as $\left|\widehat{\sigma}_n^2 - \sigma^2\right| = O_p(1)$, $\mathrm{NIC}(W)$ will perform as well as its infeasible counterpart $\mathrm{NIC}^*(W)$ when n is sufficiently large.

Based on the above properties of $\mathrm{NIC}^*(W)$ and $\mathrm{NIC}(W)$, we now study the risk properties of the NIC. Similar to the previous procedure, we first study the loss based on the infeasible weight-choosing criterion $\mathrm{NIC}^*(W)$; then we show that the loss based on $\mathrm{NIC}^*(W)$ and $\mathrm{NIC}(W)$ are asymptotically identical given that the estimated variance $\widehat{\sigma}_n^2$ satisfies some order conditions. In the existing model averaging literature, one desirable property is the asymptotic optimality, which refers to the property that the model averaging weight selected by the weight-choosing criterion asymptotically minimizes the unknown loss function $L_n(W)$. Define $\widehat{W}_n = \arg\min_{W\in\mathscr{H}_S}\mathrm{NIC}^*(W)$ and $\xi = \inf_{W\in\mathscr{H}_S} R_0(W)$, the following theorem shows that the NMA based on the infeasible weight-choosing criterion $\mathrm{NIC}^*(W)$ is asymptotically optimal.

Theorem 3. If Assumptions 1–5 hold, S is fixed, and $\xi > 0$, then

$$\frac{L_n(\widehat{W}_n)}{\inf_{W\in\mathscr{H}_S} L_n(W)} = 1 + O_p(n^{-\frac{1}{2}}). \quad (12)$$

Apart from Assumptions 1–5, Theorem 3 also requires that $\xi > 0$. Such a restriction means that all the candidate models, as well as their weighted combinations, are misspecified, so that the true conditional mean cannot be approximated with arbitrary

[1] In particular, for fixed S, we can easily show that $\widehat{\sigma}_n^2(W) \to_p \sigma^2 + R_0(W)$ based on the techniques used in the proof of Proposition 1.

accuracy by finitely many candidate models. Such a restriction will not be required in the following case with diverging S. Besides optimality, Theorem 3 also provides the convergence speed of the ratio between $L_n(\widehat{W}_n)$ and the unknown optimal loss $\inf_{W\in\mathscr{H}_S} L_n(W)$.

The optimality result in Theorem 3 is derived in the case of finite S, that is, the number of the candidate models does not increase with the sample size n. In many applications, however, we usually have more candidate models in the model set as the sample size increases. So, it is also worthwhile to investigate the properties of model averaging estimation when the number of candidate models increases with the sample size. To ensure that optimality holds when $\lim_{n\to\infty} S = \infty$, we make the following additional assumptions.

Assumption 6.

1. $\sup_{s\geq 1} Ef_s^2(X,\boldsymbol{\theta}_{s,0}) < \infty$;
2. let $\|\cdot\|$ denote the Euclidean norm, $\sup_{1\leq s\leq S} ||\widehat{\boldsymbol{\theta}}_{s,n} - \boldsymbol{\theta}_{s,0}|| = O_p(n^{\alpha_n-\frac{1}{2}})$ for some sequence $\{\alpha_n\}_{n=1}^{\infty}$ which satisfies the condition in Assumption 6(5);
3. for all s, $K_s^{-1}\left\|\partial f_s(X,\boldsymbol{\theta}_s)/\partial\boldsymbol{\theta}_s\right\| \leq M(X)$ for some measurable function $M(X)$ satisfying $EM^2(X) < \infty$;
4. $\sup_{1\leq s\leq S}\left|\pi_{s,n}\right| = O_p(n^{-\frac{1}{2}})$;
5. define $\xi_n = \inf_{W\in\mathscr{H}_S} L_n(W)$, $\sup_{1\leq s\leq S} K_s n^{\alpha_n-\frac{1}{2}}\xi_n^{-1} = o_p(1)$, $S^{\frac{1}{2}}n^{-\frac{1}{2}}\xi_n^{-1} = o_p(1)$.

We briefly discuss Assumption 6. Assumption 6(1) requires that the second moments of all the candidate models are uniformly bounded. Assumption 6(2) makes restrictions on the uniform convergence speed of the estimated parameters. Obviously, when S is fixed, $||\widehat{\boldsymbol{\theta}}_{s,n} - \boldsymbol{\theta}_{s,0}||$ is $O_p(n^{-\frac{1}{2}})$ and so is $\sup_{1\leq s\leq S}||\widehat{\boldsymbol{\theta}}_{s,n} - \boldsymbol{\theta}_{s,0}||$; whereas when S is diverging and $\sup_{s\geq 1} K_s = \infty$, assuming $\sup_{1\leq s\leq S}||\widehat{\boldsymbol{\theta}}_{s,n} - \boldsymbol{\theta}_{s,0}|| = O_p(n^{-\frac{1}{2}})$ may be too strong and can be violated in some situations. As the alternative, a weaker condition is posed, as in Assumption 6(2). Assumption 6(3) requires $||\partial f_s(X,\boldsymbol{\theta}_s)/\partial\boldsymbol{\theta}_s||/K_s$ to be uniformly bounded by a function with a finite second moment. For example, when $f_s(X,\boldsymbol{\theta}_s) = \prod_{k=1}^{K_s} x_k^{\theta_k}$, where $x_k > 0$ for all $1 \leq k \leq K_s$, we have $||\partial f_s(X,\boldsymbol{\theta}_s)/\partial\boldsymbol{\theta}_s|| = \left|f_s(X,\boldsymbol{\theta}_s)\right|(\sum_{k=1}^{K_s}\log^2 x_k)^{\frac{1}{2}}$. If further $E\left(f_s(X,\boldsymbol{\theta}_s)\sup_k\left|\log x_k\right|\right)^2 < \infty$, then the $M(X)$ can be taken as $\left|f_s(X,\boldsymbol{\theta}_s)\right|\sup_k\left|\log x_k\right|$. Assumption 6(4) requires that $\sup_{1\leq s\leq S}\left|\pi_{s,n}\right|$ is $O_p(n^{-\frac{1}{2}})$. Note that when S is fixed, $\sup_{1\leq s\leq S}\left|\pi_{s,n}\right|$ is $O_p(n^{-1})$ under Assumptions 1–5. When S diverges to infinity with n, $\sup_{1\leq s\leq S}\left|\pi_{s,n}\right|$ may be of higher order than $O_p(n^{-1})$, but we still require its decreasing speed is faster than $n^{-\frac{1}{2}}$. Assumption 6(5) makes some restrictions on the number of candidate models S as well as the number of parameters K_s. It is required that both $S^{\frac{1}{2}}$ and $\sup_{1\leq s\leq S} K_s n^{\alpha_n}$ do not increase faster than $n^{\frac{1}{2}}\xi_n$. A necessary condition for the above requirements to hold is that $n^{-\frac{1}{2}}\xi_n^{-1} = o_p(1)$, which indicates that the loss function decreases at a speed slower than $n^{-\frac{1}{2}}$. Similar conditions are also assumed in Zhang et al. (2016).

When Assumptions 1, 2, and 6 hold, we have the following theorem.

Theorem 4. If Assumptions 1, 2, and 6 hold, then

$$\frac{L_n(\widehat{W}_n)}{\inf_{W\in\mathscr{H}_S} L_n(W)} = 1 + O_p\left(\left(\sup_{1\leq s\leq S} K_s n^{\alpha_n} + S^{\frac{1}{2}}\right)n^{-\frac{1}{2}}\xi_n^{-1}\right), \tag{13}$$

holds when $\lim_{n\to\infty} S = \infty$.

Asymptotic optimality describes the properties of the loss function under the model averaging method. Next, we discuss the behavior of the model averaging weight $\widehat{W}_n$ when S is fixed. Since $\mathscr{H}_S$ is compact, it is a direct result that there exists a subsequence $\{\widehat{W}_{n_i}\}_{i=1}^{\infty}$ such that $\widehat{W}_{n_i}$ converges to some well-defined limit in $\mathscr{H}_S$. But such a result is not satisfactory because it does not imply any properties of the limiting weight. Here we provide some further results on the weight vector selected by the NIC*(W). Define $\mathscr{A} = \{W^* \in \mathscr{H}_S | R_0(W^*) \leq R_0(W), \forall W \in \mathscr{H}_S\}$. $\mathscr{A}$ is the set of weights that minimize $R_0(W)$ on $\mathscr{H}_S$, which is nonempty due to the compactness of $\mathscr{H}_S$. We show that the distance between model averaging weight $\widehat{W}_n$ and the set $\mathscr{A}$ will be arbitrarily small as sample size n increases. Specifically, we have the following theorem.

Theorem 5. If Assumptions 1–5 hold and $\xi > 0$, S is fixed, then $\inf_{W\in\mathscr{A}}||\widehat{W}_n - W|| \to_p 0$.

Remark 4. Theorem 5 implies that when $\mathscr{A}$ has a unique element, that is, there is a unique weight vector W^* such that $R_0(W)$ is minimized at W^*, then the model averaging weight vector $\widehat{W}_n$ converges in probability to such an optimal weight vector. Such a result can be used to obtain many useful conclusions. For example, consider the situation where the sth model is $f_s(X,\boldsymbol{\theta}_s) = \theta_0 + \sum_{k=1}^{K_s}\theta_k x_k$ for $1 \leq s \leq S$ and $1 \leq K_1 < K_2 < \cdots < K_S$. Suppose Assumptions 1–5 hold, then $\widehat{\boldsymbol{\theta}}_{s,n} \to_p \boldsymbol{\theta}_{s,0}$ for all s. Suppose further at least one of $\boldsymbol{\theta}_{S,0}(K_{S-1}+1),\cdots,\boldsymbol{\theta}_{S,0}(K_S)$ does not degenerate to zero. Denote $\boldsymbol{\theta}(W) = \sum_{s=1}^{S} w_s(\boldsymbol{\theta}_{s,0}^T, \mathbf{0}_s^T)^T$, where $\mathbf{0}_s$ is a $(K_S - K_s)\times 1$ column vector with all elements being 0, $\boldsymbol{\theta}(W) \neq \boldsymbol{\theta}_{S,0} = \boldsymbol{\theta}(e_S)$ for any $W \in \mathscr{H}_S$ and $W \neq e_S$. Then we have

$$E\left(\mu - \left(1, x_1, \ldots, x_{K_S}\right)\boldsymbol{\theta}_{S,0}\right)^2 < E\left(\mu - \left(1, x_1, \ldots, x_{K_S}\right)\boldsymbol{\theta}(W)\right)^2$$

for all $W \in \mathscr{H}_S$ and $W \neq e_S$. This implies that $W^* = e_S$ uniquely minimizes $R_0(W)$ on $\mathscr{H}_S$ and according to Theorem 5, we have $\widehat{W}_n \to_p e_S$. Consequently, there hold $\widehat{w}_{n,s} \to_p 0$ for $s < S$, $\widehat{w}_{n,S} \to_p 1$, and $\sum_{s=1}^{S}\widehat{w}_{n,s}\widehat{\boldsymbol{\theta}}_{s,n} \to_p \boldsymbol{\theta}_{S,0}$.

Remark 5. Interestingly, the NMA does not always assign weight 1 to the largest model even when the largest model nests all the remaining models if a more general nonlinear model set is considered. We now provide an example. Consider the case where the true conditional mean is $\mu(X) = kx_1x_2$ and there are only two candidate models $f_1(X,\boldsymbol{\theta}_1) = x_1^{\theta_1}$ and $f_2(X,\boldsymbol{\theta}_2) = x_1^{\theta_1}x_2^{\theta_2}$, where $x_1, x_2 \sim U(0,1)$, and x_1 and x_2 are independent of each other. Obviously, $f_1(X,\boldsymbol{\theta}_1)$ is nested by $f_2(X,\boldsymbol{\theta}_2)$. $R_0(W) = E(\mu - w_1f_1(X,\boldsymbol{\theta}_{1,0}) - w_2f_2(X,\boldsymbol{\theta}_{2,0}))^2$, where $\boldsymbol{\theta}_{i,0}$ minimizes $E(\mu - f_i(X,\boldsymbol{\theta}_{i,0}))^2$. Let $\alpha_{11} = E(\mu - f_1(X,\boldsymbol{\theta}_{1,0}))^2$, $\alpha_{22} = E(\mu - f_2(X,\boldsymbol{\theta}_{2,0}))^2$, $\alpha_{12} = E(\mu - f_1(X,\boldsymbol{\theta}_{1,0}))(\mu - f_2(X,\boldsymbol{\theta}_{2,0}))$, and we have $R_0(W) = (\alpha_{11}+\alpha_{22}-2\alpha_{12})w_1^2 - 2(\alpha_{22}-\alpha_{12})w_1 + \alpha_{22}^2$. When $0 < \frac{\alpha_{22}-\alpha_{12}}{\alpha_{11}+\alpha_{22}-2\alpha_{12}} < 1$, $0 < w_1^* < 1$ and thus $0 <$

$w_2^* < 1$ hold, so both models are assigned with nondegenerating weights. Since $\alpha_{11} + \alpha_{22} - 2\alpha_{12} \geq 0$, $0 < \frac{\alpha_{22}-\alpha_{12}}{\alpha_{11}+\alpha_{22}-2\alpha_{12}} < 1$ holds as long as $\alpha_{22} - \alpha_{12} > 0$ holds strictly. Simple calculation leads to $\boldsymbol{\theta}_{1,0} = \frac{2\sqrt{2}k^{-\frac{1}{2}}-1}{2-\sqrt{2}k^{-\frac{1}{2}}}$, $\boldsymbol{\theta}_{2,0} = (\frac{2k^{-\frac{1}{3}}-1}{2-k^{-\frac{1}{3}}}, \frac{2k^{-\frac{1}{3}}-1}{2-k^{-\frac{1}{3}}})^T$. On this condition, let $\boldsymbol{\theta}_{i,0,j}$ be the jth element of $\boldsymbol{\theta}_{i,0}$, and we have

$$\begin{aligned}
\alpha_{22} - \alpha_{12} &= E\left(x_1^{\boldsymbol{\theta}_{1,0,1}} - x_1^{\boldsymbol{\theta}_{2,0,1}} x_2^{\boldsymbol{\theta}_{2,0,2}}\right)\left(kx_1x_2 - x_1^{\boldsymbol{\theta}_{2,0,1}} x_2^{\boldsymbol{\theta}_{2,0,2}}\right) \\
&= kEx_1^{1+\boldsymbol{\theta}_{1,0,1}} x_2 - Ex_1^{\boldsymbol{\theta}_{1,0,1}+\boldsymbol{\theta}_{2,0,1}} x_2^{\boldsymbol{\theta}_{2,0,2}} \\
&\quad - kEx_1^{1+\boldsymbol{\theta}_{2,0,1}} x_2^{1+\boldsymbol{\theta}_{2,0,2}} + Ex_1^{2\boldsymbol{\theta}_{2,0,1}} x_2^{2\boldsymbol{\theta}_{2,0,2}} \\
&= \frac{k}{2\left(2+\boldsymbol{\theta}_{1,0,1}\right)} - \frac{1}{\left(\boldsymbol{\theta}_{1,0,1}+\boldsymbol{\theta}_{2,0,1}+1\right)\left(\boldsymbol{\theta}_{2,0,1}+1\right)} \\
&\quad - \frac{k}{\left(2+\boldsymbol{\theta}_{2,0,1}\right)\left(2+\boldsymbol{\theta}_{2,0,2}\right)} \\
&\quad + \frac{1}{\left(1+2\boldsymbol{\theta}_{2,0,1}\right)\left(1+2\boldsymbol{\theta}_{2,0,2}\right)}.
\end{aligned}$$

$\alpha_{22} - \alpha_{12}$ is positive when $k < 1$ and k is close to 1.

Remark 6. Theorem 5 also has another important implication. In the supplementary materials, we show that when S is fixed, $\sup_{W\in\mathscr{H}_S} |L_n(W) - R_0(W)| = O_p(n^{-\frac{1}{2}})$ (in Lemma 2) and $R_0(\widehat{W}_n) \to_p \xi$ (in the proof of Theorem 5) hold under Assumptions 1–5. When $\xi > 0$, $L_n(e_s)/L_n(\widehat{W}_n) \to_p R_0(e_s)/\xi$. If $\inf_{1\leq s\leq S} R_0(e_s) > \xi$, that is, the optimal approximation is not achieved by any single model, we have

$$p\lim\left(\inf_{1\leq s\leq S} L_n(e_s)\right)/L_n(\widehat{W}_n) > 1.$$

With this condition, model selection is strictly inferior to NMA asymptotically.

Remark 7. In Theorem 5, we show that the model averaging weight $\widehat{W}_n$ will be arbitrarily close to the set $\mathscr{A}$ as sample size n increases. In many situations, empirical researchers also care about the risk function $R_n(W)$, which is the conditional expectation of $L_n(W)$ on $\mathbf{X}_n$. Now we further provide a result on the relationship between $\widehat{W}_n$ and the weights that minimize $R_n(W)$. Given any sequence of positive real numbers $a = \{a_n\}_{n=1}^{\infty}$ such that $a_n = o(1)$ and $a_n^{-1}n^{-\frac{1}{2}} = o(1)$, define $\mathscr{A}_{n,a} = \left\{W^* \in \mathscr{H}_S | R_n(W^*) - \inf_{W\in\mathscr{H}_S} R_n(W) \leq a_n\right\}$. Obviously, $\mathscr{A}_{n,a}$ contains the weights that asymptotically minimize $R_n(W)$. Different from $\mathscr{A}$, $\mathscr{A}_{n,a}$ depends on the sequence a, the sample size n as well as the observed sample realization $\mathbf{X}_n$. In the supplementary materials, we show that if Assumptions 1–5 hold, S is fixed, $\xi > 0$, and $\sup_{n\geq 1} E(||\sqrt{n}(\widehat{\boldsymbol{\theta}}_{s,n} - \boldsymbol{\theta}_{s,0})||^2) < \infty$ for all $1 \leq s \leq S$, then for any a that satisfies the above mentioned properties, there holds

$$\inf_{W\in\mathscr{A}_{n,a}} \left\|\widehat{W}_n - W\right\| \to_p 0. \tag{14}$$

Such a result implies that, the distance between model averaging weight $\widehat{W}_n$ and $\mathscr{A}_{n,a}$ tends to 0 as sample size n increases.

Up to now, we have extensively discussed the properties of the NMA based on the weight-choosing criterion NIC*(W). However, as we have discussed previously, the variance of the error term σ^2 usually requires estimation, and we have to use the NIC(W) instead of the infeasible criterion in practice. Since the difference between the NIC(W) and NIC*(W) lies only in the variance term, it is natural that the NMA results based on the NIC(W) do not differ much from the results based on the NIC*(W) as long as the estimator of σ^2 is reasonable. The following theorem provides a thorough description of the properties of NIC(W).

Theorem 6. Define $\widetilde{W}_n = \arg\min_{W\in\mathscr{H}_S}$ NIC(W) with $\widehat{\sigma}_n^2 = \widehat{\sigma}_n^2(s^*)$ for some $s^* \leq S$. We have

1. If Assumptions 1–5 hold, S is fixed and $\xi > 0$, then the results of Theorems 3 and 5 hold when $\widehat{W}_n$ is replaced with $\widetilde{W}_n$; if $\sup_{n\geq 1} E(||\sqrt{n}(\widehat{\boldsymbol{\theta}}_{s,n} - \boldsymbol{\theta}_{s,0})||^2) < \infty$ further holds for all $1 \leq s \leq S$, then (14) also holds when $\widehat{W}_n$ is replaced with $\widetilde{W}_n$;
2. If Assumptions 1, 2, and 6 hold, $\lim_{n\to\infty} S = \infty$, and $|f_{s^*}(X, \theta_{s^*})| \leq m(X)$ holds uniformly on Θ_{s^*} with $Em^2(X) < \infty$, then the results of Theorem 4 hold when $\widehat{W}_n$ is replaced with $\widetilde{W}_n$.

4. Simulation

In this section, we conduct extensive simulations to evaluate the finite-sample performance of the NIC and compare it with popular model selection and model averaging methods. In particular, the model selection methods include AIC, BIC, and Takeuchi information criterion (TIC). The model averaging methods include the smoothed AIC (SAIC) and the smoothed BIC (SBIC). For each candidate model, we estimate the parameters using the nonlinear least-squares method as in (3).

For the sth model, the AIC and BIC criteria are given by $\text{AIC}_n(s) = n\log(\widehat{\sigma}_n^2(s)) + 2K_s$ and $\text{BIC}_n(s) = n\log(\widehat{\sigma}_n^2(s)) + \log(n)K_s$, where $\widehat{\sigma}_n^2(s)$ is given in Section 3. The TIC criterion is given by $\text{TIC}_n(s) = n\log(\widehat{\sigma}_n^2(s)) + 2\text{tr}(\widehat{A}_s^{-1}\widehat{B}_s)$, where

$$\widehat{A}_s = -\sum_{t=1}^{n} \partial\ell_{s,t}^2(\widehat{\widetilde{\boldsymbol{\theta}}}_s)/\partial\widetilde{\boldsymbol{\theta}}_s\partial\widetilde{\boldsymbol{\theta}}_s^T,$$

$$\widehat{B}_s = \sum_{t=1}^{n} \left(\partial\ell_{s,t}(\widehat{\widetilde{\boldsymbol{\theta}}}_s)/\partial\widetilde{\boldsymbol{\theta}}_s\right)\left(\partial\ell_{s,t}(\widehat{\widetilde{\boldsymbol{\theta}}}_s)/\partial\widetilde{\boldsymbol{\theta}}_s^T\right),$$

where $\widetilde{\boldsymbol{\theta}}_s^T = (\boldsymbol{\theta}_s^T, \sigma^2)$, $\widehat{\widetilde{\boldsymbol{\theta}}}_s = \left(\widehat{\boldsymbol{\theta}}_s^T, \widehat{\sigma}^2\right)$, and $\ell_{s,t}(\widetilde{\boldsymbol{\theta}}_s) = -\frac{1}{2}\log 2\pi\sigma^2 - (y_t - f_s(X_t, \boldsymbol{\theta}_s))^2/2\sigma^2$. The weights of SAIC and SBIC for the sth model are

$$\exp\left(-\frac{1}{2}\text{AIC}_n(s)\right)/\sum_{s=1}^{S}\exp\left(-\frac{1}{2}\text{AIC}_n(s)\right)$$

and

$$\exp\left(-\frac{1}{2}\text{BIC}_n(s)\right)/\sum_{s=1}^{S}\exp\left(-\frac{1}{2}\text{BIC}_n(s)\right),$$

respectively. The estimated variance in the NIC is taken as the smallest estimated variance among all the candidate models. To evaluate the performance of different model selection and averaging methods, the loss function is calculated as $\sum_{t=1}^{n}\left(\mu(X_t)-\widehat{y}_t\right)^2$, where $\mu(X_t)$ is the unknown conditional mean and $\widehat{y}_t$ is the estimation of $\mu(X_t)$ under different methods. We repeat the simulation 1000 times and the averaged losses are calculated as the risks. For comparison purposes, we report the relative risk by dividing the risk of other methods by that of NMA.

Assume the data we observe is $\{X_t, y_t\}, t = 1,\ldots,n$, where $X_t = (x_{t1},\ldots,x_{tK})$ is the K-dimensional covariate vector. The true data generating process is given by

$$y_t = \prod_{k=1}^{K} x_{tk}^{\alpha_k} + \varepsilon_t, \tag{15}$$

where $x_{tk} \overset{\text{iid}}{\sim} \text{Unif}(0.5, 1.5)$ across k and t, $\alpha_k = k^{\delta}$, δ is a constant controlling the elasticity mechanism of x_{tk}, and ϵ_t is the random error to be specified. In particular, when $\delta < 0$, the exponent of x_{tk} decreases with k, and the exponent of x_{tk} is a constant 1 across k when $\delta = 0$. The following four cases are studied.

- Case 1: $\delta = -0.25$, misspecified scenario.
- Case 2: $\delta = 0$, misspecified scenario.
- Case 3: $\delta = -0.25$, correctly specified scenario.
- Case 4: $\delta = 0$, correctly specified scenario.

For each case, the candidate model set is given by

$$\left\{ f_s(X_t, \boldsymbol{\theta}_s) = \theta_0 \prod_{k=1}^{s} x_{tk}^{\theta_k}, s = 1, 2, \ldots, S \right\}, \tag{16}$$

where S is the number of candidate models. The misspecified scenario corresponds to $K = 10$ and $S = 5$ and the correctly specified scenario corresponds to $K = S = 10$. For each case, we consider the sample sizes $n = 100$, $n = 200$, and $n = 500$, and $\epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$ or $\sigma \cdot t(4)$, and vary the value of σ such that the population R^2 ranges from 0.1 to 0.9 with increment 0.1. In the misspecified scenario, all candidate models are misspecified, whereas in the correctly specified scenario, the correct model is inside the candidate model set.

The results are presented in Figures 1–4. From Figure 1, we observe that the NIC achieves the lowest risk in most situations, with the improvement most prominent when $R^2 \leq 0.5$. Interestingly, the model selection methods AIC and BIC are inferior to their model averaging counterparts SAIC and SBIC in all scenarios, showing the benefits of model averaging. In many scenarios, the heavy tail error leads to a slightly larger gain for NIC compared with the case of normal error. As n increases, different methods' performance becomes more similar when R^2 is large, but NIC still outperforms the other approaches when R^2 is small.

Another observation is that TIC seems to perform worse than AIC, under this misspecified setting. Although TIC was shown to be superior to AIC asymptotically under model misspecification (Burnham and Anderson 2002), it is not always the case for finite sample scenarios. For example, Yanagihara (2006) showed that in some cases, the frequencies of selecting the true models of TIC are less than that of AIC; as an estimator of Kullback–Leibler divergence, TIC has larger bias than AIC. For our particular data generation process, the calculation of TIC involves estimating the 4th moment of the error term which does not exist for the t distribution with degrees of freedom 4. This could also lead to the worse performance of TIC than that of AIC. See Table 3.4 in Konishi and Kitagawa (2008) for a detailed comparison.

In general, we observe similar messages for Cases 2–4. Another notable finding is that, even in the correctly specified scenario, using the nonlinear model averaging with the NIC leads to a smaller risk compared with the model selection methods.

5. Empirical Application: Predicting the Wage

In this section, we revisit the empirical example in Hansen and Racine (2012) and apply the NMA with the NIC to the prediction of the individual wage. The empirical data[2] come from the Wooldridge (2003) cross-sectional dataset "wage1," which contains 526 observations taken from the US Current Population Survey for the year 1976. The goal is to predict the log of average hourly earnings (the dependent variable) using 20 explanatory variables: educ, exper, tenure, nonwhite, female, married, numdep, smsa, northcen, south, west, construc, ndurman, trcommpu, trade, services, profserv, profocc, clerocc, and servocc. Following Hansen and Racine (2012), in addition to the 20 original variables, we also consider the following nine possible interaction terms: educ×nonwhite, educ×female, educ×married, exper×nonwhite, exper×female, exper×married, tenure×nonwhite, tenure× female, and tenure×married.

For individual i, let w_i represent the average hourly earnings and x_{ik} represent the kth explanatory variables according to the order introduced above. The *linear candidate model set* contains 30 nested models $\mathcal{M}^l = \{M_1^l,\ldots,M_{30}^l\}$, where the mth ($1 \leq m \leq 30$) candidate model M_m^l is given by

$$\log w_i = \theta_0 + \sum_{k<m} \theta_k x_{ik} + \varepsilon_i. \tag{17}$$

The above constructing procedure of the model set is the same as in Hansen and Racine (2012). The linear candidate model set $\mathcal{M}^l$ ranges from the null model with only the intercept to the model including all original variables as well as the nine interaction terms.

To study the nonlinear effects of the continuous variables (education, experience, and tenure) on the log-wage, we introduce possible nonlinear factors including $\psi_{11}\widetilde{x}_{i1}^{\psi_{12}}$, $\psi_{21}\widetilde{x}_{i2}^{\psi_{22}}$, and $\psi_{31}\widetilde{x}_{i3}^{\psi_{32}}$. These factors describe the nonlinear impacts of the continuous variables on the individual log-wage in a flexible way. For example, when the first nonlinear factor is added into the model, the marginal return of education on the log-wage, which is $\partial \log w_i/\partial x_{i1}$, now becomes $\theta_1 + \psi_{11}\psi_{12}\widetilde{x}_{i1}^{\psi_{12}-1}$. When $\psi_{11} > 0$ and $\psi_{12} \in (0,1)$ hold, $\partial^2 \log w_i/\partial x_{i1}^2 < 0$,

[2]The data are available at *http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html*

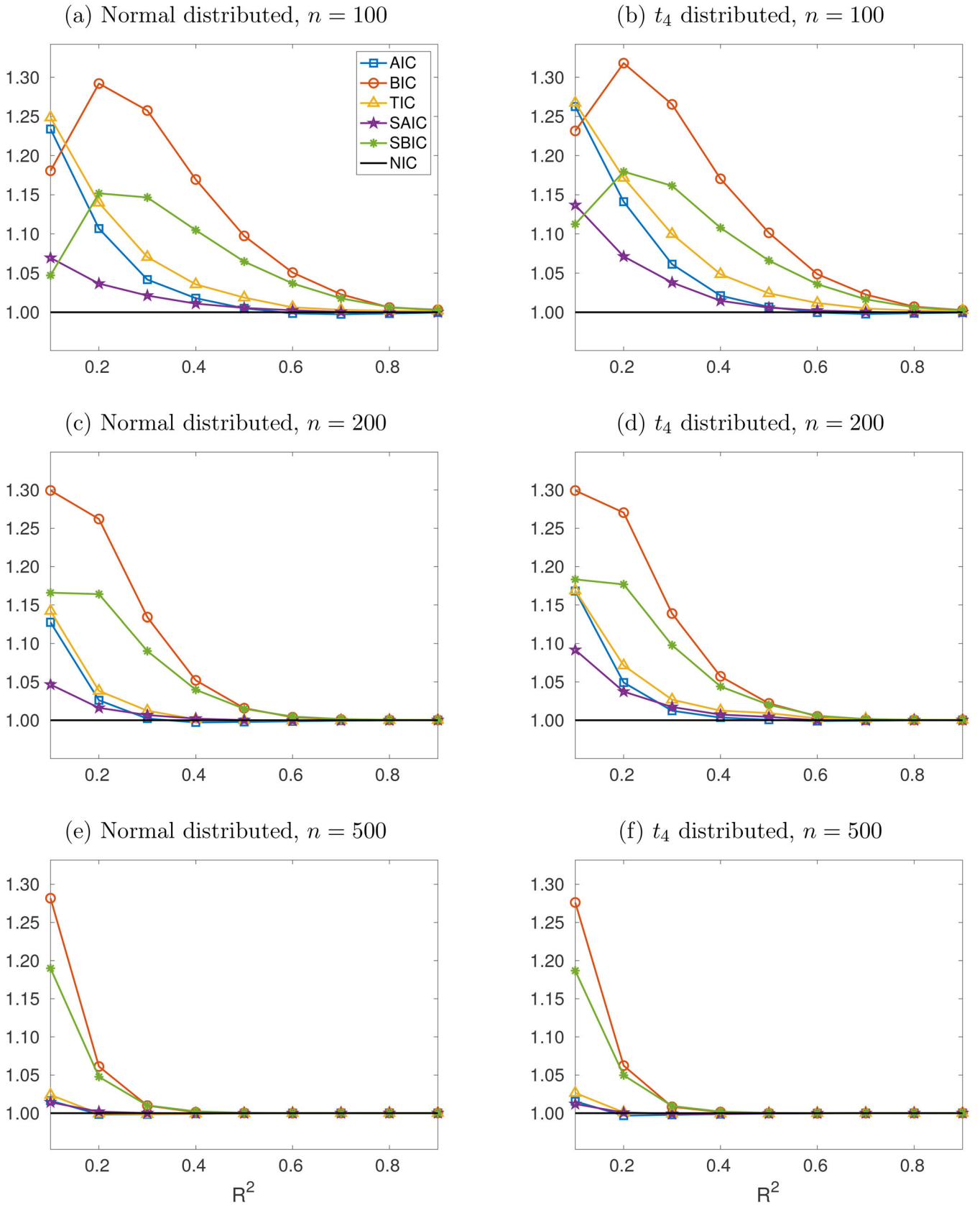

Figure 1. Relative risk when comparing with NMA for Case 1.

implying the diminishing marginal returns of education. When $\psi_{11} > 0$ and $\psi_{12} \in (1, 2)$, the marginal return of education increases as the education years increases, but the increasing speed gradually drops. Finally, for $\psi_{11} > 0$ and $\psi_{12} > 2$, both the marginal return and its increasing speed rises with that of years of education. Note that when $\psi_{11} = 0$, the nonlinear factor vanishes and log-wage depends on education linearly; when $\psi_{12} = 2$, the nonlinear factor degenerates to the quadratic

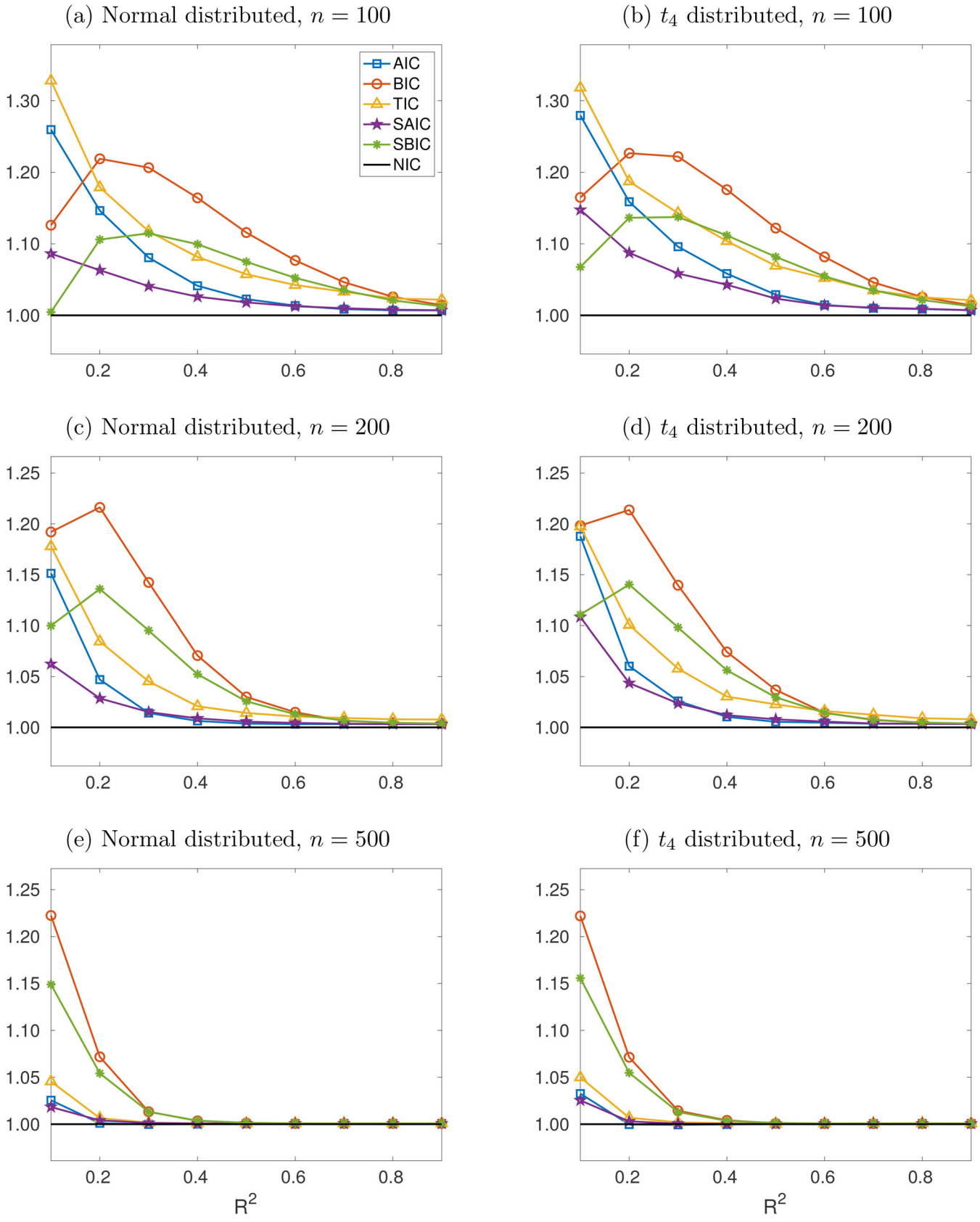

Figure 2. Relative risk when comparing with NMA for Case 2.

term of education. So our specification is general enough to nest some widely applied setups.

The *nonlinear candidate models set*, $\mathcal{M}^n = \{M_1^n, M_2^n, M_3^n, M_{12}^n, M_{13}^n, M_{23}^n, M_{123}^n\}$, contains seven models, where M_k^n refers to the candidate with the kth nonlinear factor, $M_{k_1k_2}^n$ refers to the candidate with the k_1 and k_2th nonlinear factors, and M_{123}^n refers to the largest candidate model containing all three nonlinear terms. More specifically,

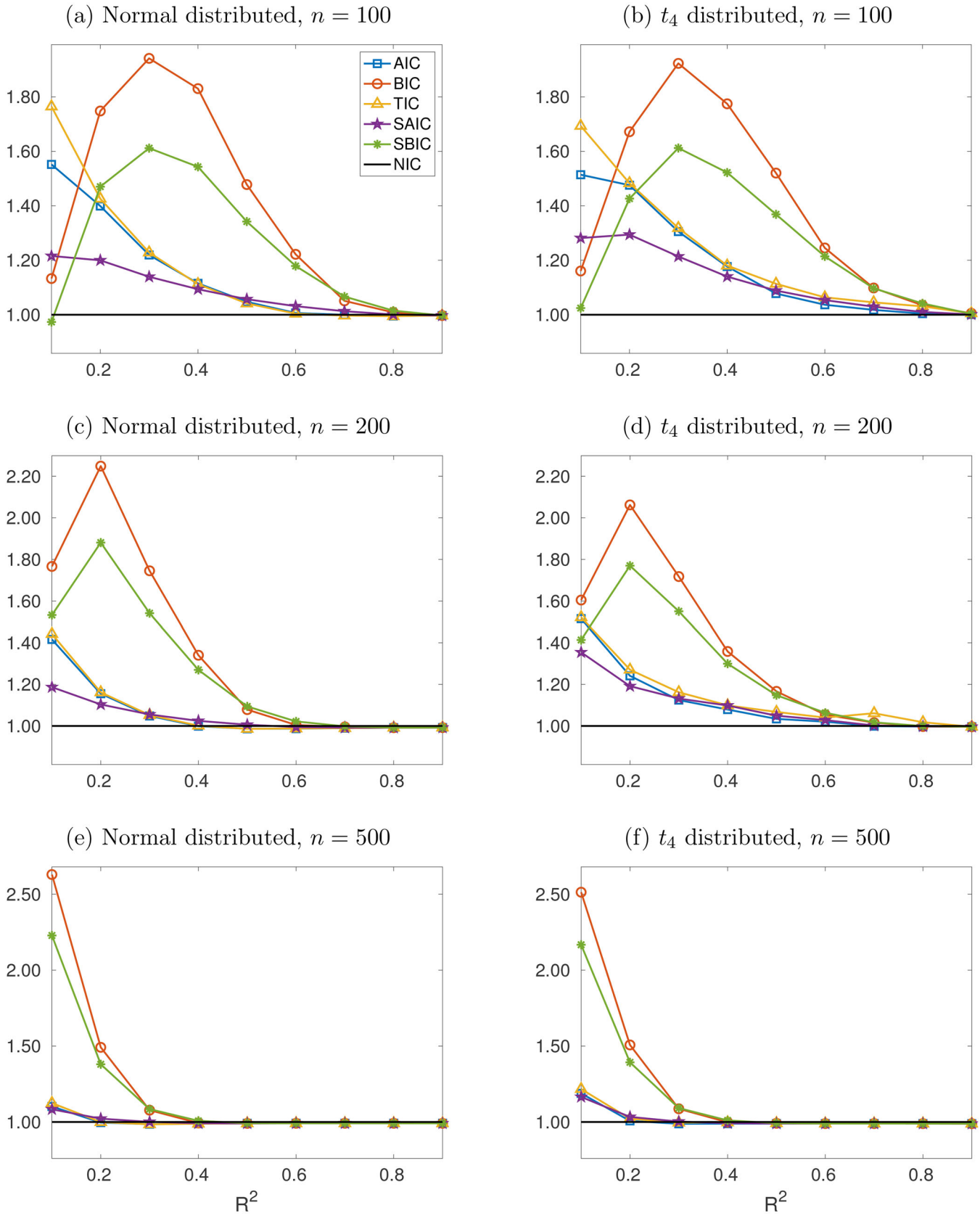

Figure 3. Relative risk when comparing with NMA for Case 3.

we have

$$\log w_i = \theta_0 + \sum_{k<30} \theta_k x_{ik} + \sum_{k \in G_s} \psi_{k1} \widetilde{x}_{ik}^{\psi_{k2}} + \varepsilon_i, \qquad (18)$$

where $G_s \in \mathcal{G}$ and $\mathcal{G} = \{\{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$. Note that unlike the linear candidate models, the nonlinear candidate models are not nested to one another.

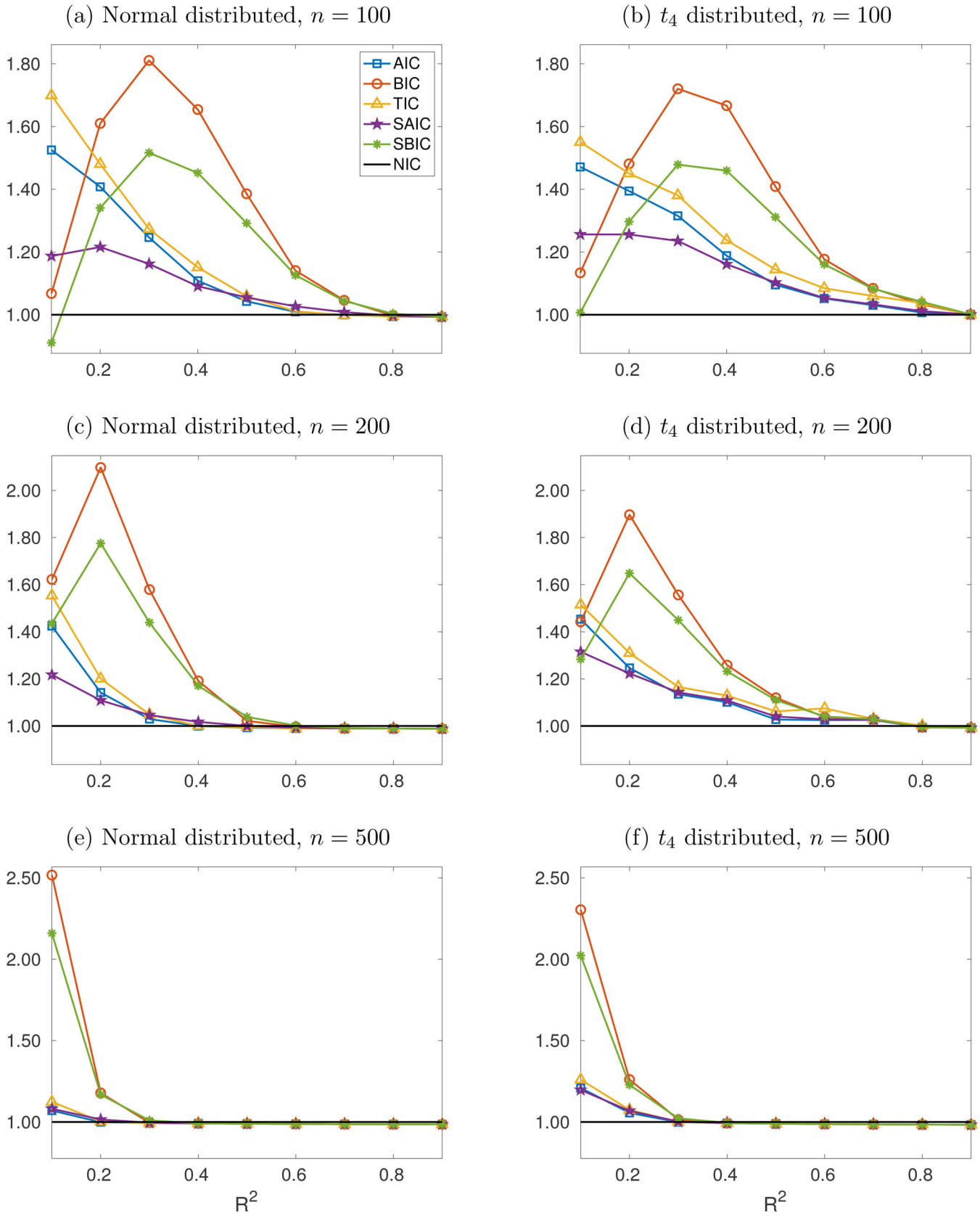

Figure 4. Relative risk when comparing with NMA for Case 4.

Now, the *full candidate model set* $\mathcal{M} = \mathcal{M}^l \cup \mathcal{M}^n$ has a total of 37 models, all of which are estimated using the nonlinear least-squares method in (3).

Following Hansen and Racine (2012), we randomly choose n_{train} observations as the training set, and calculate the mean squared prediction error (MSPE) on the remaining n_{test} test

Table 1. Relative mean and median of MSPE over 1000 random splits.

n_{train}		FULL	AIC	BIC	TIC	SAIC	SBIC	MMA	JMA
100	Mean	1.2079	1.1686	1.0283	1.1340	1.0716	1.0064	0.9520	0.9293
	Median	1.1763	1.0969	1.0782	1.1060	1.0435	1.0623	0.9766	0.9714
200	Mean	1.0690	1.0632	1.1585	1.0606	1.0408	1.1395	1.0125	1.0107
	Median	1.0636	1.0581	1.1577	1.0565	1.0379	1.1409	1.0129	1.0123
300	Mean	1.0454	1.0528	1.1553	1.0492	1.0389	1.1305	1.0249	1.0239
	Median	1.0486	1.0557	1.1641	1.0508	1.0422	1.1384	1.0257	1.0262
400	Mean	1.0378	1.0423	1.0631	1.0433	1.0343	1.0609	1.0306	1.0295
	Median	1.0365	1.0427	1.0598	1.0420	1.0336	1.0575	1.0278	1.0265
500	Mean	1.0329	1.0415	1.0521	1.0433	1.0303	1.0525	1.0352	1.0356
	Median	1.0365	1.0442	1.0445	1.0394	1.0289	1.0458	1.0373	1.0338

Table 2. Model averaging weights of different models.

n_{train}	Linear	M_1^n	M_2^n	M_3^n	M_{12}^n	M_{13}^n	M_{23}^n	M_{123}^n
100	0.6434	0.0315	0.0234	0.0814	0.0282	0.0801	0.0643	0.0477
200	0.4850	0.0378	0.0180	0.0694	0.0425	0.1013	0.1416	0.1044
300	0.3836	0.0423	0.0113	0.0611	0.0377	0.0736	0.2311	0.1593
400	0.3475	0.0240	0.0065	0.0383	0.0246	0.0816	0.2922	0.1853
500	0.3251	0.0095	0.0010	0.0072	0.0074	0.0709	0.3326	0.2463

observations. We vary n_{train} from 100 to 500 with increment 100 to see the effect of sample size increasing.

We evaluate the performances of the NIC as well as other model selection and averaging methods including AIC, BIC, TIC, SAIC, and SBIC considered in the simulation section, evaluated on the full candidate model set. In addition, we consider the performance of FULL (the largest nonlinear model M_{123}^n), MMA (Hansen 2007), and JMA (Hansen and Racine 2012) on the linear candidate model set $\mathcal{M}^l$. We repeat the random splitting 1000 times for each method and calculate the mean and median of MSPE. Table 1 reports the mean and median of MSPE for the competing methods relative to those of the NIC. It is clear that when $n_{\text{train}} \geq 200$, the NMA has the lowest MSPE in terms of both mean and median among all methods considered. It is interesting to note that when $n_{\text{train}} = 100$, the linear model averaging methods MMA and JMA perform better than the NIC. As we have more training data, the advantage of the NIC over linear model averaging becomes more evident, which indicates there are possible nonlinear effects.

Now, we report the average weights of different models for the NIC in Table 2, where "linear" represents the total weights assigned to the linear candidate models. From the table, we can see that as n_{train} increases, the total weight assigned to linear candidate models decreases, and the weights corresponding to M_{23}^n and M_{123}^n increase monotonically. This may indicate that the nonlinear factors corresponding to experience and tenure play an important role in the model, and considering the nonlinear regression model is critical.

6. Concluding Remarks

This article considered the NMA framework and advocated the use of the NIC as the weight-choosing criterion. We proved the optimality of the NIC, and showed that the model-averaging weights selected by minimizing NIC converge to a well-defined limit.

Extensive simulation studies illustrated that the NIC outperformed competing methods in most situations. The empirical application of wage predication also demonstrates the superiority of the new method over alternatives. One interesting future work is to extend the current approach to high-dimensional settings. We expect the NIC leads to sparse solutions, as the MMA does for linear regression models (Feng, Liu, and Okui 2020). Another possible research direction is to handle the case where the errors are heteroscedastic or autocorrelated.

Supplementary Materials

The online supplement contains the proofs, the codes for simulations and the empirical application, and the data for the empirical application.

Acknowledgments

We thank the co-editor, associate editor, and two anonymous referees for their constructive comments which greatly improved the quality and scope of the article.

Funding

This research was partially supported by NSF CAREER grant DMS-2013789 (Feng), JSPS KAKENHI grant number JP16K03590 and JP19K01582 (Liu), and Outstanding Innovative Talents Cultivation Funded Programs 2018 of Renmin University of China (Yao). All authors contributed equally to this work.

References

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *International Symposium on Information Theory*, pp. 610–624. [1]

Ando, T. and Li, K.-C. (2014), "A Model-Averaging Approach for High-Dimensional Regression," *Journal of the American Statistical Association*, 109, 254–265. [2]

——— (2017), "A Weighted-Relaxed Model Averaging Approach for High-Dimensional Generalized Linear Models," *The Annals of Statistics*, 45, 2654–2679. [2]
Bates, D. M., and Watts, D. G. (1988), *Nonlinear Regression Analysis and Its Applications* (Vol. 2), New York: Wiley. [2]
Bates, J. M., and Granger, C. W. (1969), "The Combination of Forecasts," *Journal of the Operational Research Society*, 20, 451–468. [1]
Burnham, K. P., and Anderson, D. R. (2002), *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach* (2nd ed.), New York: Springer. [8]
Chen, J., Li, D., Linton, O., and Lu, Z. (2018), "Semiparametric Ultra-High Dimensional Model Averaging of Nonlinear Dynamic Time Series," *Journal of the American Statistical Association*, 113, 919–932. [2]
Cheng, T. C. F., Ing, C. K., and Yu, S. H. (2015), "Toward Optimal Model Averaging in Regression Models With Time Series Errors," *Journal of Econometrics*, 189, 321–334. [1]
Cheng, X., and Hansen, B. E. (2015), "Forecasting With Factor-Augmented Regression: A Frequentist Model Averaging Approach," *Journal of Econometrics*, 186, 280–293. [2]
Claeskens, G., and Hjort, N. L. (2003), "The Focused Information Criterion," *Journal of the American Statistical Association*, 98, 900–916. [1]
——— (2008), *Model Selection and Model Averaging*, Cambridge: Cambridge University Press. [1,2]
Clyde, M., and George, E. I. (2004), "Model Uncertainty," *Statistical Science*, 19, 81–94. [1]
Dardanoni, V., De Luca, G., Modica, S., and Peracchi, F. (2015), "Model Averaging Estimation of Generalized Linear Models With Imputed Covariates," *Journal of Econometrics*, 184, 452–463. [1]
De Luca, G., Magnus, R. J., and Peracchi, F. (2018), "Weighted-Average Least Squares Estimation of Generalized Linear Models," *Journal of Econometrics*, 204, 1–17. [2]
Draper, D. (1995), "Assessment and Propagation of Model Uncertainty," *Journal of the Royal Statistical Society*, Series B, 57, 45–70. [1]
Feng, Y., Liu, Q., and Okui, R. (2020), "On the Sparsity of Mallows Model Averaging Estimator," *Economics Letters*, 187, 1–5. [13]
Foster, D. P., and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975. [1]
Gao, Y., Zhang, X., Wang, S., and Zou, G. (2016), "Model Averaging Based on Leave-Subject-Out Cross-Validation," *Journal of Econometrics*, 192, 139–157. [1]
Hansen, B. E. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [1,3,5,13]
——— (2014), "Model Averaging, Asymptotic Risk, and Regressor Groups," *Quantitative Economics*, 5, 495–530. [1]
Hansen, B. E., and Racine, J. S. (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46. [1,8,12,13]
Heckman, J. J., Lochner, L. J., and Todd, P. E. (2008), "Earnings Functions and Rates of Return," *Journal of Human Capital*, 2, 1–31. [2]
Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401. [1]
Jennrich, R. I. (1969), "Asymptotic Properties of Non-Linear Least Squares Estimators," *The Annals of Mathematical Statistics*, 40, 633–643. [3,4,5]
Konishi, S., and Kitagawa, G. (2008), *Information Criteria and Statistical Modeling*, New York: Springer. [8]
Kuersteiner, G., and Okui, R. (2010), "Constructing Optimal Instruments by First-Stage Prediction Averaging," *Econometrica*, 78, 697–718. [2]
Lemieux, T. (2006), "The 'Mincer Equation' Thirty Years After Schooling, Experience, and Earnings," in *Jacob Mincer a Pioneer of Modern Labor Economics*, ed. S. Grossbard, Boston, MA: Springer, pp. 127–145. [2]
Liu, Q., and Okui, R. (2013), "Heteroscedasticity-Robust C_p Model Averaging," *The Econometrics Journal*, 16, 463–472. [1]
Liu, Q., Okui, R., and Yoshimura, A. (2016), "Generalized Least Squares Model Averaging," *Econometric Reviews*, 35, 1692–1752. [1,3]
Liu, Q., Yao, Q., and Zhao, G. (2020), "Model Averaging Estimation for Conditional Volatility Models With an Application to Stock Market Volatility Forecast," *Journal of Forecasting*, 39, 841–863. [2]
Lu, X., and Su, L. (2015), "Jackknife Model Averaging for Quantile Regressions," *Journal of Econometrics*, 188, 40–58. [2]
Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675. [1,4,5]
Mincer, J. (1996), "Changes in Wage Inequality, 1970–1990," NBER Working Paper 5823. [2]
Moral-Benito, E. (2015), "Model Averaging in Economics: An Overview," *Journal of Economic Surveys*, 29, 46–75. [2]
Murphy, K. M., and Welch, F. (1990), "Empirical Age-Earnings Profiles," *Journal of Labor Economics*, 8, 202–229. [2]
Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 15–18. [1]
Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Series B, 36, 111–133. [1]
Sueishi, N. (2013), "Generalized Empirical Likelihood-Based Focused Information Criterion and Model Averaging," *Econometrics*, 1, 141–156. [2]
White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838. [1]
——— (1981), "Consequences and Detection of Misspecified Nonlinear Regression Models," *Journal of the American Statistical Association*, 76, 419–433. [3,4]
Wooldridge, J. M. (2003), *Introductory Econometrics*, Mason, OH: Thomson South-Western. [8]
Yanagihara, H. (2006), "Corrected Version of AIC for Selecting Multivariate Normal Linear Regression Models in a General Nonnormal Case," *Journal of Multivariate Analysis*, 97, 1070–1089. [8]
Yang, Y. (2001), "Adaptive Regression by Mixing," *Journal of the American Statistical Association*, 96, 574–588. [3]
Zhang, X., and Liang, H. (2011), "Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models," *The Annals of Statistics*, 39, 174–200. [2,3]
Zhang, X., Ullah, A., and Zhao, S. (2016), "On the Dominance of Mallows Model Averaging Estimator Over Ordinary Least Squares Estimator," *Economics Letters*, 142, 69–73. [1]
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. [2,3,6]
Zhang, X., Zou, G., and Carroll, R. J. (2015), "Model Averaging Based on Kullback–Leibler Distance," *Statistica Sinica*, 25, 1583–1598. [2]
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. [2]