Nonparametric Instrument Model Averaging

Jianan Chen
Department of Statistics and Applied Probability
National University of Singapore
and
Binyan Jiang
Department of Applied Mathematics
Hong Kong Polytechnic University
and
Jialiang Li*
Department of Statistics and Applied Probability
National University of Singapore

April 20, 2023

Abstract

We present a new nonparametric model averaging approach to the instrumental variable (IV) regression where the effects of multiple instruments on the endogenous variable are modeled as nonparametric functions in the reduced form equations. Even if individual IVs may have weak and nonlinear relevance to the exposure, our proposed model averaging is able to ensemble their effects with optimal weights to produce valid inference. Our analysis covers both the case in which the number of IV is fixed and the case in which the dimension of IV is diverging with sample size. This novel framework can be especially beneficial to the practical situations involving weak IVs since in many recent observational studies we may encounter a large number of instruments and their quality could range from poor to strong. Numerical studies are carried out and comparisons are made between our proposed method and a wide range of existing alternative methods.

Keywords: Endogeneity, Instrumental variable, Model Averaging, Nonparametric regression, Penalty function, Two-stage least squares.

*Corresponding author. E-mail: *stalj@nus.edu.sg*

1 Introduction

Theory and methods for estimating causal effects in presence of unmeasured confounders from observational studies are abundant and still growing these days (cf. Austin (2011), McCaffrey et al. (2013), Imai et al. (2013) and Austin and Stuart (2015), among others). An attractive solution to the endogeneity issue is to employ the so-called instrumental variable (IV), for which a two-stage least squares (2SLS) estimation procedure is usually adopted. The inference of 2SLS crucially depends on the quality of the available instruments. IVs with poor quality not only lead to imprecise estimates of the structural parameters but also affect the standard diagnostic statistics, that we use to assess these estimates, to be unreliable. In many modern datasets we may collect a large number of potentially useful IVs and their quality could vary in a wide range. How to effectively conduct a satisfactory inference with these IVs is an important research question. We propose a nonparametric model averaging approach to contribute towards this topic.

A qualified IV should satisfy the *relevance* condition, with a moderate to high degree of correlation with the endogenous variables (Hall et al., 1996; Greenland, 2000). However, when facing large number of IVs this assumption might be violated for some instruments. We stress that the use of strong instruments is critical to the validity of the inferences using 2SLS (Bollen, 1989; Bound et al., 1995; Kline, 1998). Weak IVs are known to undermine the causal inferences with misleading results (Nelson and Startz, 1990). See Bound et al. (1995); Shea (1996); Donald and Newey (2001); Hall and Peixe (2003); Hall et al. (2007); Hansen et al. (2008) for some earlier developments. All these works only examine the linear association between IV and the endogenous variable. In this work we will relax the linearity assumption and consider a more general functional dependence in the reduced form equation. In fact, a relevant IV with strong nonlinear dependency on the endogenous variable may appear to be weak when mistakenly used in a linear model. Our methodology thus offers greater flexibility to the IV regression.

Strengthening instruments often comes with the expense of a reduced data size (Angrist and Krueger, 1995; Small and Rosenbaum, 2008; Baiocchi et al., 2010; Zubizarreta et al., 2013; Keele and Morgan, 2016). On the other hand, model averaging does not require data removal, and can easily incorporate weak IVs. See Yang (2003); Moral-Benito (2015); Hjort and Claeskens (2003) for some earlier works in this area. Compared to model selection, model averaging is less sensitive to model mis-specification and structural assumptions. It has the advantage of avoiding the need to defend the choice of a single "best" model. A key step for model averaging is to estimate the optimal weights assigned to sub-models (Sloughter et al., 2010; Koop and Korobilis, 2012; Wang et al., 2016). Some authors also considered model averaging for the traditional 2SLS setting. For example, built on the linear instrumental variable regression framework, Martins and Gabriel (2014) proposed to obtain the model averaging weights via a direct smoothing of information criteria attained in the two stage least squares estimation procedure. Other related approaches include Clyde et al. (2011); Koop et al. (2012); Yu et al. (2014); Zeugner et al. (2015); Burgess et al. (2018). Nonparametric model averaging only became popular in recent years (Li et al. (2015, 2018); Huang and Li (2018); Fang et al. (2022); Li et al. (2022)). But from our review there has been no work of using nonparametric model averaging for IV regression yet.

Our proposed nonparametric model averaging method can accommodate high-dimensional IVs. Recent application of genetic biomarkers, typically single nucleotide polymorphisms (SNPs), as instruments in Mendelian randomization studies (Didelez and Sheehan, 2007; Wehby et al., 2008; Lawlor et al., 2008; Lin et al., 2015) has boosted research works on

IVs with very high dimensions. Belloni et al. (2014) conducted an overview on how model selection methods can be adapted to provide better inference on causal parameters. Belloni et al. (2012) proposed the post-Lasso estimator to alleviate the shrinkage bias. The Adaptive Lasso method was also adopted to select instruments in the first stage of 2SLS (Caner and Fan, 2010). Fan and Liao (2014) considered the endogeneity issue with diverging number of covariates and provided a necessary technical condition for consistency. Lin et al. (2015) proposed regularisation methods for high-dimensional IVs and applied their approaches in genomics. Guo et al. (2018) showed that Durbin-Wu-Hausman (DWH) test maintains the correct size for high-dimensional covariates and further proposed a specification test to improve the power. Seng and Li (2022) recently considered model averaging for structural equation models and offered some empirical findings under high-dimensional IV setting. However, none of these authors considered the nonparametric functional dependence for IVs. With diverging number of IVs, we will explicitly allow the number of nonparametric IV models to diverge with sample size. A regularized step is adopted to determine the model weights and a few familiar penalty functions are recommended for this purpose. Our procedure is easy to implement and also enjoys solid theoretical support.

2 Methods

We suppose that all variables are centralised and therefore leave intercept terms out in this paper. The true model for explaining the causal effects of covariate X on the response variable Y is

$$Y = X\beta + \epsilon, \tag{1}$$

where X is an endogenous variable correlated with the model error ϵ. The relationship can be fully characterized if we assume $E(\epsilon|X) = 0$ or, equivalently $E(Y|X) = X\beta$. However, in numerous structural econometric models or observational studies in biostatistics, the conditional expectation function is not the parameter of interest. The structural parameter or causal parameter $\beta \in \mathbb{R}$ spells a relation between Y and X, where X is endogenous and so $E(\epsilon|X) \neq 0$. The coefficient β is of interest and it is a major measure of the causal effect of X on Y in presence of unmeasured confounding effects from ϵ. We do not involve additional exogenous variables to simplify the notation and presentation. Those covariates can be handled easily by adding a projection step in practice.

When the unobserved error term ϵ is correlated with the endogenous variable X, the estimation of β using the ordinary least squares (OLS) method by fitting a linear regression on X might lead to biased estimates. The 2-stage least squares (2SLS) method can be used to obtain a consistent estimator of β. The standard IV methodology just applies traditional linear regression on X. In this paper we consider nonparametric regression on the endogenous variable X. Thus suppose there are d instrumental variables $Z_1, Z_2, \ldots, Z_d$, and we consider a fully nonparametric reduced form equation for X:

$$X = m(\mathbb{Z}) + e \tag{2}$$

where $m : \mathbb{R}^d \to \mathbb{R}$ is an unknown function, $\mathbb{Z} = (Z_1, Z_2, \ldots, Z_d)^\top$ is a d-dimensional random vector, and e is the random error term.

The following model conditions are assumed to hold in this paper.

(C1) $\mathbb{Z}$ and X are dependent, i.e. m is not a constant function.

(C2) $E(\epsilon|\mathbb{Z}) = 0$.

(C3) $E(e|\mathbb{Z}) = 0$.

The joint conditions of (C1) and (C2) specify the relationship between $\mathbb{Z}$ and the variables in model (1). In fact, $\mathbb{Z}$ is a relevant predictor of Y and its effect is through X only when the conditions are met. It implies the exogeneity of the instruments relative to the structural model (1). These are the standard requirements for 2SLS results to be valid. Condition (C3) implies the exogeneity of $\mathbb{Z}$ holds and standard nonparametric regression can be applied to fit the first stage model.

We stress that we should try our best to check these conditions in a real world study. Specifically, to verify (C1), sometimes we can use some standard statistics such as Pearson correlation for a marginal association, F-statistics, or "partial R^2" of the IVs in the first stage as useful indicators of the quality of the IVs. A common test used to check (C2) is the well-known J-test of overidentifying restrictions. There are also plenty of recent developments on this issue in the literature. For example, see Kang et al. (2016). (C3) is usually assumed so that IVs are not related to unmeasured variables that affect the exposure. It is also part of the validity for IV and can be checked similarly as (C2). However, unlike (C1), the validity assumption often cannot be completely tested and we have to carry out sensitivity analysis case-by-case.

2.1 Nonparametric Instrument Model: Single IV

In this part we consider the simple case where $d = 1$. We give results for single IV case first and will proceed to the case with multiple IVs in the next subsection. Suppose that we are given the independent data of sample size n for the response variable $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$, endogenous variable $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$ and instruments $\mathbf{Z} = (\mathbb{Z}_1, \mathbb{Z}_2, \ldots, \mathbb{Z}_n)^\top$. According to equation (1), we can write $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$ is an n-vector.

Let $\mathcal{Z}$ be the support of the instrumental variable Z and assume that $\mathcal{Z}$ is a compact set in $\mathcal{R}$. We use a J_n-dimensional equispaced B-spline basis on $\mathcal{Z}$, denoted by $\mathbf{B}(z) = (B_1(z), \ldots, B_{J_n}(z))^\top$, to approximate the function $m(z)$, that is $m(z) \approx \mathbf{B}(z)^\top\boldsymbol{\theta}$, where

$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{J_n})^\top$ is a vector of coefficients of $\mathbf{B}(z)$. See Schumaker (2007) for the definition and properties of B-spline basis.

To establish inference results, we need to assume the following conditions.

(C4) Let $s > 0$ be an integer, $\alpha \in (0, 1]$ be a constant such that $r = s + \alpha > 2$ and $L > 0$ be a positive constant. We use $\mathbb{H}(r, L)$ to denote the set of functions on $\mathscr{Z}$ such that for every $h \in \mathbb{H}(r, L)$, the sth derivative of h, denoted by $h^{(s)}$, exists and satisfies the following Hölder condition of order α:

$$|h^{(s)}(t_1) - h^{(s)}(t_2)| \leq L|t_1 - t_2|^\alpha \quad \text{for} \quad t_1, t_2 \in \mathscr{Z}.$$

We assume that Z is a random variable in $\mathscr{Z}$ with density function bounded away from 0 and infinity, and $m \in \mathbb{H}(r, L)$.

(C5) We take $J_n = \lfloor cn^{1/(2r+1)} \rfloor$ for some positive constant c, where $\lfloor t \rfloor$ is the largest integer no greater than t.

In matrix notation, denote $\mathbf{m} = (m(\mathbb{Z}_1), \ldots, m(\mathbb{Z}_n))^\top, \mathbf{B} = (\mathbf{B}(\mathbb{Z}_1), \ldots, \mathbf{B}(\mathbb{Z}_n))^\top$. From equation (2), we have the linear model $\mathbf{X} = \mathbf{m} + \mathbf{e} \approx \mathbf{B}\boldsymbol{\theta} + \mathbf{e}$ where $\mathbf{e} = (e_1, e_2, \ldots, e_n)^\top$. We then obtain the predicted value $\hat{\mathbf{m}} = \mathbf{B}\hat{\boldsymbol{\theta}}$ in the first stage of 2SLS where the OLS estimator $\hat{\boldsymbol{\theta}} = \left(\mathbf{B}^\top\mathbf{B}\right)^{-1}\mathbf{B}^\top\mathbf{X}$ and $\hat{\mathbf{m}}$ is used to substitute $\mathbf{X}$ in the second stage when fitting a linear model (1). After a standard derivation, the 2SLS estimator for β is given by

$$\hat{\beta} = (\hat{\mathbf{m}}^\top\hat{\mathbf{m}})^{-1}\hat{\mathbf{m}}^\top\mathbf{Y} \tag{3}$$

Now we are ready to state the consistency and asymptotic normality of $\hat{\beta}$ under the preceding conditions. We denote the normal distribution with mean η and covariance Ω by $N(\eta, \Omega)$, and by " $\xrightarrow{P}$ " we mean convergence in probability, by " $\xrightarrow{D}$ " we mean convergence in distribution. Further assume conditions (C6) to (C8) below.

(C6) $n^{-1}\mathbf{m}^\top\mathbf{m} \xrightarrow{P} u$ where $u > 0$ is a constant.

(C7) $n^{-1}\mathbf{m}^\top\boldsymbol{\epsilon} \xrightarrow{P} 0$.

(C8) $n^{-1/2}\mathbf{m}^\top\boldsymbol{\epsilon} \xrightarrow{D} N(0, u\sigma^2)$.

Assumptions (C6)-(C8) are standard regularity assumptions to ensure the asymptotic results in Theorem 1 below. Given that $m \in \mathbb{H}(r, L)$ and $\mathbb{Z}_1, \ldots, \mathbb{Z}_n$ are independent, (C6) can be shown to be true via standard concentration arguments; see for example Lemma A.2 of Huang et al. (2004). (C7) and (C8) are true when the noises $\epsilon_1, \ldots, \epsilon_n$ are independently normally distributed. From (3) we have:

$$\hat{\beta} = \beta + (\hat{\mathbf{m}}^\top\hat{\mathbf{m}})^{-1}\hat{\mathbf{m}}^\top\boldsymbol{\epsilon} = \beta + (\hat{\mathbf{m}}^\top\hat{\mathbf{m}})^{-1}[(\hat{\mathbf{m}} - \mathbf{m})^\top\boldsymbol{\epsilon} + \mathbf{m}^\top\boldsymbol{\epsilon}].$$

Note that $n^{-1}\hat{\mathbf{m}}^\top\hat{\mathbf{m}} \to n^{-1}\mathbf{m}^\top\mathbf{m}$ and $n^{-1}(\hat{\mathbf{m}} - \mathbf{m})^\top\boldsymbol{\epsilon} \to 0$ in probability, $n^{-1/2}\mathbf{m}^\top\boldsymbol{\epsilon}$ converges in distribution to a normal distribution, and $n^{-1/2}(\hat{\mathbf{m}} - \mathbf{m})^\top\boldsymbol{\epsilon} = o_p(1)$. By Slutsky's Theorem we can conclude that $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normally distributed with variance $Var((\mathbf{m}^\top\mathbf{m})^{-1}n^{-1/2}\mathbf{m}^\top\boldsymbol{\epsilon})$, which converges to $u^{-1}\sigma^2$ under Conditions 6 and 8. We summarize our results below:

Theorem 1 *Under conditions (C1) to (C8), as* $n \to \infty$*, we have*

$$\hat{\beta} \xrightarrow{P} \beta,$$

and

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, u^{-1}\sigma^2).$$

2.2 Nonparametric Instrument Model Averaging: Low-dimensional IV

Now we consider the case with $d > 1$ and propose the nonparametric instrument model averaging (NIMA) method. As the function m in equation (2) is hard to estimate directly, we adopt model averaging method. The proposed approach here is to adopt least squares model averaging (Hansen, 2007) to compute the weighted average of the predicted value of X from a set of submodels in the first stage, and then use the predicted $\hat{X}$ to estimate β in the second stage.

At the first stage, consider d distinct submodels to predict X with each $Z_k,\ k = 1, \ldots, d$. Specifically, for the k^{th} submodel, we regress X on Z_k in a fully unspecified functional form:

$$X = m_k(Z_k) + e_k \tag{4}$$

where $e_k = X - m_k(Z_k) = X - E[X|Z_k]$ is the corresponding error term. In what follows when we say that condition (C4) holds, it should refer to that each $m_k, k = 1, \ldots, d$ should satisfy the Hölder condition in (C4). We may then adopt similar estimation methods as in the preceding subsection. In the k^{th} submodel, we still use a B-spline basis to approximate the function $m_k(Z_k)$, that is $m_k(Z_k) \approx \mathbf{B}(Z_k)^\top \boldsymbol{\theta}_k$, where $\mathbf{B}(Z_k) = (B_1(Z_k), \ldots, B_{J_n}(Z_k))^\top$ are the B-spline basis functions, and $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kJ_n})^\top$ is a vector of coefficients of the basis.

We denote $\mathbb{Z}_{ki}$ to be the ith observation of the kth IV, $\mathbf{Z}_k = (\mathbb{Z}_{k1}, \ldots, \mathbb{Z}_{kn})^\top, \mathbf{m}_k = (m_k(\mathbb{Z}_{k1}), \ldots, m_k(\mathbb{Z}_{kn}))^\top$, $\mathbf{B}_{ki} = (B_1(\mathbb{Z}_{ki}), \ldots, B_{J_n}(\mathbb{Z}_{ki}))^\top, \mathbf{B}_k = (\mathbf{B}_{k1}, \ldots, \mathbf{B}_{kn})^\top$ where $i = 1, \ldots, n$ and $k = 1, \ldots, d$.

Then for the k^{th} submodel, we have $\mathbf{m}_k \approx \mathbf{B}_k \boldsymbol{\theta}_k$. Applying the OLS leads to the sample estimator $\hat{\boldsymbol{\theta}}_k = (\mathbf{B}_k^\top \mathbf{B}_k)^{-1} \mathbf{B}_k^\top \mathbf{X}$. Next we compute the model-based predicted value as

$\hat{\mathbf{m}}_k = \mathbf{B}_k\hat{\boldsymbol{\theta}}_k$ for each submodel. Substituting $\hat{\boldsymbol{\theta}}_k$, we have

$$\hat{\mathbf{m}}_k = \mathbf{B}_k(\mathbf{B}_k^\top\mathbf{B}_k)^{-1}\mathbf{B}_k^\top\mathbf{X}$$

from which we can see the prediction for the ith subject is

$$\hat{m}_k(\mathbb{Z}_{ki}) = \mathbf{B}_{ki}^\top(\mathbf{B}_k^\top\mathbf{B}_k)^{-1}\mathbf{B}_k^\top\mathbf{X}, i = 1, \ldots, n.$$

To construct a model average estimator of X, we seek weights to minimise

$$f(\boldsymbol{\omega}) = E\left(E(X|Z_1, \ldots, Z_d) - \sum_{k=1}^{d}\omega_k m_k(Z_k)\right)^2,$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)^\top$ are the weights corresponding to the d models. Define the optimal weight to be $\boldsymbol{\omega}^* = (\omega_1^*, \ldots, \omega_d^*)^\top = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} f(\boldsymbol{\omega})$. Denote $u_{kl} = E[m_k(Z_k)m_l(Z_l)]$ for $k, l = 1, \ldots, d$, $\mathbb{M} = (m_1(Z_1), \ldots, m_d(Z_d))$, $\mathbf{U} = E[\mathbb{M}^\top\mathbb{M}] = (u_{kl})_{d\times d}$, $v_k = E[m_k(Z_k)X]$, and $\mathbf{v} = (v_1, \ldots, v_d)^\top$. It can be easily shown that

$$\boldsymbol{\omega}^* = \mathbf{U}^{-1}\mathbf{v}.$$

The optimal weights $\boldsymbol{\omega}^*$ can be estimated by minimising the empirical least square function $Q(\boldsymbol{\omega}) = \|\mathbf{X} - \sum_{k=1}^{d}\omega_k\hat{\mathbf{m}}_k\|^2$.

Denote $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_d), \hat{\mathbf{M}} = (\hat{\mathbf{m}}_1, \ldots, \hat{\mathbf{m}}_d), \hat{\boldsymbol{\omega}} = (\hat{\omega}_1, \ldots, \hat{\omega}_d)^\top$. Then the closed form solution for the estimated weights is

$$\hat{\boldsymbol{\omega}} = (\hat{\mathbf{M}}^\top\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}^\top\mathbf{X}. \tag{5}$$

To establish the consistency of the estimated weights, we need to modify condition (C6) given in the preceding section and impose an additional condition.

(C6') With probability tending to 1, $n^{-1}\mathbf{m}_k^\top\mathbf{m}_l \rightarrow u_{kl}$ for all $k, l = 1, \ldots, d$, and $\mathbf{U} = (u_{k,l})_{1\leq k,l\leq d}$ is positive definite such that there exists an infinitesimal $\rho_n \succ n^{-\frac{r}{3(2r+1)}}$ such that $\rho_n \leq \lambda_1(\mathbf{U}) \leq \lambda_d(\mathbf{U}) \leq \rho_n^{-1}$, where $\lambda_1(\mathbf{U})$ and $\lambda_d(\mathbf{U})$ are the smallest eigenvalue and the largest eigenvalue of $\mathbf{U}$, respectively.

(C9) $n^{-1}\mathbf{m}_k^\top \mathbf{X} \xrightarrow{P} v_k$ for $k = 1, \ldots, d$.

(C6') is a multivariate version of (C6), with an additional bounded constraint for $\mathbf{U}$ to ensure the estimability of $\mathbf{U}$. In particular, the infinitesimal ρ_n in the condition indirectly allows the true $m_k(\cdot)$ functions to converge to zero at a slow rate, corresponding to the case with weak instruments. The order of ρ_n is chosen to ensure that the smallest eigenvalues of $\mathbf{U}$ and $\mathbf{U}^{-1}$ has a higher order than the estimation error. By the weak law of large numbers, (C9) holds when $m_k \in \mathbb{H}(r, L)$ and $(\mathbb{Z}_1, X_1), \ldots, (\mathbb{Z}_n, X_n)$ are independent.

Theorem 2 *Under conditions (C1) to (C5), (C6') and (C9), as $n \to \infty$, we have*

$$\hat{\boldsymbol{\omega}} \xrightarrow{P} \boldsymbol{\omega}^*.$$

The proof is given in the supplementary material. In the literature of model averaging, justification of weight optimality is very necessary. Following Theorem 2, our weight estimates are consistent to the true yet unattainable optimal weights.

Using the estimated weight $\hat{\boldsymbol{\omega}}$, the model averaging prediction for $\mathbf{X}$ is

$$\hat{\mathbf{X}} = \hat{\mathbf{M}}\hat{\boldsymbol{\omega}} = \hat{\mathbf{M}}(\hat{\mathbf{M}}^\top \hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}^\top \mathbf{X}.$$

We then proceed to the second stage and regress $\mathbf{Y}$ on $\hat{\mathbf{X}}$ to obtain the final NIMA estimate $\hat{\beta}$ given by

$$\hat{\beta} = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^\top \mathbf{Y}. \tag{6}$$

To establish the consistency and asymptotic normality of the model average estimator, we need to assume the following conditions modified from the preceding subsection.

(C7') $n^{-1}\mathbf{m}_k^\top \boldsymbol{\epsilon} \xrightarrow{P} 0$ for $k = 1, \ldots, d$.

(C8') $n^{-1/2}\mathbf{m}_k^\top \boldsymbol{\epsilon} \xrightarrow{D} N(0, u_{kk}\sigma^2)$ for $k = 1, \ldots, d$.

(C7') and (C8') are direct extensions of (C7) and (C8) under the multivariate case.

Theorem 3 *Under conditions (C1) to (C5), (C6') to (C8') and (C9), as $n \to \infty$, we have*

$$\hat{\beta} \xrightarrow{P} \beta,$$

and

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, [\mathbf{v}^\top \mathbf{U}^{-1} \mathbf{v}]^{-1} \sigma^2).$$

Theorem 3 reduces to Theorem 1 when $d = 1$. Chamberlain (1987) and Chen et al. (2020) found combinations of instruments to improve the estimation efficiency of $\hat{\beta}$. Our paper focuses more on finding the "best" linear combinations of models (for predicting $\mathbf{X}$) based on different functional forms of IVs. Notice that the true model weights are given as $\omega^* = \mathbf{U}^{-1}\mathbf{v}$. Consequently, from Theorem 3 we have the asymptotic variance of $\sqrt{n}\hat{\beta}$ is $[\mathbf{v}^\top \mathbf{U}^{-1} \mathbf{v}]^{-1} \sigma^2 = [\omega^{*\top} \mathbf{U} \omega^*]^{-1} \sigma^2$. With some simple calculation, it can be shown that when an additional IV is served in 2SLS, the asymptotic variance of $\hat{\beta}$ becomes smaller, leading to a more efficient estimator. In the next subsection, we address the case where the number of IVs d is growing to infinity. To deal with the high dimensionality, a penalized approach is introduced to obtain a regularized weight estimator at the model averaging step.

2.3 Nonparametric Instrument Model Averaging: High-dimensional IV

In this subsection we extend the proposed NIMA method to the case where $d \to \infty$. It is known that the performance of the 2SLS method deteriorates drastically or becomes inapplicable as the dimension of instruments increases. We thus adopt a regularization step to cope with the high dimensionality at the model averaging stage in the NIMA procedure.

We use the same spline approximation method in the preceding section to obtain the d nonparametric models $\{\hat{\mathbf{m}}_k : k = 1, \cdots, d\}$. To obtain the weights at the model averaging stage, a regularized weight estimator $\hat{\boldsymbol{\omega}}_d$ is given by

$$\hat{\boldsymbol{\omega}}_d = \underset{\boldsymbol{\omega} \in \mathbb{R}^d}{argmin} \left\{ \frac{1}{2n} \|\mathbf{X} - \sum_{k=1}^{d} \omega_k \hat{\mathbf{m}}_k\|_2^2 + \sum_{k=1}^{d} p_\lambda(|\omega_k|) \right\} \tag{7}$$

where $p_\lambda(\cdot)$ is a sparsity-inducing penalty function to be specified below, and $\lambda > 0$ is a tuning parameter that controls the strength of the regularization.

We consider the following three common choices of the penalty function $p_\lambda(t)$ for $t \geq 0$: (a) the L_1 penalty or Lasso (Tibshirani, 1996), $p_\lambda(t) = \lambda t$; (b) the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001),

$$p_\lambda(t) = \lambda \int_0^t \{I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\} d\theta, \ a > 2;$$

and (c) the minimax concave penalty (MCP) (Zhang et al., 2010),

$$p_\lambda(t) = \int_0^t \frac{(a\lambda - \theta)_+}{a} d\theta, \ a > 1.$$

The SCAD and MCP penalties have an additional tuning parameter a to control the shape of the function. Based on the original literature, we choose $a = 3.7$ for SCAD penalty and $a = 3$ for MCP penalty. These penalty functions have been widely used in high-dimensional sparse modeling and their properties are well understood in ordinary regression models (e.g., Fan and Lv (2010)). Moreover, the fact that these penalties belong to the class of quadratic spline functions on $[0, \infty)$ allows for a closed-form solution to the corresponding penalized least squares problem in each coordinate, leading to very efficient implementation via coordinate descent (e.g., Mazumder et al. (2011)).

In this paper, we assume that the optimal weights for combining the marginal regressions are sparse with respect to the marginal regressions, i.e., most elements in $\boldsymbol{\omega}^*$ are zero.

Denote the index set for the nonzero weights as S and its complement as S^c, i.e., $\omega_i^* = 0$ iff $i \in S^c$. We assume that the cardinality of S is $|S| = p$. We remark that the sparsity assumption with exact zeros here can be further relaxed to $\sum_{i \in S^c} |\omega_i^*| = O\big(\big(\frac{\log(dn)}{n}\big)^{\frac{r}{2r+1}}\big)$. Under such an approximate sparse assumption, the convergence results provided in Theorem 4 below will still be valid; see the technical discussions after the proof of Theorem 4 in the supplementary material.

For any two index sets $S_1, S_2 \in \{1, \ldots, d\}$, and a $d \times d$ matrix $\mathbf{A}$, we shall use $\mathbf{A}_{S_1,S_2}$ to denote the corresponding submatrix with the rows indexed by S_1 and columns indexed by S_2. To establish the consistency of $\hat{\boldsymbol{\omega}}_d$, we make the following regularity assumptions:

(C5') We take $J_n = \Big\lfloor c\big(\frac{n}{\log(dn)}\big)^{1/(2r+1)}\Big\rfloor$ for some positive constant c, where $\lfloor t \rfloor$ is the largest integer no greater than t.

(C6") Let $\mathbf{U}_{S,S}$ be the sub-matrix of $\mathbf{U} = E(\mathbf{M}^\top \mathbf{M})$ with row and column indexed by S. We assume that $\mathbf{U}_{S,S}$ is positive definite such that there exists an infinitesimal $\rho_{d,n} \succ \big(\frac{\log(dn)}{n}\big)^{-\frac{r}{3(2r+1)}}$ such that $\rho_{d,n} \leq \lambda_1(\mathbf{U}_{S,S}) \leq \lambda_d(\mathbf{U}_{S,S}) \leq \rho_{d,n}^{-1}$, where $\lambda_1(\mathbf{U}_{S,S})$ and $\lambda_d(\mathbf{U}_{S,S})$ are the smallest eigenvalue and the largest eigenvalue of $\mathbf{U}_{S,S}$, respectively.

(C10) Let e_k be defined as in (4). We assume that for all $k = 1, \ldots, d$, and given Z_k, we have $E(e_k) = 0$, $Var(e_k) \leq \bar{\sigma}^2$ for some $\bar{\sigma}^2 < \infty$, and there exists a constant $a > 0$ such that $E|e_k|^t \leq \bar{\sigma}^2 t! a^{t-2}/2$ hold for all $t \geq 2$. Similarly, we assume that there exist positive constants a_ϵ and σ_ϵ^2 such that given X, $E|\epsilon|^t \leq \sigma_\epsilon^2 t! a_\epsilon^{t-2}/2$ hold for all $t \geq 2$.

(C11) There exists a constant $\kappa \in (0, 1)$ such that

$$\max_{i \in S^c} \|[\mathbf{M}^\top \mathbf{M}]_{i,S}([\mathbf{M}^\top \mathbf{M}]_{S,S})^{-1}\|_1 \leq 1 - \kappa.$$

Notice that different from the fixed dimensional case, we have to introduce an additional $\log(dn)$ term in the order of J_n in (C5') to balance the estimation accuracy under the high dimensional settings. (C6") is an analog to (C6') imposed for the variables in the active set S. The moment condition (C10) is imposed to control the tail behavior of the noises, and as a result, ensure the validity of classical concentration results (c.f. Bernstein's inequality in Lin and Bai (2011)) which plays an important role in establishing uniform convergence under high dimensionality. (C11) is an irrepresentability condition for model selection consistency; see for example Theorem 1 in Zou (2006) for further discussions.

Denote the ith element of $\hat{\boldsymbol{\omega}}_d$ as $\hat{\omega}_{d,i}$. The following theorem establishes the consistency of $\hat{\boldsymbol{\omega}}_d$ under the sparse case.

Theorem 4 *Suppose Assumptions (C1) to (C4), (C5'), (C6"), (C10) and (C11) hold, and assume that*

$$\rho_{n,d}^{-1} p \left(\frac{\log(dn)}{n}\right)^{\frac{r}{2r+1}} \to 0 \quad \text{and} \quad \rho_{n,d}^{-1} \left(\frac{\log(dn)}{n}\right)^{\frac{r}{2r+1}} \sqrt{p} \|\boldsymbol{\omega}^*\|_1 \to 0.$$

Let

$$\lambda = c_p \rho_{n,d}^{-1} \left(p \left(\frac{\log(dn)}{n}\right)^{\frac{r}{2r+1}} + \left(\frac{\log(dn)}{n}\right)^{\frac{r}{2r+1}} \sqrt{p} \|\boldsymbol{\omega}^*\|_1 \right),$$

for some large enough constant $c_p > 0$. We have:

(i) With probability tending to 1, $\hat{\omega}_{d,i} = 0$ for all $i \in S^c$.

(ii) $\|\hat{\boldsymbol{\omega}}_d - \boldsymbol{\omega}^*\|_2 = O_p\left(\rho_{n,d}^{-1} \sqrt{\frac{p \log(dn)}{n}} + \rho_{n,d}^{-1} \left(\frac{\log(dn)}{n}\right)^{\frac{r}{2r+1}} \sqrt{p} \|\boldsymbol{\omega}^*\|_1 \right)$.

In Theorem 4 we provide an explicit estimation error bound for the high dimensional case and that distinguishes from Theorem 2. We remark that the additional assumptions in Theorem 4 would hold under the fixed dimensional case, and hence the error bound derived in Theorem 4(ii) also holds for the fixed dimensional case. Theorem 4(i) indicates that the zero weights can be consistently identified, and Theorem 4(ii) yields an upper bound for the overall estimation error of the weights. Practically, the tuning parameter is usually chosen via the cross validation based on the squared loss. Without loss of generality, suppose we have selected $\bar{p}$ instrumental variables with nonzero weights: $Z_{d_1}, Z_{d_2}, \ldots, Z_{d_{\bar{p}}}$, $1 \leq d_1, d_2, \ldots, d_{\bar{p}} \leq d$. From Theorem 4 we have that with probability tending to 1, $\bar{p} \to p$. Following a similar idea of hybrid Lasso, we can further construct the least squared estimator using these selected instrumental variables, i.e.,

$$\hat{\boldsymbol{\omega}}_{\bar{p}} = \left(\mathbf{M}_{\bar{p}}^\top \mathbf{M}_{\bar{p}}\right)^{-1} \mathbf{M}_{\bar{p}}^\top \mathbf{X},$$

which is obtained by replacing $\hat{\mathbf{M}}$ by $\mathbf{M}_{\bar{p}} = (\hat{\mathbf{m}}_{d_1}, \ldots, \hat{\mathbf{m}}_{d_{\bar{p}}})$ in (5).

The model averaging prediction for $\mathbf{X}$ is then given by:

$$\hat{\mathbf{X}}_{\bar{p}} = \mathbf{M}_{\bar{p}} \hat{\boldsymbol{\omega}}_{\bar{p}} = \mathbf{M}_{\bar{p}} (\mathbf{M}_{\bar{p}}^\top \mathbf{M}_{\bar{p}})^{-1} \mathbf{M}_{\bar{p}}^\top \mathbf{X}.$$

We then proceed to the second stage and regress $\mathbf{Y}$ on $\hat{\mathbf{X}}_{\bar{p}}$ to obtain the final estimates $\hat{\beta}_{\bar{p}}$ given by

$$\hat{\beta}_{\bar{p}} = (\hat{\mathbf{X}}_{\bar{p}}^\top \hat{\mathbf{X}}_{\bar{p}})^{-1} \hat{\mathbf{X}}_{\bar{p}}^\top \mathbf{Y}. \tag{8}$$

The following theorem provides an upper bound for the estimation error of $\hat{\beta}_{\bar{p}}$:

Theorem 5 *Under the assumptions of Theorem 4, and assume that*

$$\rho_{n,d}^{-1} \sqrt{\frac{p \log(dn)}{n \|\boldsymbol{\omega}^*\|_2^2}} + \rho_{n,d}^{-1} \left(\frac{\log(dn)}{n}\right)^{\frac{r}{2r+1}} \frac{\sqrt{p}\|\boldsymbol{\omega}^*\|_1}{\|\boldsymbol{\omega}^*\|_2} \to 0. \tag{9}$$

Let $\hat{\beta}_{\bar{p}}$ be defined as in (8). *Given $\bar{p}$, we have:*

$$|\hat{\beta}_{\bar{p}} - \beta| = O_p\left(\rho_{n,d}^{-1} \sqrt{\frac{p \log(dn)}{n \|\boldsymbol{\omega}^*\|_2^2}}\right). \tag{10}$$

Condition (9) ensures that the signal strength $\|\boldsymbol{\omega}^*\|_2$ has a higher order than the estimation error in Theorem 4(ii). Interestingly, the estimation upper bound provided in (10) suggests that the estimation error of $\hat{\beta}_{\bar{p}}$ is proportional to $\|\boldsymbol{\omega}^*\|_2^{-1}$. Intuitively, a larger $\|\boldsymbol{\omega}^*\|_2$ could reflect the strengths of the combined instrumental variables, and hence lead to more accurate estimation of β. Assume that size of the active set p is a fixed constant. Similar to Theorem 3, given the selected instrumental variables, we can establish the consistency and asymptotic normality of the regularized model average estimator. Let $\mathbf{v}_{\bar{p}}$ and $\mathbf{U}_{\bar{p}}$ be defined as $\mathbf{v}$ and $\mathbf{U}$ in Conditions (C5) and (C6') respectively, with the instrumental variables replaced by $Z_{d_1}, \ldots, Z_{d_{\bar{p}}}$. We have the following asymptotic results.

Corollary 1 *Assume that the assumptions of Theorem 3 hold with the selected $\bar{p}$ instrumental variables $Z_{d_1}, \ldots, Z_{d_{\bar{p}}}$, we have, as $n \to \infty$,*

$$\sqrt{n}(\hat{\beta}_{\bar{p}} - \beta) \xrightarrow{D} N(0, [\mathbf{v}_{\bar{p}}^\top \mathbf{U}_{\bar{p}}^{-1} \mathbf{v}_{\bar{p}}]^{-1} \sigma^2).$$

3 Simulation

We next conduct simulation studies to examine the performance of the proposed nonparametric model averaging estimation method. In the following cases, Y is generated according to equation (1) with $\beta = 1$, and X is generated according to different nonparametric models given below. We generate the error terms $\binom{\epsilon}{e} \sim N(\binom{0}{0}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix})$. The IVs $(Z_1, \ldots, Z_d)$ are generated from $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a compound symmetric matrix with variance 1 and covariance ρ.

Case 1. Linear model ($d = 5$):

$$X = 0.08Z_1 + 0.06Z_2 + 0.05Z_3 + 0.08Z_4 + 0.08Z_5 + e$$

Case 2. Nonparametric additive model ($d = 5$):

$$\begin{aligned} X &= 0.08Z_1^3 e^{\sin(50Z_1)} + 0.06e^{Z_2}\cos(50Z_2) + 0.05Z_3^3 e^{Z_3} \\ &\quad +0.04(e^{-2Z_4} + e^{2Z_4}) + 0.08Z_5^3 e^{Z_5} + e \end{aligned}$$

Case 3. Nonparametric additive model with interaction terms ($d \geq 5$):

$$\begin{aligned} X &= 0.08Z_1^3 e^{\sin(50Z_1)} + 0.06e^{Z_2}\cos(50Z_2) + 0.05Z_3^3 e^{Z_3} + 0.04(e^{-2Z_4} + e^{2Z_4}) \\ &\quad +0.08Z_5^3 e^{Z_5} + 0.06Z_5^3\cos(50Z_1) + 0.05e^{2Z_1+\sin(50Z_2)} + e \end{aligned}$$

Case 4. Nonparametric additive model with interaction terms and high-dimensional covariates ($d > 5$):

$$\begin{aligned} X &= 0.08Z_1^3 e^{\sin(50Z_1)} + 0.06e^{Z_2}\cos(50Z_2) + 0.05Z_3^3 e^{Z_3} + 0.04(e^{-2Z_4} + e^{2Z_4}) \\ &\quad +0.08Z_5^3 e^{Z_5} + 0.06Z_5^3\cos(50Z_1) + 0.05e^{2Z_1+\sin(50Z_2)} + \frac{1}{d^2}\sum_{k=6}^{d} Z_k + e \end{aligned}$$

In all these cases, the strengths of the IVs are designed to be weak to moderate, relative to the order of the error magnitude. Such a specification may be more plausible in a real high-dimensional IV studies. For example, using single nucleotide polymorphism (SNP) as IVs for bio-medical research, it is commonly observed that massive number of SNPs only have weak association with the exposures. Comparisons are made with various competing methods. In Cases 1 and 2, we also considered the effects of applying different splines in the first stage, and examine whether our procedure is sensitive to these functional basis choices. In Case 3, we included the linear model averaging method of Martins and Gabriel (2014) for comparison. We considered 2 different ways of model screening for their approach, namely "trimming" and "split-sample" screenings, to reduce the number of candidate models for averaging. In terms of the criteria used to select models as well as to compute the weights of the candidate models, the relevant moment selection criteria (RMSC), model selection criteria (MSC) and generalised R-squared (GR^2) as defined in Martins and Gabriel (2014) are applied. The AIC and BIC penalty terms are used to compute the RMSC and MSC respectively in the comparison. This method, however, is infeasible for comparison when d is enormous, due to the huge number of candidate models. Instead, in Cases 3 and 4 with diverging d values, we compare the performance of our proposed model average estimator with that of 2SLS coupled with existing model selection method and linear method without model selection. We note that $d-5$ IVs in cases 3 and 4 only have very weak effects (equal to zero in case 3 while approaching to zero with increasing d in case 4) on X. To study the impacts of different penalty functions, we consider three familiar penalty functions: Lasso, Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP). In summary, the following methods are compared:

Naive: one-stage OLS regressing Y on X directly;

NIMA($\cdot$): the proposed nonparametric instrument model average method applying different splines in stage one, "$\cdot$" refers to the splines applied where "bs" is the B-spline, "ps" is the P-spline in Eilers et al. (1996) and "cr" is the cubic regression spline in Wood (2017);

$MA \cdot lin$: the parametric model average method using linear regression in stage one;

$RMSC_t$: the model average method in Martins and Gabriel (2014) using RMSC for model selection and weight computation, with "trimming" screening procedure;

MSC_t: the model average method in Martins and Gabriel (2014) using MSC for model selection and weight computation, with "trimming" screening procedure;

GR_t^2: the model average method in Martins and Gabriel (2014) using GR^2 for model selection and weight computation, with "trimming" screening procedure;

$RMSC_s$: the model average method in Martins and Gabriel (2014) using RMSC for model selection and weight computation, with "split-sample" screening procedure;

MSC_s: the model average method in Martins and Gabriel (2014) using MSC for model selection and weight computation, with “split-sample” screening procedure;

GR_s^2: the model average method in Martins and Gabriel (2014) using GR^2 for model selection and weight computation, with “split-sample” screening procedure;

NIMA.Lasso: the nonparametric instrument model average method using Lasso for IV selection in stage one;

NIMA.SCAD: the nonparametric instrument model average method using SCAD for IV selection in stage one;

NIMA.MCP: the nonparametric instrument model average method using MCP for IV selection in stage one;

$MS \cdot Lasso$: the parametric model average using the IVs selected by Lasso in stage one, implemented by `R` package `glmnet`;

We report the simulation results over 500 simulations for Cases 1 to 4. In all cases, the coverage of the 95% confidence interval based on the asymptotic distribution of our proposed model average estimators derived in the previous section is close to the nominal level, outperforming the naive least squares estimation method. Due to the endogeneity issue, the naive method performs badly in most cases, with large bias and low coverage probability.

In cases 1 and 2, different spline bases perform quite similarly and also provide satisfactory estimation results for the proposed nonparametric model averaging. Thus for all the following cases we only report results from B-splines. The results for case 1 are included in supplementary file. In both cases, we can see that the parametric model averaging with linear models perform quite well with small bias and satisfactory coverage, especially in case 1 where the true model is linear. However, in case 2 (Table 1), we notice that the estimated causal parameter may have a much larger standard errors for the linear method, comparing to our proposed NIMA method. Even if we increase the sample size, the efficiency loss of using parametric model averaging is still non-negligible. This traditional approach is thus not as favorable as our proposed nonparametric model averaging, when the true IV effects are nonlinear.

In case 3 we compare our NIMA method with more parametric MA methods in Table 2 when $d = 5$. In such a low-dimensional setting, we do not need the various regularization methods and can include all the IVs in the 2 stage regression. The $RMSC_t$, MSC_t, GR_t^2, $RMSC_s$, MSC_s and GR_s^2 methods all lead to very small estimation bias. Nonetheless, our method yields relatively smaller standard errors and the coverage rate of the 95% confidence interval is closer to the nominal level.

We then increase d to 50, 100 and 200 in case 3 and report the estimation results of NIMA with various penalty functions in Table 3. It seems with sample size fixed at $n = 500$, the estimation performance deteriorates as d increases. In all cases, MCP appears to be slightly better than the other two penalty methods, with a smaller bias and a slightly higher coverage. MCP also tends to selects smaller number of IVs than other methods.

We then examine the most difficult case 4 in Table 4 where large number of very weak IVs are present in the model. Our proposed nonparametric model averaging perform quite well, producing consistent estimates for the causal parameter. The parametric model averaging with linear model using only the first 5 IVs does not perform so well, with slightly larger bias and lower coverage rate. The model selection method using Lasso can also achieve consistency for the parameter estimates. However, the standard errors from model selection are always much greater than those from our proposed model averaging.

Table 1: *Results for Case 2 (with $\rho = 0, 0.3, 0.6, 0.9$ and $d = 5$) using different splines. Bias, standard deviation (sd), standard error (se) and coverage probability (CP) of the β_e estimates. True value $\beta = 1$.*

ρ	n		$Naive$	MA	$NIMA('bs')$	$NIMA('ps')$	$NIMA('cr')$
0	200	bias	0.214	0.009	0.023	0.022	0.024
		sd	0.111	0.192	0.077	0.076	0.078
		se	0.069	0.187	0.075	0.075	0.077
		CP	26.2%	95.8%	94.2%	94.2%	94.0%
	500	bias	0.179	0.014	0.008	0.008	0.009
		sd	0.075	0.132	0.045	0.045	0.048
		se	0.044	0.135	0.047	0.047	0.049
		CP	10.0%	96.2%	94.8%	94.8%	95.4%
	1000	bias	0.161	0.007	0.005	0.004	0.005
		sd	0.062	0.106	0.033	0.033	0.035
		se	0.031	0.105	0.033	0.033	0.035
		CP	4.8%	95.2%	95.4%	95.4%	95.2%
0.3	200	bias	0.207	0.025	0.020	0.020	0.021
		sd	0.114	0.171	0.076	0.076	0.079
		se	0.069	0.173	0.076	0.076	0.078
		CP	25.6%	94.6%	94.2%	94.4%	93.4%
	500	bias	0.173	0.010	0.008	0.007	0.008
		sd	0.074	0.117	0.047	0.047	0.049
		se	0.044	0.122	0.047	0.047	0.049
		CP	12.6%	96.2%	95.0%	95.0%	95.2%
	1000	bias	0.159	0.005	0.005	0.005	0.006
		sd	0.059	0.092	0.033	0.033	0.035
		se	0.031	0.093	0.033	0.033	0.035
		CP	6.0%	96.2%	95.2%	95.2%	94.6%
0.6	200	bias	0.187	0.008	0.019	0.018	0.018
		sd	0.115	0.174	0.078	0.078	0.078
		se	0.069	0.176	0.075	0.075	0.076
		CP	35.4%	95.2%	93.6%	94.0%	93.4%
	500	bias	0.156	0.008	0.003	0.003	0.003
		sd	0.079	0.121	0.048	0.048	0.051
		se	0.044	0.124	0.047	0.047	0.049
		CP	17.0%	96.2%	94.8%	94.8%	94.2%
	1000	bias	0.150	0.001	0.004	0.004	0.003
		sd	0.057	0.103	0.033	0.033	0.034
		se	0.031	0.091	0.033	0.033	0.035
		CP	5.2%	95.8%	95.0%	95.0%	95.2%
0.9	200	bias	0.177	0.027	0.016	0.016	0.015
		sd	0.110	0.195	0.072	0.072	0.073
		se	0.069	0.187	0.073	0.073	0.074
		CP	35.0%	94.8%	95.6%	95.6%	95.0%
	500	bias	0.143	0.015	0.007	0.007	0.008
		sd	0.077	0.144	0.046	0.046	0.047
		se	0.044	0.136	0.046	0.046	0.047
		CP	24.2%	96.0%	95.0%	95.0%	93.8%
	1000	bias	0.123	0.002	0.000	0.000	0.000
		sd	0.056	0.100	0.032	0.032	0.033
		se	0.031	0.099	0.032	0.032	0.034
		CP	14.2%	95.8%	95.4%	95.4%	95.8%

Table 2: *Results for Case 3 (with $\rho = 0, 0.3, 0.6, 0.9$ and $d = 5$). Bias, standard deviation (sd), standard error (se) and coverage probability (CP) of the β_e estimates. True value $\beta = 1$.*

ρ	n		$Naive$	$NIMA$	$RMSC_t$	MSC_t	GR_t^2	$RMSC_s$	MSC_s	GR_s^2
0	200	bias	0.175	0.026	0.026	0.026	0.026	0.034	0.043	0.045
		sd	0.091	0.070	0.179	0.174	0.173	0.383	0.328	0.329
		se	0.069	0.075	0.261	0.253	0.254	0.397	0.373	0.375
		CP	34.8%	94.6%	99.8%	99.8%	99.8%	99.6%	99.4%	99.4%
	500	bias	0.148	0.011	0.013	0.012	0.012	0.024	0.024	0.026
		sd	0.067	0.047	0.134	0.130	0.131	0.213	0.196	0.200
		se	0.044	0.047	0.192	0.187	0.188	0.275	0.262	0.264
		CP	18.0%	94.8%	99.4%	99.6%	99.6%	99.0%	99.4%	99.6%
	1000	bias	0.137	0.005	0.011	0.011	0.011	0.016	0.019	0.019
		sd	0.055	0.034	0.111	0.108	0.108	0.165	0.156	0.157
		se	0.031	0.033	0.148	0.144	0.145	0.210	0.202	0.203
		CP	8.8%	93.8%	99.2%	99.4%	99.2%	99.0%	99.0%	99.0%
0.3	200	bias	0.160	0.018	0.022	0.023	0.022	0.034	0.039	0.039
		sd	0.101	0.077	0.163	0.159	0.159	0.425	0.388	0.415
		se	0.070	0.076	0.233	0.228	0.229	0.375	0.359	0.360
		CP	39.8%	93.6%	99.2%	99.2%	99.4%	98.4%	98.4%	99.0%
	500	bias	0.145	0.008	0.008	0.009	0.009	0.025	0.028	0.030
		sd	0.067	0.047	0.112	0.121	0.113	0.196	0.180	0.184
		se	0.044	0.048	0.164	0.161	0.161	0.237	0.230	0.231
		CP	19.4%	95.2%	99.2%	99.2%	99.2%	99.6%	99.8%	99.8%
	1000	bias	0.129	0.004	0.003	0.004	0.004	0.002	0.003	0.004
		sd	0.050	0.033	0.088	0.086	0.086	0.144	0.134	0.136
		se	0.031	0.034	0.124	0.122	0.122	0.177	0.173	0.173
		CP	7.0%	94.4%	99.6%	99.6%	99.6%	98.8%	98.6%	98.6%
0.6	200	bias	0.150	0.012	0.017	0.017	0.017	0.046	0.054	0.059
		sd	0.096	0.076	0.170	0.164	0.163	0.323	0.291	0.306
		se	0.070	0.075	0.230	0.226	0.227	0.364	0.352	0.353
		CP	45.2%	95.2%	99.2%	99.2%	99.2%	99.8%	99.8%	99.8%
	500	bias	0.121	0.004	0.010	0.011	0.011	0.029	0.032	0.034
		sd	0.067	0.046	0.110	0.110	0.110	0.213	0.207	0.216
		se	0.044	0.047	0.160	0.158	0.158	0.239	0.234	0.234
		CP	33.0%	95.4%	99.6%	99.8%	99.8%	99.2%	99.6%	99.6%
	1000	bias	0.116	0.003	0.007	0.007	0.007	0.013	0.013	0.013
		sd	0.049	0.034	0.082	0.082	0.082	0.135	0.135	0.137
		se	0.031	0.033	0.121	0.119	0.120	0.175	0.172	0.172
		CP	13.0%	94.8%	99.6%	99.6%	99.6%	99.4%	99.4%	99.4%
0.9	200	bias	0.149	0.012	0.017	0.018	0.018	0.044	0.065	0.062
		sd	0.102	0.073	0.170	0.168	0.168	0.462	0.440	0.504
		se	0.070	0.073	0.248	0.245	0.246	0.427	0.415	0.413
		CP	47.2%	95.0%	99.6%	99.4%	99.4%	98.6%	98.6%	98.6%
	500	bias	0.111	0.002	0.001	0.001	0.001	0.006	0.015	0.017
		sd	0.063	0.045	0.132	0.132	0.132	0.285	0.281	0.290
		se	0.044	0.046	0.174	0.173	0.174	0.276	0.271	0.272
		CP	37.4%	94.8%	99.4%	99.4%	99.4%	99.2%	99.2%	99.4%
	1000	bias	0.098	0.001	0.001	0.001	0.001	0.012	0.014	0.015
		sd	0.051	0.034	0.095	0.095	0.095	0.166	0.164	0.165
		se	0.031	0.032	0.132	0.131	0.131	0.199	0.198	0.198
		CP	23.6%	94.0%	99.8%	99.8%	99.8%	99.0%	99.0%	98.8%

Table 3: *Results for Case 3 (with $\rho = 0$ and $n = 500$). Bias, standard deviation (sd), standard error (se) and coverage probability (CP) of the β_e estimates. Number selected (NS) is the number of instrumental variables Z_k which are selected, $5 < k \leq d$ under high dimension case. True value $\beta = 1$.*

d		$Naive$	$NIMA.Lasso$	$NIMA.SCAD$	$NIMA.MCP$
50	bias	0.150	0.025	0.017	0.015
	sd	0.069	0.047	0.046	0.046
	se	0.044	0.047	0.047	0.047
	CP	20.4%	92.8%	94.6%	95.8%
	NS		27.8	7.4	5.7
100	bias	0.146	0.035	0.019	0.015
	sd	0.069	0.048	0.048	0.047
	se	0.044	0.046	0.046	0.047
	CP	18.0%	87.2%	91.8%	92.6%
	NS		44.9	11.4	8.0
200	bias	0.147	0.055	0.027	0.021
	sd	0.068	0.052	0.051	0.052
	se	0.044	0.046	0.046	0.046
	CP	18.0%	74.8%	89.4%	89.8%
	NS		78	17.2	11.0

Table 4: *Results for Case 4 (with $\rho = 0, 0.3, 0.8$ and $n = 500$). Bias, standard deviation (sd), standard error (se) and coverage probability (CP) of the β_e estimates. Number selected (NS) is the number of instrumental variables Z_k which are selected, $5 < k \leq d$ under high dimension case. True value $\beta = 1$.*

				NIMA				
ρ	d		*Naive*	*Lasso*	*SCAD*	*MCP*	*MA*	$MS \cdot Lasso$
0	50	bias	0.154	0.027	0.018	0.017	0.068	0.045
		sd	0.067	0.045	0.045	0.045	0.097	0.118
		se	0.044	0.047	0.047	0.047	0.097	0.116
		CP	15.8%	92.4%	94.2%	95.0%	88.2%	93.8%
		NS		24.9	6.1	4.7		5.8
	100	bias	0.152	0.035	0.018	0.014	0.095	0.056
		sd	0.071	0.050	0.050	0.050	0.087	0.117
		se	0.044	0.046	0.047	0.047	0.081	0.109
		CP	17.0%	85.8%	90.4%	91.6%	74.2%	90.0%
		NS		38.3	9.6	6.8		8.3
	200	bias	0.153	0.058	0.030	0.022	0.127	0.068
		sd	0.070	0.053	0.052	0.050	0.079	0.119
		se	0.044	0.046	0.046	0.047	0.064	0.106
		CP	18.2%	70.8%	85.2%	90.4%	49.6%	85.4%
		NS		68.1	16.2	9.9		9.4
0.3	50	bias	0.136	0.015	0.012	0.012	0.052	0.022
		sd	0.065	0.046	0.046	0.046	0.092	0.114
		se	0.044	0.047	0.047	0.047	0.090	0.105
		CP	22.8%	93.4%	94.4%	94.8%	90.2%	91.4%
		NS		28.5	16.5	13.5		6.6
	100	bias	0.139	0.027	0.020	0.019	0.082	0.029
		sd	0.066	0.048	0.047	0.047	0.079	0.112
		se	0.044	0.046	0.047	0.047	0.077	0.103
		CP	21.6%	91.4%	93.0%	93.2%	78.2%	92.8%
		NS		44.1	25.1	18.7		9.1
	200	bias	0.140	0.039	0.023	0.020	0.112	0.035
		sd	0.067	0.051	0.050	0.049	0.071	0.110
		se	0.044	0.046	0.047	0.047	0.063	0.099
		CP	21.0%	83.6%	89.6%	91.2%	57.0%	90.2%
		NS		67.4	31.0	21.1		11.1
0.8	50	bias	0.114	0.009	0.008	0.008	0.043	0.012
		sd	0.068	0.047	0.047	0.047	0.093	0.121
		se	0.044	0.046	0.046	0.046	0.092	0.112
		CP	34.2%	93.6%	93.6%	93.8%	91.2%	94.6%
		NS		18.6	10.9	10.1		6.4
	100	bias	0.121	0.014	0.011	0.011	0.074	0.016
		sd	0.063	0.046	0.046	0.046	0.082	0.111
		se	0.044	0.046	0.046	0.046	0.078	0.110
		CP	30.8%	93.6%	93.8%	93.6%	81.8%	94.6%
		NS		30.2	15.7	14.0		8.0
	200	bias	0.117	0.019	0.014	0.013	0.090	0.019
		sd	0.065	0.047	0.046	0.046	0.074	0.128
		se	0.044	0.045	0.045	0.046	0.063	0.113
		CP	33.6%	91.8%	93.2%	94.4%	67.4%	93.8%
		NS		44.0	19.1	17.9		9.5

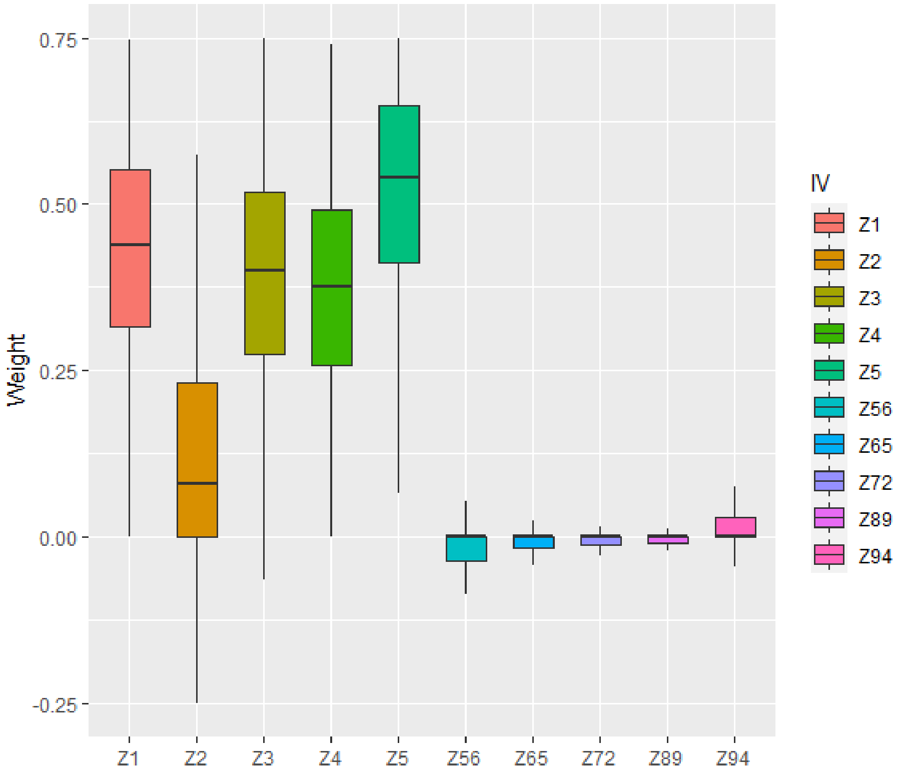

Figure 1: The boxplot of partial weights corresponding to the IVs in Case 4

An efficiency gain is again observed for this high-dimensional case.

We plotted the estimated weights for case 4 in Figure 1. In case 4, only the first 5 IVs are indeed associated with the exposure X with non-degenerating effects while all other IVs have close-to-zero effects. We randomly sample 5 IVs from $\{Z_k : 5 < k \leq d\}$ and report their weights also in Figure 1. We can see that these estimated weights are distributed in a narrow neighborhood of zero. The estimated weights from our NIMA procedure are indeed very sensible for this example.

4 Home Price Data Analysis

To illustrate the proposed nonparametric model averaging method, we consider a real case study of the effect of federal appellate court decisions regarding eminent domain on home prices. The data has recently been analyzed in Belloni et al. (2012) and Seng and Li (2022). Since this data also contains many confounders, we estimate the structural model of the form $Y = X\beta + \mathbf{X}_o^\top \boldsymbol{\beta}_o + \epsilon$ where Y denotes the log of Case-Shiller home price index, X denotes the number of pro-plaintiff appellate takings decisions, and $\mathbf{X}_o$ include whether there was any cases in that circuit-year, number of takings appellate decisions, controls for the distribution of characteristics of federal circuit court judges in a given circuit-year, circuit-specific effects, time-specific effects, and circuit-specific time trends. An appellate court decision is pro-plaintiff if the court ruled that a taking was unlawful and overturned the government?s seizure of the property in favor of the private owner. The number of pro-plaintiff decisions is thus a proxy indication of how protective a regime is of individual property rights. We are interested to estimate β which represents the effect of an additional decision upholding individual property rights on the home price.

The analysis of the effect of takings law is complicated by the possible endogeneity between takings decisions and home price as suggested by Chen and Yeh (2012) and Belloni et al. (2014, 2012); Seng and Li (2022). The instrumental variables strategy relies on the random assignment of judges to federal appellate panels which renders the exact identity of the judges and their demographics as potential instruments since they are also randomly assigned conditional on the distribution of characteristics of federal circuit court judges in a given circuit-year. We obtain a total of $d = 147$ potential instruments in this data analysis. The exogenous variables are included only in stage two of our analysis. The total sample size is $n = 183$. The data for Y, X, $\mathbb{Z}$ and $\mathbf{X}_o$ are all standardized to have mean zero and standard deviation one.

We now consider implementing our proposed methods and comparing with other models examined in the simulation studies. For the proposed nonparametric model averaging method, we use Lasso, SCAD and MCP penalties to regularize the high-dimensional model weights. Specifically, the following methods are compared.

$pLasso$: 2SLS using the IVs selected by Lasso in stage one, implemented by `R` package `glmnet`;

pEL_a: 2SLS using the IVs selected by elastic net in stage one with elastic net mixing parameter a, implemented by `R` package `glmnet`;

pAL: 2SLS using the IVs selected by adaptive Lasso in stage one, implemented by `R` package `glmnet`;

$pSCAD$: 2SLS using the IVs selected by SCAD in stage one, implemented by `R` package `ncvreg`;

$pMCP$: 2SLS using the IVs selected by MCP in stage one, implemented by `R` package `ncvreg`;

SIS: 2SLS using the IVs selected by sure independent screening (SIS) method (Fan and Lv, 2008) in stage one, implemented by `R` package `SIS`;

npLasso: 2SLS using the IVs selected by Lasso with generalized additive model in stage one, implemented by `R` package `mgcv`;

npSCAD: 2SLS using the IVs selected by SCAD with generalized additive model in stage one, implemented by `R` package `mgcv`;

npMCP: 2SLS using the IVs selected by MCP with generalized additive model in stage one, implemented by `R` package `mgcv`;

In addition to these approaches we have also tried implementing other `R` packages including `sam`, `hgm`, `npreg` which are all designed for high-dimensional variable selection. Unfortunately none of these packages can generate selection results for this data and therefore is not reported in this paper.

In Table 5, we report the estimation results for two different cases. In Case I, we estimate the effect of takings law without adjusting for the other confounding covariates. Sometimes such a marginal association study may be of interest to practitioners. The naive method produces positive estimate but is not significant. In contrast, our proposed model average method and all existing model selection methods give negative yet insignificant estimates. After adjusting for the confounding covariates in Case II, we notice that the estimated effect of takings law become positive for all methods. Our penalized nonparametric model averaging method produce relatively stronger effect estimates than most other methods. In particular, the estimator from NIMA with a Lasso penalty is significant at 0.05 level. These findings are also consistent with the earlier findings in Belloni et al. (2012), Belloni et al. (2014) and Seng and Li (2022) where parametric IV regression was conducted. In contrast, none of the other model selection methods examined in our analysis produce statistically significant results.

Table 5: *Results of estimation for the Case-Shiller home price data. $\hat{\beta}_e$ is the estimated coefficient for X_e and SE is the estimated standard error of the estimator.*

	CASE I: Unadjusted estimation			CASE II: Adjusted estimation		
Methods	$\hat{\beta}_e$	SE	$p-value$	$\hat{\beta}_e$	SE	$p-value$
NIMA.$Lasso$	-0.125	0.047	0.008	0.038	0.016	0.018
NINA.$SCAD$	-0.143	0.108	0.183	0.033	0.037	0.372
NIMA.MCP	-0.137	0.113	0.224	0.042	0.039	0.285
$Naive$	0.027	0.074	0.720	0.049	0.030	0.103
$pLasso$	-0.155	0.109	0.154	0.037	0.038	0.320
$pEL_{0.5}$	-0.140	0.111	0.207	0.027	0.038	0.486
$pEL_{\frac{1}{3}}$	-0.133	0.108	0.220	0.036	0.037	0.340
pAL	-0.159	0.112	0.154	0.034	0.039	0.384
$pSCAD$	-0.133	0.110	0.227	0.049	0.038	0.193
$pMCP$	-0.129	0.114	0.255	0.050	0.039	0.204
$EL_{0.5}$	-0.263	0.108	0.015	0.052	0.037	0.168
$EL_{\frac{1}{3}}$	-0.295	0.104	0.005	0.047	0.036	0.188
AL	-0.169	0.112	0.130	0.049	0.039	0.204
SIS	-0.129	0.114	0.257	0.050	0.048	0.301
$npLasso$	-0.081	0.102	0.432	0.040	0.042	0.345
$npSCAD$	-0.086	0.105	0.416	0.042	0.043	0.326
$npMCP$	-0.075	0.106	0.480	0.043	0.043	0.327

The estimated weights from our nonparametric model averaging and the corresponding nonparametric models for the IVs are plotted in Figures 1 and 2 of the supplementary file. These fitted models suggest the IV effects may be quite nonlinear in this case as the curves differ substantially from a straight line. Lasso assigns non-zero weights to 6 variables $\{Z_1, Z_8, Z_{18}, Z_{82}, Z_{88}, Z_{100}\}$ in the data set while SCAD and MCP select $\{Z_1, Z_{82}, Z_{88}\}$ and

$\{Z_1\}$, respectively.

SUPPLEMENTARY MATERIAL

Technical proofs and additional numerical results are included in the supplementary material.

Acknowledgements

The authors thank the Editor, the Associate Editor and the anonymous referee for very useful comments that improved the presentation of the paper. This work was partially supported by the MOE Tier 1 Academic Research Funds A-8000016-00-00 in Singapore, and the National Natural Science Foundation Grant for Young Scholars 12001459.

References

Angrist, J. D. and A. B. Krueger (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics 13*(2), 225–235.

Austin, P. C. (2011). A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behavioral Research 46*(1), 119–151.

Austin, P. C. and E. A. Stuart (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine 34*(28), 3661–3679.

Baiocchi, M., D. S. Small, S. Lorch, and P. R. Rosenbaum (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association 105*, 1285–1296.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives 28*(2), 29–50.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York: Wiley.

Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association 90*, 443–450.

Burgess, S., V. Zuber, A. Gkatzionis, and C. N. Foley (2018). Modal-based estimation via heterogeneity- penalized weighting: model averaging for consistent and efficient estimation in mendelian randomization when a plurality of candidate instruments are valid. *International Journal of Epidemiology*, 1242–1254.

Caner, M. and Q. Fan (2010). The adaptive lasso method for instrumental variable selection. Technical report, North Carolina State University.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–334.

Chen, D. and S. Yeh (2012). Growth under the shadow of expropriation? the economic impacts of eminent domain.

Chen, J., D. L. Chen, and G. Lewis (2020). Mostly harmless machine learning: learning optimal instruments in linear iv models. *arXiv preprint arXiv:2011.06158*.

Clyde, M. A., J. Ghosh, and M. L. Littman (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics 20*(1), 80–101.

Didelez, V. and N. Sheehan (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research 16*, 309–330.

Donald, S. G. and W. K. Newey (2001). Choosing the number of instruments. *Econometrica 69*(5), 1161–1191.

Eilers, P. H., B. D. Marx, et al. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science 11*(2), 89–121.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and Y. Liao (2014). Endogeneity in high dimensions. *The Annals of Statistics 42*, 872–917.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B 70*(5), 849–911.

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*(1), 101.

Fang, F., J. Li, and X. Xia (2022). Semiparametric model averaging prediction for dichotomous response. *Journal of Econometrics 229*, 219–245.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 722–729.

Guo, Z., H. Kang, T. Tony Cai, and D. S. Small (2018). Testing endogeneity with high dimensional covariates. *Journal of Econometrics 207*(1), 175–187.

Hall, A. R., A. Inoue, K. Jana, and C. Shin (2007). Information in generalized method of moments estimation and entropy-based moment selection. *Journal of Econometrics 138*(2), 488–512.

Hall, A. R. and F. P. M. Peixe (2003). A consistent method for the selection of relevant instruments. *Econometric Reviews 22*(3), 269–287.

Hall, A. R., G. D. Rudebusch, and D. W. Wilcox (1996). Judging instrument relevance in instrumental variables estimation. *International Economic Review 37*(2), 283–298.

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

Hansen, C., J. Hausman, and W. Newey (2008). Estimation with many instrumental variables. *Journal of Business and Economic Statistics 26*, 398–422.

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

Huang, J. Z., C. O. Wu, and L. Zhou (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 763–788.

Huang, T. and J. Li (2018). Semiparametric model average prediction in panel data analysis. *Journal of Nonparametric Statistics 30*, 125–144.

Imai, K., D. Tingley, and T. Yamamoto (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 176*(1), 5–51.

Kang, H., A. Zhang, T. Cai, and D. Small (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association 111*, 132–144.

Keele, L. and J. W. Morgan (2016). How Strong is Strong Enough? Strengthening Instruments Through Matching and Weak Instrument Tests. *Annals of Applied Statistics 10*, 1086–1106.

Kline, R. B. (1998). *Principles and Practice of Structural equation modeling.* New York: The Guilford Press.

Koop, G. and D. Korobilis (2012). Forecasting inflation using dynamic model averaging. *International Economic Review 53*(3), 867–886.

Koop, G., R. Leon-Gonzalez, and R. Strachan (2012). Bayesian model averaging in the instrumental variable regression model. *Journal of Econometrics 171*(2), 237–250.

Lawlor, D. A., R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. Davey Smith (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine 27*, 1133–1163.

Li, D., O. Linton, and Z. Lu (2015). A flexible semiparametric forecasting model for time series. *Journal of Econometrics 187*, 345–357.

Li, J., J. Lv, A. Wan, and J. Liao (2022). Adaboost semiparametric model averaging prediction for multiple categories. *Journal of the American Statistical Association 117*, 495–509.

Li, J., X. Xia, W. K. Wong, and D. Nott (2018). Varying-coefficient semiparametric model averaging prediction. *Biometrics 74(4)*, 1417–1426.

Lin, W., R. Feng, and H. Li (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association 110*(509), 270–288.

Lin, Z. and Z. Bai (2011). *Probability inequalities*. Springer Science & Business Media.

Martins, L. F. and V. J. Gabriel (2014). Linear instrumental variables model averaging estimation. *Computational Statistics and Data Analysis 71*, 709–724.

Mazumder, R., J. H. Friedman, and T. Hastie (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association 106*(495), 1125–1138.

McCaffrey, D. F., B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine 32*(19), 3388–3414.

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

Nelson, C. R. and R. Startz (1990). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *The Journal of Business 63*(1), S125–S140.

Schumaker, L. (2007). *Spline functions: basic theory*. Cambridge University Press.

Seng, L. and J. Li (2022). Structural equation model averaging: Methodology and application. *Journal of Business and Economic Statisticss 40*, 815–828.

Shea, J. (1996, March). Instrument relevance in multivariate linear models: A simple measure. Working Paper 193, National Bureau of Economic Research.

Sloughter, J. M., T. Gneiting, and A. E. Raftery (2010). Probabilistic wind speed forecasting using ensembles and bayesian model averaging. *Journal of the American Statistical Association 105*(489), 25–35.

Small, D. S. and P. R. Rosenbaum (2008). War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association 103*, 924–933.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B 58*(1), 267–288.

Wang, Y., F. Ma, Y. Wei, and C. Wu (2016). Forecasting realized volatility in a changing world: A dynamic model averaging approach. *Journal of Banking & Finance 64*, 136–149.

Wehby, G. L., R. L. Ohsfeldt, and J. C. Murray (2008). 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. *Statistics in Medicine 27*, 2745–2749.

Wood, S. N. (2017). *Generalized additive models: an introduction with R.* CRC press.

Yang, Y. (2003, July). Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica 13*(3), 783–809.

Yu, D., H. Wang, P. Chen, and Z. Wei (2014). Mixed pooling for convolutional neural networks. In *International Conference on Rough Sets and Knowledge Technology*, pp. 364–375. Springer.

Zeugner, S., M. Feldkircher, et al. (2015). Bayesian model averaging employing fixed and flexible priors: The bms package for r. *Journal of Statistical Software 68*(4), 1–37.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Zubizarreta, J. R., D. S. Small, N. K. Goyal, S. Lorch, and P. R. Rosenbaum (2013). Stronger Instruments via Integer Programming in An Observational Study of Late Preterm Birth Outcomes. *Annals of Applied Statistics 7*, 25–50.