Model Selection and Model Averaging in Nonparametric Instrumental Variables Models*

Chu-An Liu† and Jing Tao‡

February 3, 2016

Abstract

This paper considers the problem of choosing the regularization parameter and the smoothing parameter in nonparametric instrumental variables estimation. We propose a simple Mallows' C_p-type criterion to select these two parameters simultaneously. We show that the proposed selection criterion is optimal in the sense that the selected estimate asymptotically achieves the lowest possible mean squared error among all candidates. To account for model uncertainty, we introduce a new model averaging estimator for nonparametric instrumental variables regressions. We propose a Mallows criterion for the weight selection and demonstrate its asymptotic optimality. Monte Carlo simulations show that both selection and averaging methods generally achieve lower root mean squared error than other existing methods. The proposed methods are applied to two empirical examples, the effect of class size question and Engel curve.

Keywords: Ill-posed inverse problem, Mallows criterion, Model averaging, Model selection, Nonparametric instrumental variables, Series estimation

JEL Classification: C14, C26, C52.

*We are indebted to Bruce Hansen for encouragement and valuable comments. We thank Jack Porter and Xiaoxia Shi for constructive comments and suggestions, and Xiaohong Chen for providing relevant results and helpful comments. We also thank Joachim Freyberger, Arthur Lewbel, Tatsushi Oka, Jeffrey Racine, Denis Tkachenko, and conference participants of CMES 2013, SETA 2014, TER 2015, and ESWC 2015 for their discussions and suggestions. All errors remain the authors'.

†Institute of Economics, Academia Sinica. Email: caliu@econ.sinica.edu.tw.

‡Department of Economics, University of Washington. Email: jingtao@uw.edu.

1 Introduction

Empirical research in economics is often concerned with estimation of causal relations between variables. In many applications, some of the regressors are endogenous, and hence the linear instrumental variables (IV) methods are widely used to identify and estimate the structural effects of interest. The linear parametric model, however, imposes strong assumptions about the population model structure that could be potentially misspecified. It is natural to generalize the linear instrumental variables approach to a flexible nonparametric framework. An important challenge of the empirical implementation of nonparametric instrumental variables methods is the selection of the regularization parameter and the smoothing parameter. This paper proposes a simple Mallows' C_p-type criterion to simultaneously select the regularization parameter and the smoothing parameter and presents a theoretical justification.

This paper deals with the nonparametric instrumental variables model

$$y_i = g(x_i) + e_i, \tag{1.1}$$

$$\mathrm{E}(e_i|z_i) = 0, \tag{1.2}$$

where y_i is a scalar dependent variable, g is an unknown structural function of interest, x_i is a vector of potentially endogenous variables, z_i is a vector of instruments, and e_i is an unobserved random variable. To recover nonlinearities from conditional expectations, it creates an ill-posed inverse problem. That is, the solution of g is not continuous in the reduced form estimators. Consequently, a consistent estimator of g need not result from replacing unknown population quantities with consistent estimators. To achieve consistency, one needs to regularize the mapping from reduced form estimators to the structural function g. Regularization is controlled by the regularization parameter that makes the mapping continuous.

In this paper we propose a novel criterion for the selection of the regularization parameter and the smoothing parameter in sieve or series estimators, where the regularization parameter is the number of terms in the linear approximation to g and the smoothing parameter is the number of series expansion terms for instruments. The proposed criterion is a simple Mallows' C_p-type criterion, which is an estimate of the mean squared error. Thus, it aims to balance the model fit and model complexity by elaborating the number of series expansion terms. One attractive advantage of the proposed criterion is computational ease.

The question we consider in this paper is more complicated than the model selection in nonparametric series regression models. We first need to deal with endogenous variables and then simultaneously choose the regularization parameter and the smoothing parameter. To tackle the difficulty, we follow Chen and Christensen (2013) and Hansen (2015) to develop bounds on the estimated matrices with endogenous variables. We then introduce a nonlinear penalty term to account for the interaction effect between the regularization parameter and the smoothing parameter. We show that this method is asymptotically optimal in the sense of achieving the lowest possible mean squared error among all candidates. Our contributions to the literature on model selection are two-fold. First, we extend the asymptotic optimality in Li (1987) to allow for possibly endoge-

nous variables. Second, we generalize the results of Donald and Newey (2001) to a nonparametric structural function of interest.

As an alternative to model selection, a model averaging estimator considers the uncertainty across different models as well as the model bias from each candidate model. In most cases, the model averaging estimator achieves lower risk and performs better than model selection estimators in finite samples. In this paper, we introduce a new nonparametric IV model averaging estimator and propose a Mallows criterion for the weight selection. To the best of our knowledge, this is the first work that considers the model averaging in nonparametric instrumental variables models. We demonstrate the asymptotic optimality of the proposed averaging estimator and provide some numerical evidence that the model averaging estimator performs relatively better than the model selection estimator.

The proposed model selection criterion and model averaging criterion depend on unknown population parameters. In practice, we replace the unknown parameters by the sample estimates. We compare the finite sample performance of proposed model selection and model averaging estimators with other existing methods. Simulation studies show that the proposed methods with plug-in estimators generally produce the lower root mean squared error as compared to other data-driven selection criteria for different sample sizes and degrees of endogeneity. As an empirical illustration, we consider the estimation of the effect of class size on students' performance and the estimation of an Engel curve for food. We find that our estimates are robust to the choice of basis functions, while other estimates are sensitive to the choice of basis functions.

We now discuss the related literature. There is a growing body of literature on nonparametric instrumental variables models; see Horowitz (2011) for a literature review. The two main nonparametric IV approaches are sieve or series estimators and kernel estimators. The sieve or series estimator has been developed by Newey and Powell (2003), Ai and Chen (2003), Blundell, Chen, and Kristensen (2007), Horowitz (2011), Horowitz (2012) and Newey (2013), while the kernel estimator has been developed by Hall and Horowitz (2005), Darolles, Fan, Florens, and Renault (2011), and Gagliardini and Scaillet (2012b). Chen and Pouzo (2012) study the nonparametric estimation in a large class of conditional moment restriction models with possible nonsmooth moments. Most of these studies, however, do not provide theoretically justified empirical methods for the regularization parameter and the smoothing parameter selection.

There is a large literature on model selection for regression models; see Claeskens and Hjort (2008) for a literature review. The approach we use in this paper is closely related to that of choosing the number of series expansion terms in regression models. Shibata (1980, 1981) demonstrates that model selection estimators based on the Akaike information criterion or the final prediction criterion achieve asymptotic optimality in homoskedastic autoregressive models. Ing and Wei (2003, 2005) extend Shibata's results for same-realization predictions. Li (1987) demonstrates the asymptotic optimality of the Mallows criterion, cross-validation, and generalized cross-validation in homoskedastic linear regression models. Andrews (1991) extends Li's results to the heteroskedastic linear regression models. Shao (1997) provides a general framework to discuss the asymptotic optimality of various model selection procedures. In this paper, we extend the existing literature

on the asymptotic optimality of model selection to regression models in the presence of endogenous variables.

Our paper is also related to the literature on instrumental variables selection. Donald and Newey (2001) and Donald, Imbens, and Newey (2009) consider the instrumental variables selection problem under the assumption that all instruments are valid. They choose instruments to minimize the higher-order asymptotic mean squared error. Andrews (1999), Andrews and Lu (2001), and Hong, Preston, and Shum (2003) investigate the problem of searching for the largest set of valid instruments. Okui (2011) proposes a shrinkage method for instrumental variable estimation. Carrasco (2012) introduces modified instrumental variable estimators based on regularizing the covariance matrix of instruments. Belloni, Chen, Chernozhukov, and Hansen (2012) develop Lasso and post-Lasso methods for constructing the optimal instruments in linear instrumental variables models.

In the model averaging literature, Hansen (2007) introduces the Mallows model averaging estimator and demonstrates its asymptotic optimality for nested and homoskedastic linear regression models. Wan, Zhang, and Zou (2010) extend Hansen's optimal result to the case of continuous weights and non-nested models. Kuersteiner and Okui (2010) propose model averaging criteria to construct the optimal instruments for linear instrumental variables estimation. Hansen and Racine (2012) propose the jackknife model averaging estimator and demonstrate the asymptotic optimality in heteroskedastic linear regression models. Zhang, Wan, and Zou (2013) generalize Hansen and Racine's results to linear regression models with lagged dependent variables. Liu and Okui (2013) propose the Heteroskedasticity-Robust C_p estimator and demonstrate its optimality in the linear regression models with heteroskedastic errors. To our knowledge, the averaging estimator has not been explored before in nonparametric IV models.

The existing literature on model selection in nonparametric instrumental variables models is comparatively small. Sueishi (2012) develops a model selection criterion based on a loss function spanned by instruments. Horowitz (2014) considers a modified nonparametric IV estimator, which uses the same number of series terms for regressors and instruments, and proposes an adaptive procedure for the regularization parameter selection. Breunig and Johannes (2015) propose an adaptive estimator for a linear functional of the structural function in nonparametric IV models. Centorrino (2015) develops a cross-validation criterion for the regularization parameter for kernel nonparametric instrumental variables estimators. A paper written concurrently with ours, Chen and Christensen (2015), establishes the optimal sup-norm convergence rates for spline and wavelet nonparametric IV estimators, and proposes a sup-norm adaptive Lepski-type procedure for choosing the regularization parameter.

The rest of the paper is organized as follows. Section 2 presents the nonparametric instrumental variables model, the approximating models, and the sieve nonparametric IV estimators. Section 3 introduces a model selection criterion for nonparametric IV models and provides an asymptotic optimality theory. Section 4 introduces a nonparametric IV model averaging estimator and presents the optimality theory for the averaging estimator. Section 5 presents the results of Monte Carlo experiments. Section 6 presents the empirical applications, and Section 7 concludes the paper.

Proofs and figures are presented in the Appendix.

Notation: For a $k \times k$ matrix A, $\sigma_{\max}(A)$ denotes its largest singular value, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote its largest and smallest eigenvalues, respectively, and A^- denotes its Moore-Penrose generalized inverse. Let $\|\cdot\|$ denote the Euclidean norm as $\|a\| = (tr(a'a))^{1/2}$ for vectors and the spectral norm as $\|A\| = (\lambda_{\max}(A'A))^{1/2}$ for matrices. Let $\|A\|_F = (tr(A'A))^{1/2}$ denote the Frobenius norm.

2 Nonparametric IV Model and Sieve Approximation

The model that we consider is

$$y_i = g(x_i) + e_i, \tag{2.1}$$

$$\mathrm{E}(e_i|z_i) = 0, \tag{2.2}$$

$$\mathrm{E}(e_i^2|z_i) = \sigma^2, \tag{2.3}$$

for $i = 1, ..., n$, where $y_i \in \mathbb{R}$ is a scalar dependent variable, g is an unknown structural function, $x_i \in \mathbb{R}^{d_x}$ is a vector of explanatory variables that may be correlated with e_i, $z_i \in \mathbb{R}^{d_z}$ is a vector of instruments, and e_i is an unobserved random variable.[1] It is assumed hereafter that the completeness condition is satisfied and then g uniquely identified.[2]

This model includes the nonparametric regression as a special case when $x_i = z_i$ and $g(x_i) = \mathrm{E}(y_i|x_i)$. It generalizes the nonparametric regression to allow some of the regressors x_i to be endogenous. The setup is general enough to allow for x_i to include a subset of z_i. The framework also includes the partial linear instrumental variables (IV) model as a special case

$$y_i = x_{1i}'\beta + h(x_{2i}) + e_i \tag{2.4}$$

where x_{1i} and x_{2i} are vectors of possibly endogenous variables.

It is well known that a nonparametric IV regression is an ill-posed inverse problem.[3] Taking conditional expectations on both sides of equation (2.1) with respect to z yields

$$\pi(z) \equiv \mathrm{E}(y|z) = \mathrm{E}(g(x)|z) = \int g(x) f(x|z) dx \tag{2.5}$$

where $f(x|z)$ is the conditional probability density function of x given z. The unknown function g solves the equation (2.5), which is an integral equation of the first kind; see Kress (1999). The main issue of solving this problem is that this equation is ill-posed, that is, the solution may not

[1] See Chernozhukov, Imbens, and Newey (2007), Horowitz and Lee (2007), and Gagliardini and Scaillet (2012a) for the quantile regression version of model (2.1)–(2.2).

[2] See Newey and Powell (2003), Blundell, Chen, and Kristensen (2007), Darolles, Fan, Florens, and Renault (2011), Andrews (2011), D'Haultfoeuille (2011), and Chen, Chernozhukov, Lee, and Newey (2014) for the identification results.

[3] The ill-posed inverse problem could be eliminated essentially if the structural function g is known to belong to a compact set and we restrict the estimator $\hat{g}$ to belong to this set.

exist or may not be continuous in the functions $\pi(z)$ and $f(x|z)$. Hence, g could not be estimated consistently by plugging in consistent estimates $\hat{\pi}(z)$ and $\hat{f}(x|z)$ in the equation (2.5).

To achieve consistency, it is necessary to regularize the mapping from reduced form estimators to the structural function. There is a variety of regularization approaches; see Kress (1999), Carrasco, Florens, and Renault (2007), and Centorrino, Feve, and Florens (2015). The regularization method we used is series truncation, which is a modified Petrov-Galerkin method.[4] We consider the series estimation that specifies the number of terms in a linear approximation to regularize the ill-posed inverse problem. The number of series expansion terms is called the regularization parameter, which makes the mapping continuous. The purpose of this paper is to construct a data-driven criterion for the regularization parameter selection.

We now consider a sequence of approximating models $m = 1, ..., M$, where the mth model uses J_m explanatory variables and K_m instruments, and M may go to infinity with the sample size n. We use m to denote a pair of explanatory variables and instruments, and $\mathcal{M}_n = \{1, ..., M\}$ a set of all pairs considered. Let $p_{mi} = p^{J_m}(x_i) = (p_1(x_i), \ldots, p_{J_m}(x_i))'$ be a $J_m \times 1$ vector of functions from a series expansion, such as power series or regression splines. Similarly, let $q^{K_m}(z_i)$ be a $K_m \times 1$ vector of functions from a series expansion such that $q_{mi} = q^{K_m}(z_i) = (q_1(z_i), \ldots, q_{K_m}(z_i))'$. Here we use J_m to denote the regularization parameter and K_m to denote the smoothing parameter.[5] The approximating models could be nested or non-nested. The models are nested if for $m_2 > m_1$, the pair $\left(p^{J_{m_2}}(x_i), q^{K_{m_2}}(z_i)\right)$ contains the pair $\left(p^{J_{m_1}}(x_i), q^{K_{m_1}}(z_i)\right)$ as a special case.

The mth approximating model is

$$y_i = p^{J_m}(x_i)'\beta_m + r_{mi} + e_i, \tag{2.6}$$

where $\beta_m = (\beta_1, \ldots, \beta_{J_m})'$ are coefficients of series expansion functions and $r_{mi} = g(x_i) - p^{J_m}(x_i)'\beta_m$ is the approximation error. Let $\mathcal{X}_m$ and $\mathcal{Z}_m$ be $n \times J_m$ and $n \times K_m$ matrices whose (i, j) elements are $p_j(x_i)$ and $q_j(z_i)$, respectively. In matrix notation, $y = g + e = \mathcal{X}_m\beta_m + r_m + e$, where $y = (y_1, \ldots, y_n)'$, $g = (g(x_1), \ldots, g(x_n))'$, $r_m = (r_{m1}, \ldots, r_{mn})'$, and $e = (e_1, \ldots, e_n)'$.

We follow Newey and Powell (2003), Blundell, Chen, and Kristensen (2007), and Newey (2013) and estimate the unknown structural function by nonparametric instrumental variables estimation. For $m = 1, ..., M$, we assume that there exists β_m such that

$$\mathrm{E}\left(\mathrm{E}\left(g(x_i) - p^{J_m}(x_i)'\beta_m | z_i\right)^2\right) \to 0 \tag{2.7}$$

when $J_m \to \infty$ and $K_m \to \infty$ as $n \to \infty$. This implies that $g(x_i)$ can be approximated by a series estimator, as in $g(x_i) \approx \sum_{j=1}^{J_m} \beta_j p_j(x_i)$ for all m. We then plug the approximation for $g(x_i)$ into

[4] Another regularization method is to add a penalty term to the minimization problem; see Newey and Powell (2003) and Blundell, Chen, and Kristensen (2007). This method, however, introduces more nuisance parameters. We therefore do not consider this method in this paper.

[5] Note that Blundell, Chen, and Kristensen (2007) and Chen and Pouzo (2012) use k and J, Chen and Reiss (2011) use m and J, and Chen and Christensen (2013) use J and K to denote the regularization and smoothing parameters, respectively.

(2.5) and obtain

$$\mathrm{E}(y_i|z_i) \approx \sum_{j=1}^{J_m} \beta_j \mathrm{E}\left(p_j(x_i)|z_i\right). \tag{2.8}$$

This equation suggests a nonparametric estimator, which is similar to the two-stage least squares estimator. In the first stage, we use a series estimator for $\mathrm{E}\left(p_j(x_i)|z_i\right)$, that is, we regress $\mathcal{X}_m$ on $\mathcal{Z}_m$ and obtain

$$\widehat{\mathrm{E}}\left(p_j(x_i)|z_i\right) = q^{K_m}(z_i)' \left(\sum_{i=1}^{n} q^{K_m}(z_i) q^{K_m}(z_i)' \right)^{-} \sum_{i=1}^{n} q^{K_m}(z_i) p_j(x_i). \tag{2.9}$$

In the second stage, the nonparametric IV estimator is the solution to the following minimization problem

$$\hat{S}(\beta_m) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J_m} \beta_j \widehat{\mathrm{E}}\left(p_j(x_i)|z_i\right) \right)^2 \tag{2.10}$$

which has a closed form solution as

$$\hat{\beta}_m = \left(\mathcal{X}_m' \mathcal{P}_m \mathcal{X}_m \right)^{-} \mathcal{X}_m' \mathcal{P}_m y \tag{2.11}$$

where $\mathcal{P}_m = \mathcal{Z}_m (\mathcal{Z}_m' \mathcal{Z}_m)^{-} \mathcal{Z}_m'$ is the projection matrix constructed by the series expansion function $\mathcal{Z}_m$. Note that the orthogonal series estimator of Horowitz (2011) corresponds to the nonparametric IV estimator with $J_m = K_m$. The nonparametric IV estimator of the unknown structural function is given by $\hat{g}_m = \mathcal{X}_m \hat{\beta}_m$. Under some regularity conditions, $\hat{g}_m$ is a consistent estimator of g as $n, J_m, K_m \to \infty$; see Newey and Powell (2003), Blundell, Chen, and Kristensen (2007), and Chen and Pouzo (2012).

The nonparametric IV estimator defined in (2.11) is not just a traditional two-stage least squares estimator with some flexible series expansion functions. In practice, the number of series terms can vary across different applications and data sets to account for more nonlinearity in both the first and second stage regressions. The point of the nonparametric IV estimator is not to just let both J_m and K_m increase with the sample size to approximate the unknown structural function, but also to let J_m grow appropriately slowly to regularize the ill-posed inverse problem. The empirical method of selecting J_m and K_m is described in the next section.

We follow Blundell, Chen, and Kristensen (2007) and Chen and Pouzo (2012) to define a sieve measure of ill-posedness. Denote the conditional expectation operator $T : L^q(x) \to L^q(z)$ as

$$Tg(z) = \mathrm{E}(g(x_i)|z_i = z) \tag{2.12}$$

for $1 \leq q \leq \infty$. The operator T is an integral operator mapping from one set of functions to another. Assume $g \in \mathcal{G}$ where $\mathcal{G}$ is a Besov space of functions. For $m = 1, ..., M$, we then define a measure of ill-posedness as

$$\tau_{mn} = \sup_{g \in \mathcal{G}_{mn} : \|g\|_{L^2(z)} \neq 0} \frac{\|g\|_{L^2(x)}}{\|Tg\|_{L^2(z)}} \tag{2.13}$$

where $\mathcal{G}_{mn}$ is the sieve space of $\mathcal{G}$. It is obvious that $\tau_{mn} \geq 1$ for all m and $\tau_{mn} = 1$ if x_i is exogenous. The nonparametric IV model is said to be mildly ill-posed if $\tau_{mn} = O(J_m^{\iota/d_x})$ and severely ill-posed if $\tau_{mn} = O(\exp(\frac{1}{2}J_m^{\iota/d_x}))$ for some $\iota > 0$.

3 Model Selection

We first describe the selection criterion of the regularization parameter and the smoothing parameter and then present a theoretical justification. As the series estimators are invariant to a nonsingular linear transformation of the approximating functions, we re-normalize $\mathcal{X}_m$ and $\mathcal{Z}_m$ so that $(p_1(x_i), \ldots, p_{J_m}(x_i))'$ and $(q_1(z_i) \ldots, q_{K_m}(z_i))'$ are orthonormal basis functions. This can be achieved by replacing q_{mi} by $\tilde{q}_{mi} = \mathrm{E}(q_{mi}q_{mi}')^{-1/2}q_{mi}$. Without loss of generality, we assume hereafter that these transformations have been made.

3.1 Loss Function and Selection Criterion

We now introduce some notations that we will use to characterize the selection criterion. Let $\hat{Q}_{z,m} = \mathcal{Z}_m'\mathcal{Z}_m/n$ be a $K_m \times K_m$ matrix as an estimate of $Q_{z,m} = \mathrm{E}(q_{mi}q_{mi}')$. Similarly, $Q_{x,m} = \mathrm{E}(p_{mi}p_{mi}')$. Define $\zeta_{z,m} = \sup_{z\in\mathcal{Z}}(q^{K_m}(z)'Q_{z,m}^{-1}q^{K_m}(z))^{1/2}$ the largest normalized Euclidean length of the instrument vector. Under standard conditions for series regression, $\zeta_{z,m}$ will be a bounded function of the dimension K_m. For example, $\zeta_{z,m}^2 = O(K_m^2)$ for a power series and $\zeta_{z,m}^2 = O(K_m)$ for a spline expansion. Similarly, $\zeta_{x,m} = \sup_{z\in\mathcal{Z}}(p^{J_m}(x)'Q_{x,m}^{-1}p^{J_m}(x))^{1/2}$. Let $\zeta_m = \max(\zeta_{x,m}, \zeta_{z,m})$. We also define the array of constants $\Psi_{mn} = \zeta_m\sqrt{(\log K_m)/n}$, which appear frequently in our bounds.

Let $\hat{\Gamma}_m = \mathcal{X}_m'\mathcal{Z}_m/n$ be a $J_m \times K_m$ matrix as an estimate of $\Gamma_m = \mathrm{E}(p_{mi}q_{mi}')$. Define ρ_{mn} be the smallest singular value of $Q_{x,m}^{-1/2}\Gamma_m Q_{z,m}^{-1/2}$. For each $g \in \mathcal{G}_{mn}$, define the $L^2(z)$ orthogonal projection of $Tg(\cdot)$ onto $\mathcal{Z}_m$ as

$$\Pi_{K_m}Tg(\cdot) = q^{K_m}(\cdot)'\mathrm{E}\left(q^{K_m}(z)Tg(z)\right) = q^{K_m}(\cdot)'\mathrm{E}\left(q^{K_m}(z)g(x)\right). \tag{3.1}$$

By the variational characterization of singular values, it follows

$$\rho_{mn} = \inf_{g\in\mathcal{G}_{mn}:\|g\|_{L^2(x)}=1} \|\Pi_{K_m}Tg\|_{L^2(z)} \leq 1. \tag{3.2}$$

Note that by the definition of τ_{mn}, we have $\rho_{mn} \leq \tau_{mn}^{-1}$. Similar to τ_{mn}, when $x_i = z_i$, we have $\rho_{mn} = 1$. We will later discuss the relation between ρ_{mn} and τ_{mn}.

We define the mean squared error as $L_n(m) = (g - \hat{g}_m)'(g - \hat{g}_m)/n$. The goal is to select the model, a pair of (J_m, K_m), to minimize the squared loss function $L_n(m)$. We first rewrite the nonparametric IV estimator as

$$\hat{\beta}_m = (\hat{\Gamma}_m\hat{Q}_{z,m}^{-}\hat{\Gamma}_m')^{-}\hat{\Gamma}_m\hat{Q}_{z,m}^{-}\mathcal{Z}_m'y/n \tag{3.3}$$

$$\hat{g}_m = \mathcal{X}_m\hat{\beta}_m = \hat{D}_m y \tag{3.4}$$

where $\hat{D}_m = \mathcal{X}_m(\hat{\Gamma}_m\hat{Q}^-_{z,m}\hat{\Gamma}'_m)^-\hat{\Gamma}_m\hat{Q}^-_{z,m}\mathcal{Z}'_m/n$. Note that $\hat{D}_m$ can be simplified as $\hat{D}_m = \mathcal{P}_m = \mathcal{Z}_m(\mathcal{Z}'_m\mathcal{Z}_m)^-\mathcal{Z}'_m$ in the special case where $x_i = z_i$. However, unlike $\mathcal{P}_m$, the matrix $\hat{D}_m$ is neither symmetric nor idempotent in general, which complicates the analysis. To address this difficulty, we follow Chen and Christensen (2013) and Hansen (2015) to develop the useful bounds on the estimated matrix $\hat{D}_m$ in the appendix.

Define the residuals as $\hat{e}_m = y - \mathcal{X}_m\hat{\beta}_m$. Expanding the sum of squared errors, we have

$$\begin{aligned}\frac{1}{n}\hat{e}'_m\hat{e}_m &= \frac{1}{n}(e + g - \hat{g}_m)'(e + g - \hat{g}_m) \\ &= \frac{1}{n}(g - \hat{g}_m)'(g - \hat{g}_m) + \frac{1}{n}e'e + \frac{2}{n}e'(g - \hat{g}_m) \\ &= L_n(m) + \frac{1}{n}e'e + \frac{2}{n}e'(I - \hat{D}_m)r_m - \frac{2}{n}e'\hat{D}_m e.\end{aligned}$$

Note that the second term does not depend on m, and the third term is related to the approximation error. In the proof of Theorem 1, we show that the third term converges to zero faster than $L_n(m)$. Thus, the fourth term is the dominant term that depends on m. The idea behind the proposed Mallows' C_p-type criterion is to approximate the mean squared error by the sum of squared errors and a penalty term, an estimate of the fourth term.

The proposed criterion function is

$$C_n(m) = \frac{1}{n}\hat{e}'_m\hat{e}_m + \frac{2}{n}\sigma^2\rho^{-1}_{mn}\sqrt{J_m K_m}, \tag{3.5}$$

where ρ_{mn} is defined in (3.2). The first term of the criterion measures the model fit, while the second term of the criterion measures the model complexity and serves as a penalty term. Thus, the criterion aims to balance the model fit and model complexity. Unlike the traditional model selection criterion, the penalty term is a nonlinear function of J_m and K_m, which account for the interaction effect between the regularization parameter and the smoothing parameter. We choose the model with the smallest $C_n(m)$. One attractive advantage is that the criterion is quite easy to compute. The criterion function defined in (3.5) can also be used for the partial linear IV model (2.4).[6] For the special case when $x_i = z_i$, then $\rho^{-1}_{mn} = 1$ and the criterion function can be simplified as $C_n(m) = \hat{e}'_m\hat{e}_m + 2\sigma^2 K_m$, which is just the traditional Mallows criterion for the linear regression model.

3.2 Asymptotic Optimality

Li (1987) has established conditions under which the Mallows criterion achieves the asymptotic optimality in the sense that the mean squared error of the selected estimator is asymptotically equivalent to the lowest mean squared error among all candidates. We now extend Li's results to the case of the model in the presence of endogenous variables. Our result also generalizes

[6]The partial linear IV model is $y_i = g(x_i) + e_i = x'_{1i}\beta + h(x_{2i}) + e_i$ and $\mathrm{E}(e_i|z_i) = 0$ where $x_i = (x'_{1i}, x'_{2i})'$. Suppose that x_{1i} and x_{2i} are $J_1 \times 1$ and $J_2 \times 1$ vectors of possibly endogenous variables, respectively. Let $p^{J_{2m}}(x_{2i})$ and $q^{K_m}(z_i)$ be $J_{2m} \times 1$ and $K_m \times 1$ vectors of functions from a series expansion. Then the criterion function for the partial linear IV model is defined as $C_n(m) = \frac{1}{n}\hat{e}'_m\hat{e}_m + \frac{2}{n}\sigma^2\rho^{-1}_{mn}\sqrt{(J_1 + J_{2m})K_m}$.

the asymptotic optimality in Donald and Newey (2001) to the case of a nonparametric structural function of interest.

We first follow Donald and Newey (2001) and consider a nonparametric reduced form relationship between the endogenous variable x_i and the exogenous variables z_i. Recall that $p_{mi} = p^{J_m}(x_i)$. Let $f_{mi} = f^{J_m}(z_i) = \mathrm{E}(p_{mi}|z_i)$ be a $J_m \times 1$ vector of conditional expectation functions. Then, $p_{mi} = f_{mi} + u_{mi} = \mathrm{E}(p_{mi}|z_i) + u_{mi}$, and $\mathrm{E}(u_{mi}|z_i) = 0$ by construction. In matrix notation, we write $\mathcal{X}_m = F_m + u_m$, where $F_m = (f_{m1}, \ldots, f_{mn})'$ and $u_m = (u_{m1}, \ldots, u_{mn})$.

We next consider an approximation of the conditional squared error $\mathrm{E}(L_n(m)|Z)$ where $Z = (z_1, \ldots, z_n)$. Ideally, we might consider $E(L_n(m)|Z)$. Unfortunately, it is not easy for us to study the conditional squared error directly. This is because the mean squared error $L_n(m)$ is a function of $\hat{D}_m$ and we are not able to separate regressors and instruments from $\hat{D}_m$ and take the conditional expectation of $\hat{D}_m$. Thus, we introduce a function $R_n(m)$ as an approximation of $\mathrm{E}(L_n(m)|Z)$. Define $\phi_m^2 = \mathrm{E}\left(r_{mi}^2|z_i\right)$ and $\phi_{m,\ell} = \mathrm{E}\left(r_{mi}r_{\ell i}|z_i\right)$. Let $D_m = F_m(\Gamma_m Q_{z,m}^- \Gamma_m')^- \Gamma_m Q_{z,m}^- Z_m'/n$ and $\tilde{D}_m = (\Gamma_m Q_{z,m}^- \Gamma_m')^- \Gamma_m Q_{z,m}^- Z_m'/n$. The approximation of $\mathrm{E}(L_n(m)|Z)$ is defined as

$$R_n(m) = \frac{\phi_m^2}{n} tr((I - D_m)'(I - D_m)) + \frac{\sigma^2}{n} tr(D_m' D_m) + \frac{1}{n}\mathrm{E}(e'\tilde{D}_m' u_m' u_m \tilde{D}_m e|Z), \tag{3.6}$$

where the last term captures the higher moment of e_i and u_{mi}. In the proof of Theorem 1, we show that $\sup_{m\in\mathcal{M}_n} |L_n(m)/R_n(m) - 1| \xrightarrow{p} 0$, that is, $L_n(m)$ and $R_n(m)$ are asymptotically equivalent.

We now state the assumptions. For some positive integers p and N, the following conditions hold almost surely. Here p indicates the smoothness of the function g.

Assumption 1. *(i) $\{y_i, x_i, z_i\}$ are independent and identically distributed (i.i.d.). (ii) The supports of x and z are compact. (iii) $0 < \sigma^2 < \infty$. (iv) $\mathrm{E}(|e_i|^{4(N+1)}|z_i) < \infty$. (v) $\mathrm{E}(\|u_{mi}\|^{4(N+1)}|z_i) < \infty$ for all m.*

Assumption 2. *(i) g is point identified. (ii) $\mathrm{E}((e_i, u_{mi}')'(e_i, u_{mi}')|z_i)$ is constant. (iii) $\mathrm{E}(e_i^2 u_{mi}|z_i) = 0$. (iv) For each K_m there exists π_{K_m} such that $\mathrm{E}(\|f(z) - \pi_{K_m} q^{K_m}(z)\|^2) = O(\sqrt{K_m/n})$ as $n \to \infty$, $K_m \to \infty$, and $K_m/n \to 0$.*

Assumption 1 specifies the data are *i.i.d.* and imposes some moment conditions. Assumption 2 concerns the approximation of the conditional expectation function and the instruments. Assumption 2 (i) is satisfied if the completeness condition holds. Assumption 2 (ii) imposes the homoskedasticity condition. Assumption 2 (iii) concerns the third moment condition. This condition holds if the joint conditional distribution of e_i and u_{mi} is symmetric around zero. Assumption 2 (iv) requires that the unknown reduced form can be approximated arbitrarily well for large enough n and K_m. Assumptions 1 and 2 are similar to Assumptions 1–3 of Donald and Newey (2001).

Assumption 3. *(i) $J_m, K_m \to \infty$ as $n \to \infty$ and $J_m \leq K_m$. (ii) $K_m = O(J_m)$ and $J_m^2/n = o(1)$. (iii) $\lambda_{\min}(\mathrm{E}(p_{mi}p_{mi}')) > \underline{\lambda} > 0$ for all m. (iv) $\lambda_{\min}(\mathrm{E}(q_{mi}q_{mi}')) > \underline{\lambda} > 0$ for all m. (v) $\rho_{mn}^{-1} J_m/n \to \infty$ for all m.*

Assumption 3 concerns the sieve bases. Assumption 3 (i) specifies that the model is over-identified or just-identified. Assumption 3 (ii) is a mild restriction on the relationship between J_m and K_m, which is standard in the literature. Assumptions 3 (i)–(iv) are satisfied by many wildly used sieve bases such as spline and polynomial series. Assumption 3 (v) restricts the rate of increase of J_m as the sample size increases. Assumptions 3 (i)–(iv) are similar to Assumptions 3 (ii) and 4 (ii) of Chen and Christensen (2013). Assumption 3 (v) is similar to Assumption 6 of Horowitz (2014).

Assumption 4. *(i)* $\sup_{m\in\mathcal{M}_n} \rho_{mn}^{-1}\zeta_m\sqrt{(\log J_m)/n} \to 0$ *(ii)* $0 < \phi_{m,\ell} < \infty$ *for all* m.

Assumption 4 concern the model complexity and the approximation error. Assumption 4 (i) puts a bound on the number of series terms relative to the sample size and indirectly bounds the number of models. Assumption 4 (ii) states that the approximation error is nonzero for all finite dimensional models, and thus all models are approximations. This is quite standard in the nonparametric literature. Assumption 4 (i) is similar to Assumptions 1 of Hansen (2015) and Assumption 3 of Chen and Christensen (2015). Assumption 4 (ii) is similar to Assumptions 4 of Hansen (2015).

Assumption 5. *(i)* $\limsup_{n\to\infty m\in\mathcal{M}_n} \sigma_{\max}(D_m) < \infty$. *(ii)* $\limsup_{n\to\infty m\in\mathcal{M}_n} \sigma_{\max}(I - D_m) < \infty$.

Assumption 6. $\sum_{m\in\mathcal{M}_n} (nR_n(m))^{-(N+1)} \to 0$.

Assumptions 5 and 6 are quite standard in the nonparametric optimality literature. Assumptions 5 and 6 correspond to conditions (A.1) and (A.3) of Li (1987). Note that the choice of N involves a trade-off between the conditional moment bound in Assumption 1 (iv) and the summation of approximate risk function in Assumption 6. For the nested model selection problem, we can replace Assumption 6 with weaker conditions; see Assumption 7.

The following result shows the asymptotic optimality of the proposed model selection criterion for the nonparametric IV model.

Theorem 1. *Suppose Assumptions 1–6 hold. Let* $\hat{m} = \underset{m\in\mathcal{M}_n}{\operatorname{argmin}} C_n(m)$. *Then*

$$\frac{L_n(\hat{m})}{\inf_{m\in\mathcal{M}_n} L_n(m)} \xrightarrow{p} 1.$$

Theorem 1 shows that the mean squared error of the nonparametric IV estimator with selected m is asymptotically equivalent to that of the infeasible best estimator in the class of nonparametric IV estimators considered in the set of models $\mathcal{M}_n$. This result generalizes the asymptotic optimality in Li (1987) to allow for possibly endogenous variables.

The proof of Theorem 1 is not a trivial extension of that of Theorem 2.1 of Li (1987). We first need to deal with the endogenous variables and consider the approximation error in the first stage. Second, in order to apply Whittle's inequality, we need to show that $\|\hat{D}_m - D_m\|$ is negligible compared with $R_n(m)$ uniformly for any $m \in \mathcal{M}_n$.

A necessary condition for Assumption 6 is

$$\inf_{m\in\mathcal{M}_n} nR_n(m) \to \infty \tag{3.7}$$

almost surely as $n \to \infty$. This assumption implies that all finite dimensional models are approximations, and thus no finite dimensional model is correctly specified. We can use (3.7) to obtain a crude bound for Assumption 6 under the nested model setup. Suppose sieve basis functions $p^{J_m}(x_i)$ and $q^{K_m}(z_i)$ are nested. The nested sieve basis means that $p^{J_{m_2}}(x_i)$ contains $p^{J_{m_1}}(x_i)$ as a special case for $m_2 > m_1$. For example, the power series or a sequence of splines where the knots are set sequentially. Observe that $nR_n(m) \geq \sigma^2 tr(D_m'D_m) \geq \sigma^2 J_m K_m \geq \sigma^2 J_m^2$. Let $N = 1$ and pick $a_n \to \infty$ so that $a_n(\inf nR_n(m))^{-2} \to 0$. Then we have

$$\begin{aligned}\sum_{m\in\mathcal{M}_n} (nR_n(m))^{-2} &\leq \sum_{m=1}^{a_n} (nR_n(m))^{-2} + \sigma^{-4} \sum_{m=a_n+1}^{M} J_m^{-4} \\ &\leq a_n \left(\inf_{m\in\mathcal{M}_n} nR_n(m)\right)^{-2} + \sigma^{-4} \sum_{m=a_n+1}^{\infty} J_m^{-4} \to 0.\end{aligned}$$

This shows Assumption 6. We now summarize the result as follows.

Assumption 7. *(i)* $\inf_{m\in\mathcal{M}_n} nR_n(m) \to \infty$. *(ii) The pair* $\left(p^{J_m}(x_i), q^{K_m}(z_i)\right)$ *is nested.*

Corollary 1. *Suppose Assumptions 1–5 and 7 hold. Let* $\hat{m} = \underset{m\in\mathcal{M}_n}{\operatorname{argmin}} C_n(m)$. *Then*

$$\frac{L_n(\hat{m})}{\inf_{m\in\mathcal{M}_n} L_n(m)} \xrightarrow{p} 1.$$

3.3 Implementation

In practice, both σ^2 and ρ_{mn}^{-1} in the selection criterion (3.5) are unknown. We follow Hansen (2007) and propose to estimate σ^2 by $\hat{\sigma}_M^2 = \hat{e}_M'\hat{e}_M/n$, where $\hat{e}_M = y - \mathcal{X}_M\hat{\beta}_M$ are the residuals from the largest approximating model. Theorem 2 of Hansen (2007) shows that $\hat{\sigma}_M^2$ is consistent for σ^2 for the special case when $x_i = z_i$. We conjecture that the consistency will extend to the case in the presence of endogenous variables. Thus, Theorem 1 and Corollary 1 continue to hold when σ^2 is replaced by $\hat{\sigma}_M^2$.

To obtain a consistent estimator of ρ_{mn} is more challenging. Recall the definition of ρ_{mn} in (3.2). We can estimate ρ_{mn} by the sample analog

$$\hat{\rho}_{mn} = \inf_{g_n\in\mathcal{G}_{mn}:\|g_n\|=1} \sqrt{\frac{1}{n}\sum_{i=1}^{n} (q^{K_m}(z_i)'\hat{\mathrm{E}}\left(q^{K_m}(z_i)g_n(x_i)\right))^2}$$

where $q^{K_m}(z_i)'\hat{\mathrm{E}}\left(q^{K_m}(z_i)g_n(x_i)\right)$ is a nonparametric estimate of the orthogonal projection for any fixed $g_n \in \mathcal{G}_{mn}$. Then, $\hat{\rho}_{mn}^2$ is the smallest eigenvalue of $\hat{\Gamma}_m\hat{Q}_{z,m}^{-}\hat{\Gamma}_m'$. However, our simulation shows that $\hat{\rho}_{mn}^{-1}$ is not a stable estimate for ρ_{mn}^{-1}.

An alternative estimator of ρ_{mn}^{-1} is the sieve measure of ill-posedness τ_{mn} since $\rho_{mn}^{-1} \geq \tau_{mn}$ by the definition of τ_{mn}. The sieve measure of ill-posedness τ_{mn} can be easily estimated by

$$\hat{\tau}_{mn} = \sup_{g_n \in \mathcal{G}_{mn}: g_n \neq 0} \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(g_n(x_i))^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathrm{E}}\left(g_n(x_i)|z_i = z\right))^2}}$$

where $\hat{\mathrm{E}}\left(g_n(x_i)|z_i = z\right)$ is a nonparametric estimate of the conditional expectation $\mathrm{E}(g(x_i)|z_i = z)$ for any fixed $g_n \in \mathcal{G}_{mn}$. That is, $\hat{\tau}_{mn}$ is the largest eigenvalue of $((\mathcal{X}_m'\mathcal{X}_m/n)(\hat{\Gamma}_m\hat{Q}_{z,m}^{-}\hat{\Gamma}_m)^{-})$.

As shown in Lemma 3.1 in Chen and Christensen (2013), we have $\rho_{mn}^{-1} \leq \tau_{mn}$ when the sieve space of $q^{K_m(z_i)}$ contains the closed linear subspace in $L^2(z)$ generated by the eigenfunction (orthonormal) base for $L^2(z)$. Therefore, it follows that $\rho_{mn}^{-1} = \tau_{mn}$. Thus, Theorem 1 and Corollary 1 continue to hold as long as $\hat{\tau}_{mn}$ is consistent for ρ_{mn}^{-1}. However, it is hard to verify if the condition for $\rho_{mn}^{-1} = \tau_{mn}$ holds or not. Nevertheless, our simulation results show that $\hat{\tau}_{m,n}$ is a more stable estimator than $\hat{\rho}_{mn}^{-1}$. In practice, we recommend to use $\hat{\tau}_{mn}$ as an estimator to approximate ρ_{mn}^{-1}.

Besides σ^2 and ρ_{mn}^{-1}, we also need to specify the set of models $\mathcal{M}_n$ to implement the selection criterion. That is, we have to first choose the highest order for polynomial series or the maximum number of knots for regression splines. Assumption 4 (i) indirectly puts a bound on the number of models. However, it does not provide us a clear rule to select the initial set of models. One possible way to tackle this problem is to follow the literature on adaptive estimation, for example, Horowitz (2014) and Breunig and Johannes (2015), and choose $\mathcal{M}_n$ empirically. It is unclear if the asymptotic optimality will still hold with the adaptive choice $\mathcal{M}_n$. Such an investigation is beyond the scope of this paper and is left for future research. In the simulations, we increase the initial set of models and find that the relative performance of proposed estimators and other existing methods is not sensitive to the choice of $\mathcal{M}_n$.

4 Model Averaging

In this section, we introduce a new nonparametric IV model averaging estimator. The proposed averaging estimator generalizes the Mallows model averaging estimator of Hansen (2007) to allow for possibly endogenous variables. Let $w = (w_1, ..., w_M)'$ be a weight vector with $w_m \geq 0$ and $\sum_{m=1}^{M} w_m = 1$. That is, $w \in \mathcal{H}_n$ where $\mathcal{H}_n = \{w \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\}$. The nonparametric IV model averaging estimator of g is defined as

$$\hat{g}(w) = \sum_{m=1}^{M} w_m \hat{g}_m = \sum_{m=1}^{M} w_m \hat{D}_m y = \hat{D}(w)y \tag{4.1}$$

where $\hat{D}(w) = \sum_{m=1}^{M} w_m \hat{D}_m$.[7] The averaging estimator includes the nonparametric IV estimator from the mth approximating model as a special case by setting the weight vector w to equal the

[7] As an alternative averaging estimator for nonparametric IV models, we may consider constructing optimal instruments from the first stage and then forming the averaging estimator. Although it is feasible in implementation, the efficiency gain of using this two-step averaging estimator is unknown. We therefore do not consider this extension in our analysis.

unit weight vector w_m^0 where the mth element is one and others are zeros.

Define the mean squared error of the averaging estimator as $L_n(w) = (g - \hat{g}(w))'(g - \hat{g}(w))/n$. The goal is to select a weight vector to minimize the squared loss function $L_n(w)$.

The averaging residual vector is

$$\hat{e}(w) = y - \hat{g}(w) = \sum_{m=1}^{M} w_m \hat{e}_m = \hat{e}w \tag{4.2}$$

where $\hat{e} = (\hat{e}_1, ..., \hat{e}_M)$ is a $n \times M$ matrix of residuals. Define $\hat{b}_m = (I - \hat{D}_m)g = (I - \hat{D}_m)r_m$. Thus, $g - \hat{g}(w) = (I - \hat{D}(w))g - \hat{D}(w)e = \hat{b}(w) - \hat{D}(w)e$ where $\hat{b}(w) = \sum_{m=1}^{M} w_m \hat{b}_m$. To construct the criterion function for the weight selection, we adopt the same strategy and expand the sum of squared errors

$$\begin{aligned}
\frac{1}{n}\hat{e}(w)'\hat{e}(w) &= \frac{1}{n}(e + g - \hat{g}(w))'(e + g - \hat{g}(w)) \\
&= \frac{1}{n}(g - \hat{g}(w))'(g - \hat{g}(w)) + \frac{1}{n}e'e + \frac{2}{n}e'(g - \hat{g}(w)) \\
&= L_n(w) + \frac{1}{n}e'e + \frac{2}{n}e'\hat{b}(w) - \frac{2}{n}e'\hat{D}(w)e.
\end{aligned}$$

Similar to the case of model selection, we approximate the mean squared error $L_n(w)$ by the sum of squared errors and a penalty term, an estimate of the fourth term, since the second term does not depend on w and the third term converges to zero faster than $L_n(w)$. The criterion function for the nonparametric IV model averaging estimator is

$$C_n(w) = \frac{1}{n}w'\hat{e}'\hat{e}w + \frac{2\sigma^2}{n}\sum_{m=1}^{M} w_m \rho_{mn}^{-1}\sqrt{J_m K_m}, \tag{4.3}$$

and an approximation of the conditional squared error $\mathrm{E}(L_n(w)|Z)$ is

$$\begin{aligned}
R_n(w) = &\frac{1}{n}\sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell \phi_{m,\ell} tr((I - D_m)'(I - D_\ell)) \\
&+ \frac{\sigma^2}{n}\sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell tr(D_m' D_\ell) + \frac{1}{n}\sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell \mathrm{E}(e'\tilde{D}_m' u_m' u_\ell \tilde{D}_\ell e|Z).
\end{aligned} \tag{4.4}$$

The proposed nonparametric IV model averaging estimator is defined as

$$\hat{g}(\hat{w}) = \sum_{m=1}^{M} \hat{w}_m \hat{g}_m = \hat{g}\hat{w} \tag{4.5}$$

$$\hat{w} = \underset{w \in \mathcal{H}_n}{\operatorname{argmin}} C_n(w) \tag{4.6}$$

where $\hat{g} = (\hat{g}_1, ..., \hat{g}_M)$ is the $n \times M$ matrix of estimates. Note that the criterion function $C_n(w)$ is a quadratic function of the weight vector, and thus the weight vector can be found numerically via quadratic programming for which numerical algorithms are available for most programming languages.

For the special case when $x_i = z_i$, the criterion function can be simplified as $C_n(w) = w'\hat{e}'\hat{e}w + 2\sigma^2 \sum_{m=1}^{M} w_m K_m$, which is Mallows model averaging estimator proposed by Hansen (2007). If we consider the unit weight vector w_m^0, then the averaging estimator simplifies to a selection estimator. Thus, $C_n(w_m^0)$ is equivalent to the selection criterion proposed in (3.5) and its minimizer $\hat{w}_m^0$ equals $\hat{m} = \text{argmin}_{m \in \mathcal{M}_n} C_n(m)$. Therefore, the nonparametric IV model averaging estimator $\hat{g}(\hat{w})$ is a generalization of the model selection estimator considered in the previous section.

To demonstrate the asymptotic optimality, we follow Hansen (2007) and Hansen and Racine (2012) and restrict the weights w_m to a discrete set, that is, $w \in \mathcal{H}_n^*$ where $\mathcal{H}_n^* = \{0, \frac{1}{N}, \frac{2}{N}, ..., 1\}$ for some positive integer N. The set $\mathcal{H}_n^*$ is a subset of $\mathcal{H}_n$, and we can make this restriction less binding by making N sufficiently large as long as the conditional moment bounds in Assumption 1 hold. In practice, we set $\mathcal{H}_n^* = \mathcal{H}_n$ and minimize the criterion $C_n(w)$ subject to $\mathcal{H}_n$.

The following result shows the asymptotic optimality of the proposed nonparametric IV model averaging estimator.

Assumption 8. $\sum_{w \in \mathcal{H}_n^*} (nR_n(w))^{-(N+1)} \to 0.$

Theorem 2. *Suppose Assumptions 1–5 and 8 hold. Let* $\hat{w} = \underset{w \in \mathcal{H}_n^*}{\text{argmin}} C_n(w)$. *Then*

$$\frac{L_n(\hat{w})}{\inf_{w \in \mathcal{H}_n^*} L_n(w)} \xrightarrow{p} 1.$$

Theorem 2 shows that the mean squared error of the nonparametric IV model averaging estimator with selected weights $\hat{w}$ is asymptotically equivalent to that of the nonparametric IV model averaging estimator with the infeasible optimal weights. This means the proposed nonparametric IV model averaging estimator (4.5) is asymptotically optimal in the class of averaging estimators (4.1) where the weight vector is restricted to the discrete set $\mathcal{H}_n^*$, which is a boarder class of estimators in $\mathcal{M}_n$. This result generalizes the asymptotic optimality of the averaging estimator in Hansen (2007) to the regression model in the presence of endogenous variables.

Assumption 8 is the counterpart of Assumption 6. This condition is similar to Condition (A.6) of Hansen and Racine (2012). A necessary condition for Assumption 8 is

$$\xi_n = \inf_{w \in \mathcal{H}_n} nR_n(w) \to \infty \tag{4.7}$$

almost surely as $n \to \infty$. This is similar to Condition (15) of Hansen (2007) and Condition (A.7) of Hansen and Racine (2012). It requires that all finite dimensional models are approximations, and thus no finite dimensional model is correctly specified. We follow Hansen and Racine (2012) and use (4.7) to obtain a primitive condition for Assumption 8. The idea is to limit the number of models for each dimension instead of placing the restriction on the number of models or the dimension of the largest model.

Let q_{rn} be the number of models that have exactly r parameters, i.e., $q_{rn} = \#\{m : J_m + K_m = r\}$, for example, if we consider a sequence of nested models with $J_m = K_m$, then $q_{2n} = q_{4n} = \cdots = 1$.

If we consider all possible pairs of $\left(p^{J_m}(x_i), q^{K_m}(z_i)\right)$ with a $dim(x_i) = dim(z_i) = 1$, then $q_{rn} = r/2$ for r is even and $q_{rn} = (r-1)/2$ for r is odd. Let $\bar{q}_n = \max_{r \leq \bar{r}} q_{rn}$ be the largest number of models of any given dimension where $\bar{r} = \max_{m \in \mathcal{M}_n}\{J_m + K_m\}$ is the largest number of parameters among all candidate models. We impose the restriction on the rate of growth of $\bar{q}_n$ as follows.

Assumption 9. *(i)* $\xi_n = \inf_{w \in \mathcal{H}_n} nR_n(w) \to \infty$. *(ii)* $\bar{q}_n = o(\xi_n^{1/N})$.

Theorem 3. *Suppose Assumptions 1–5 and 9 hold. Let* $\hat{w} = \underset{w \in \mathcal{H}_n^*}{\operatorname{argmin}} C_n(w)$. *Then*

$$\frac{L_n(\hat{w})}{\inf_{w \in \mathcal{H}_n^*} L_n(w)} \xrightarrow{p} 1.$$

Theorem 3 shows that the proposed nonparametric IV model averaging estimator is asymptotically optimal under primitive conditions. Assumptions 1 (iii)–(iv) and Assumption 9 together imply a trade-off between the number of non-nested models and the moments of the error. When higher moments of the error are finite, we impose fewer restrictions on the number of models for each dimension.

5 Simulation Study

In this section, we investigate the finite sample mean squared error of the selection and averaging estimators via Monte Carlo experiments. The simulation design is similar to that of Horowitz (2012).

5.1 Simulation Setup

We consider the following data generating process

$$\begin{aligned} y_i &= g(x_i) + \sigma_e e_i \\ g(x) &= \sum_{j=1}^{\infty} (-1)^{j+1} j^{-2} \sin(j\alpha\pi x) \\ x_i &= \Phi(u_{1i} + u_{2i}) \\ z_i &= \Phi(u_{1i}) \\ e_i &= \lambda u_{2i} + (1-\lambda) u_{3i} \end{aligned}$$

where u_{1i}, u_{2i}, u_{3i} are generated from independent standard normal distributions and $\Phi(\cdot)$ is the cumulative standard normal distribution function. We set $\sigma_e = 0.5$. The parameter λ measures the degree of endogeneity and is varied between 0.1 and 0.9. The parameter α controls the wavelength of the sine function, and we vary α from 1 to 4. For computational purposes, the series in the function g are truncated at $j = 100$. The function g for different α are displayed in Figure 1. The sample size n is varied between 50 and $1,250$.

The basis functions $p^{J_m(x_i)}$ and $q^{K_m(z_i)}$ are either Legendre polynomials or a third order B-spline. Legendre polynomials are centered and scaled to be orthonormal on $[0,1]$, and B-splines are orthonormalized by the Gram-Schmidt procedure. The order of polynomials is varied from $\ell = 1, 2, ..., \ell_n$, and the number of knots of B-splines is varied from $\ell = 0, 1, ..., \ell_n$. We set $\ell_n = 2n^{1/5}$ for $p^{J_m(x_i)}$ and $\ell_n = 2.5n^{1/5}$ for $q^{K_m(z_i)}$. We consider approximating models with $J_m = K_m$ for Horowitz's adaptive estimator and consider all possible approximating models with $J_m \leq K_m$ for other estimators in all experiments.

We consider the following estimators: (1) Horowitz's (2014) adaptive nonparametric IV estimation (Horowitze),[8] (2) the nonparametric IV estimator with Sueishi's (2012) selection criterion (Sueishi),[9] (3) the nonparametric IV Mallows model selection estimator with ρ_{mn}^{-1} estimated by $\hat{\rho}_{mn}^{-1}$ (Mallows-ρ), (4) the nonparametric IV Mallows model selection estimator with ρ_{mn}^{-1} estimated by $\hat{\tau}_{mn}$ (Mallows-τ), (5) the nonparametric IV model averaging estimator with ρ_{mn}^{-1} estimated by $\hat{\rho}_{mn}^{-1}$ (Averaging-ρ), and (6) the nonparametric IV model averaging estimator with ρ_{mn}^{-1} estimated by $\hat{\tau}_{mn}$ (Averaging-τ).

To evaluate the finite behavior of the selection and averaging estimators, we compute the root mean squared error (RMSE) by averaging across the realized values of x_i and $5,000$ random samples. We follow Hansen (2007) and normalize the RMSE by dividing by the RMSE of the infeasible optimal nonparametric IV estimator, i.e., the RMSE of the best-fitting approximating model m.

5.2 Simulation Results

The normalized RMSE are displayed in Figures 2–6 for either Legendre polynomials or B-splines. Figure 2 shows the normalized RMSE for all six estimators for $\alpha = 1$ and $n = 150$. The normalized RMSE of proposed selection estimators, Mallows-ρ and Mallows-τ, are close to that of infeasible optimal model selection, while Horowitz's adaptive estimator and Sueishi's selection estimator have large normalized RMSE for some values of λ. It is clear that the averaging estimators, Averaging-ρ and Averaging-τ, have much lower normalized RMSE than other estimators. The averaging estimators have lower normalized RMSE than 1, which means that the RMSE of averaging estimators are lower than that of the infeasible best-fitting approximating model m.

Comparing the two estimators for ρ_{mn}^{-1}, we find that the estimate $\hat{\rho}_{mn}^{-1}$ is dominated by the esti-

[8] Horowitz (2014) proposes an adaptive estimator that minimizes the sample analog of the weighted asymptotic integrated mean squared error. The adaptive estimation is a two-step procedure. Let J_n be a preliminary series truncation point. The first-stage estimator is $\tilde{g} = \mathcal{X}_{J_n}\hat{\beta}_{J_n}$ where $\hat{\beta}_{J_n} = (\mathcal{Z}'_{J_n}\mathcal{X}_{J_n})^{-}\mathcal{Z}'_{J_n}y$. For $J \leq J_n$, the second-stage estimator is $\hat{g}_J = \sum_{j=1}^{J}\hat{\beta}_j p_J(x_i)$ where $\hat{\beta}_j$ is the jth element of $\hat{\beta}_{J_n}$. The adaptive estimator is defined as $\hat{g}_{\hat{J}}$ and $\hat{J} = \operatorname{argmin} T_n(J)$ where $T_n(J) = (2/3)(\log n)n^{-2}\sum_{i=1}^{n}\left((y_i - \tilde{g}(x_i))^2 \times \sum_{j=1}^{J}((\mathcal{Z}'_J\mathcal{X}_J)^{-}q_j(z_i))^2\right) - \|\hat{g}_J\|^2$. The preliminary series truncation point is estimated by $\hat{J}_n = \operatorname{argmin}\{\hat{\tau}_J^2 J^{3.5}/n : \hat{\tau}_J^2 J^{3.5}/n - 1 \geq 0\}$ where $\hat{\tau}_J$ is the estimate of the sieve measure of ill-posedness. Horowitz (2014) shows that $\mathrm{E_A}\|\hat{g}_{\hat{J}} - g\|^2 \leq \big(2 + (4/3)\log(n)\big)\mathrm{E_A}\|\hat{g}_{J_{opt}} - g\|^2$ where $\mathrm{E_A}(x)$ is the mean of the leading term of the asymptotic expansion of the random variable x.

[9] Instead of the mean squared error, Sueishi (2012) considers a loss function spanned by instruments to evaluate the nonparametric IV model. The loss function is defined as $\tilde{L}_n(m) = \|\mathcal{P}_m(g - \hat{g}_m)\|^2/n$. He proposes a Mallows selection criterion $\tilde{C}_n(m) = \|\mathcal{P}_m(y - \hat{g}_m)\|^2/n - \sigma^2(K_m - 2J_m)/n$ and demonstrates its asymptotic optimality. The proposed criterion, however, might not be optimal with respect to the squared loss function.

mate $\hat{\tau}_{mn}$, i.e., Mallows-ρ is dominated by Mallows-τ, and Averaging-ρ is dominated by Averaging-τ. To keep the graphs uncluttered, we only report the results of $\hat{\tau}_{mn}$ in the remaining figures.

Figures 3 and 4 show the normalized RMSE for Legendre polynomials and B-splines, respectively. We plot the normalized RMSE functions for $\alpha = 1$ and for the sample size $n = 50$, 150, 400, and $1,000$ in four panels. For Legendre polynomial basis functions, Averaging-τ has the best performance and Sueishi's selection estimator has the worst performance. Mallows-τ and Horowitz's adaptive estimator do not dominate each other uniformly. Mallows-τ has lower normalized RMSE than Horowitz's adaptive estimator for $n = 50$ and 150, while Horowitz's adaptive estimator has better performance than Mallows-τ for $n = 400$ and $1,000$. For B-splines base functions, both Averaging-τ and Mallows-τ have much lower normalized RMSE than other estimators. Sueishi's selection estimator has better performance than Horowitz's adaptive estimator for $n = 50$, 150, and 400, while both estimators have similar normalized RMSE for $n = 1,000$.

Figures 5 and 6 examine the effect of the sample size and the parameter α on the normalized RMSE for Legendre polynomials and B-splines, respectively. We plot the normalized RMSE functions for $\lambda = 0.5$ and for $\alpha = 1$, 2, 3, and 4 in four panels. As the sample size increases, the normalized RMSE of both Horowitz's and Sueishi's estimators increases. Therefore, it shows that both estimators are not asymptotically optimal in terms of RMSE. Both figures also show that both Averaging-τ and Mallows-τ are relatively unaffected by the value of α, while the performance of Horowitz's and Sueishi's estimators strongly depends on features that we do not know. It is also instructive to compare the results of Legendre polynomials and B-splines. Both Averaging-τ and Mallows-τ have similar results for both basis functions, while the Horowitz's and Sueishi's estimators are sensitive to the choice of basis functions.

Figures 7 and 8 examine the sensitivity of the choice of the set of models $\mathcal{M}_n$ on the normalized RMSE for $\alpha = 1$ and 2, respectively. The base function is a third order B-spline with the number of knots varied from $\ell = 0, 1, ..., \ell_n$. We set $\ell_n = cn^{1/5}$ for $p^{J_m(x_i)}$ and $\ell_n = (c+0.5)n^{1/5}$ for $q^{K_m(z_i)}$ and c is varied between 1 and 4. The larger c implies that the set of models $\mathcal{M}_n$ is bigger. Overall, the relative performance of four estimators is not sensitive to the choice of $\mathcal{M}_n$. The normalized RMSE of most estimators increases as c increases, while the RMSE of both Averaging-τ and Mallows-τ is close to that of infeasible optimal model selection in most ranges of the parameter space.

6 Empirical Examples

In this section, we apply the proposed methods to two empirical examples to illustrate the usefulness of the model selection and model averaging in nonparametric IV estimation. The first example is about estimating the effect of class size on students' academic performance. The second example is about estimation of an Engel curve for food.

For both examples, we consider four estimators: (1) Horowitz's (2014) adaptive nonparametric IV estimation (Horowitz), (2) the nonparametric IV estimator with Sueishi's (2012) selection criterion (Sueishi), (3) the nonparametric IV Mallows model selection estimator with ρ_{mn}^{-1} estimated by $\hat{\tau}_{mn}$ (Mallows-τ), and (4) the nonparametric IV model averaging estimator with ρ_{mn}^{-1} estimated by

$\hat{\tau}_{mn}$ (Averaging-τ). The basis functions $p^{J_m(x_i)}$ and $q^{K_m(z_i)}$ are either orthonormal Legendre polynomials or a third order orthonormal B-spline. We consider approximating models with $J_m = K_m$ for Horowitz's adaptive estimator and consider all possible approximating models with $J_m \leq K_m$ for other estimators.

6.1 Effect of Class Size

There is a large literature on the studies of the relationship between school quality and students' performance. Many studies, however, find no strong evidence that improving school resources, such as student-teacher ratio, have an expected positive effect on students' performance on standardized achievement tests; see Hanushek (1986). These empirical results tend to counter the school policy that students' performance can be improved by allocating more money to them. Angrist and Lavy (1999) used Israeli public school data to study the effect of class size on test scores. They found that reducing class size induces a significant increase in test scores in most of the specifications of the linear models. Here we use their data to reexamine the effect of class size on students' performance in a flexible nonparametric framework.

For school s and class c, we consider the following model

$$y_{sc} = g(x_{sc}, d_{sc}) + \alpha_s + u_{sc} \tag{6.1}$$

$$\mathrm{E}(\alpha_s + u_{sc} | z_{sc}, d_{sc}) = 0 \tag{6.2}$$

where y_{sc} is the average reading comprehension test score in the class, x_{sc} is the number of students in class c of school s, z_{sc} is the instrumental variable, d_{sc} is the percentage of disadvantaged students in class c of school s, α_s is an unobserved school-specific effect, and u_{sc} is an unobserved random variable. Note that x_{sc} is a potentially endogenous variable since x_{sc} is not randomly assigned, and hence it may be correlated with other determinants and potential outcomes. Angrist and Lavy (1999) use Maimonides' rule on maximum number of students in a class to construct the instrumental variable. The instrument z_{sc} for the class size x_{sc} is

$$z_{sc} = e_s / \mathrm{int}(1 + (e_s - 1)/40)$$

where e_s is the enrollment in school s and the function $\mathrm{int}(n)$ is the largest integer less than or equal to n. The data set consists of observations of $2,049$ classes of fourth-grade students that were tested in 1991.[10]

Figure 9 shows the estimate of g as a function of the class size for d_{sc} less than or equal to 10 percent based on Legendre polynomials. Sueishi's selection estimator shows that increasing class size has no effect on reading comprehension test scores. Unlike Sueishi's selection estimator, Mallows-τ, Averaging-τ, and Horowitz's adaptive estimator support the conclusion drawn from the linear model that decreasing class size has a positive effect on reading comprehension test scores. One interesting observation is that the approximating model chosen by Mallows-τ is completely

[10] The data are available at http://economics.mit.edu/faculty/angrist/data1/data/anglavy99. See Angrist and Lavy (1999) for a detailed description of the data and their source.

different from those chosen by Averaging-τ. The model chosen by Mallows-τ is $(J_m, K_m) = (3, 3)$, while the models chosen by Averaging-τ are $(J_m, K_m) = (1, 1)$, $(2, 2)$, and $(4, 4)$ with weights 0.329, 0.291, and 0.380, respectively.

Figure 10 shows the estimate of g as a function of the class size for d_{sc} less than or equal to 10 percent based on B-splines. All four estimates of g are nonlinear and nonmonotonic. The results show that increasing class size, overall, has a negative effect on test scores. We, however, find that there is a positive effect of increasing class size when the class size is larger than 37. Comparing the estimates between Figures 9 and 10, we find that Sueishi's selection estimator is sensitive to the choice of basis functions, which is consistent with the finding in our simulations.

6.2 Estimation of an Engel Curve

The nonparametric instrumental variables estimation of Engel curves has been developed by Blundell, Chen, and Kristensen (2007). They find that the Engel curve for food changes significantly after taking the endogeneity into account. Here we apply the proposed model selection and model averaging estimators to investigate the robustness of Engel curve estimates with respect to the choices of the regularization parameter and the smoothing parameter.

The model is (2.1)–(2.2) where y_i is a household's expenditure share on food, x_i is a household's total expenditure, z_i is a household's gross earning, and g is an Engel curve. The data consist of observations of $1,655$ households from the British Family Expenditure Survey.

Figure 11 shows the estimate of an Engel curve for food based on Legendre polynomials. The estimates of the Engel curve of all four estimators are linear. Horowitz's adaptive estimator, Mallows-τ, and Averaging-τ choose the approximating model $(J_m, K_m) = (2, 2)$, and Sueishi's selection estimator chooses the approximating model $(J_m, K_m) = (2, 3)$. This result is similar to the finding of Horowitz (2014) in which the orthonormal Legendre polynomials are used.

Figure 12 shows the estimate of an Engel curve for food based on B-splines. Both Mallows-τ and Averaging-τ put the whole weight on the model $(J_m, K_m) = (4, 9)$, and the estimate of the Engel curve is close to linear. Horowitz's adaptive estimator chooses the approximating model $(J_m, K_m) = (4, 4)$, and the estimate of the Engel curve is also close to linear. However, the estimate of Engel curves based on Sueishi's selection estimator is nonlinear and nonmonotonic. Comparing Figures 11 and 12, it shows that the estimates of proposed model selection and model averaging methods are relatively unaffected by the choice of basis functions.

7 Conclusion

This paper considers model selection and model averaging in nonparametric instrumental variables estimation. We propose a simple Mallows' C_p-type criterion to simultaneously select the regularization parameter and the smoothing parameter. We show that our criterion is asymptotically optimal in the sense of achieving the lowest possible mean squared error among all candidates. We also introduce a new nonparametric instrumental variables averaging estimator and demonstrate its

asymptotic optimality. Simulation results show that the proposed data-driven approaches achieve lower root mean squared error than other existing methods.

One important limitation of our results is that we restrict to the case of homoskedastic errors. As pointed out by Andrews (1991), the Mallows criterion is not optimal under heteroskedasticity. This implies the optimality of proposed methods will similarly fail under heteroskedasticity. To overcome this limitation, it is possible to consider the jackknife model averaging estimator proposed by Hansen and Racine (2012) or the Heteroskedasticity-Robust C_p estimator proposed by Liu and Okui (2013) for nonparametric IV models with heteroskedastic errors. Another possible but more challenging extension is to investigate the important problem of inference after model selection and averaging. The existing studies show that the asymptotic distributions of post model selection and averaging estimators are nonstandard and cannot be approximated by simulation; see Hjort and Claeskens (2003) and Leeb and Pötscher (2005). Thus, the traditional confidence interval based on normal approximations leads to distorted inference. It would be an important research topic to consider constructing valid confidence intervals after model selection and model averaging in nonparametric IV models.

Appendix

This appendix contains three parts. Appendix A contains the proofs of the main theorems. Appendix B provides some useful moment bounds on the estimated matrices. Appendix C contains all figures. Let C denote a generic constant that may be different in different uses. The relation $a_n \lesssim b_n$ means there exists a finite positive C such that $a_n \leq Cb_n$ for all n large enough.

A Proofs of main results

Proof of Theorem 1: The proof of Theorem 1 is an application of Theorem 2 of Whittle (1960). Observe that

$$C_n(m) - L_n(m) = \frac{1}{n}e'e + \frac{2}{n}e'(I - \hat{D}_m)r_m + \frac{2}{n}\sigma^2\rho_{mn}^{-1}\sqrt{J_m K_m} - \frac{2}{n}e'\hat{D}_m e. \tag{A.1}$$

Note that the term $e'e$ doesn't depend on m. Thus, Theorem 1 is valid if the following hold:

$$\sup_{m\in\mathcal{M}_n} |e'(I - \hat{D}_m)r_m|/(nR_n(m)) \xrightarrow{p} 0. \tag{A.2}$$

$$\sup_{m\in\mathcal{M}_n} |\sigma^2\rho_{mn}^{-1}\sqrt{J_m K_m} - e'\hat{D}_m e|/(nR_n(m)) \xrightarrow{p} 0. \tag{A.3}$$

$$\sup_{m\in\mathcal{M}_n} |L_n(m)/R_n(m) - 1| \xrightarrow{p} 0. \tag{A.4}$$

We first consider (A.2). Using the triangle inequality, we have

$$|e'(I - \hat{D}_m)r_m|/(nR_n(m)) \leq |e'D_m r_m - e'\hat{D}_m r_m|/(nR_n(m)) + |e'(I - D_m)r_m|/(nR_n(m)).$$

In Lemma 2 (iii), we show that the supremum of the first term is $o_p(1)$. For the second term, we have

$$\begin{aligned}
&P\left(\sup_{m\in\mathcal{M}_n} |e'(I - D_m)r_m|/(nR_n(m)) > \delta|Z\right) \\
&\leq \sum_{m\in\mathcal{M}_n} P\left(|e'(I - D_m)r_m|/(nR_n(m)) > \delta|Z\right) \\
&\leq \sum_{m\in\mathcal{M}_n} \frac{\mathrm{E}\left(|e'(I - D_m)r_m|^{2(N+1)}|Z\right)}{\delta^{2(N+1)}(nR_n(m))^{2(N+1)}} \\
&\leq \frac{C}{\delta^{2(N+1)}} \sum_{m\in\mathcal{M}_n} \frac{\left(\phi_m^2 tr((I - D_m)'(I - D_m))\right)^{N+1}}{(nR_n(m))^{2(N+1)}} \\
&\leq \frac{C}{\delta^{2(N+1)}} \sum_{m\in\mathcal{M}_n} (nR_n(m))^{-(N+1)}
\end{aligned}$$

where the first inequality holds by Boole's inequality, the second inequality holds by Markov's inequality, the third inequality holds by the fact that $\mathrm{E}(r_m'(I - D_m)'(I - D_m)r_m|Z) = \phi_m^2 tr((I - D_m)'(I - D_m))$ and Theorem 2 of Whittle (1960), and the fourth inequality holds by the fact that

$nR_n(m) \geq \phi_m^2 tr((I-D_m)'(I-D_m))$. Then by Assumption 6, the supremum of the second term is $o_p(1)$. Thus, we obtain (A.2).

We next consider (A.3). Note that $\mathrm{E}(e'D_m e|Z) = \sigma^2 tr(D_m)$. Using the triangle inequality, we have

$$\begin{aligned}|\sigma^2\rho_{mn}^{-1}\sqrt{J_mK_m} - e'\hat{D}_m e|/(nR_n(m)) \leq\ & |\sigma^2\rho_{mn}^{-1}\sqrt{J_mK_m} - \sigma^2 tr(D_m)|/(nR_n(m)) \\ & + |\mathrm{E}(e'D_m e|Z) - e'D_m e|/(nR_n(m)) \\ & + |e'D_m e - e'\hat{D}_m e|/(nR_n(m)). \end{aligned} \tag{A.5}$$

For the first term, observe that $tr(D_m) \leq n\lambda_{\max}(D_m) \leq n\sigma_{\max}(D_m) = n\|D_m\|$. We pre and post multiply the terms by $Q_{x,m}^{-1/2}Q_{x,m}^{1/2}$ to obtain

$$\begin{aligned}\|D_m\| &= \|F_m(\Gamma_m Q_{z,m}^{-}\Gamma_m')^{-}\Gamma_m Q_{z,m}^{-}Z_m'/n\| \\ &= \|F_m Q_{x,m}^{-1/2}(Q_{x,m}^{-1/2}\Gamma_m Q_{z,m}^{-}\Gamma_m' Q_{x,m}^{-1/2})^{-}Q_{x,m}^{-1/2}\Gamma_m Q_{z,m}^{-}Z_m'/n\| \\ &= \|F_m Q_{x,m}^{-1/2}(Q_{x,m}^{-1/2}\Gamma_m Q_{z,m}^{-1/2}Q_{z,m}^{-1/2}\Gamma_m' Q_{x,m}^{-1/2})^{-}Q_{x,m}^{-1/2}\Gamma_m Q_{z,m}^{-1/2}Q_{z,m}^{-1/2}Z_m'/n\| \\ &\leq n\|F_m Q_{x,m}^{-1/2}/n\|\|(Q_{x,m}^{-1/2}\Gamma_m Q_{z,m}^{-1/2}Q_{z,m}^{-1/2}\Gamma_m' Q_{x,m}^{-1/2})^{-}Q_{x,m}^{-1/2}\Gamma_m Q_{z,m}^{-1/2}\|\|Q_{z,m}^{-1/2}Z_m'/n\| \\ &= O_p(\rho_{mn}^{-1}\sqrt{J_mK_m})\end{aligned}$$

where the last equality holds by the fact that $\|(Q_{x,m}^{-1/2}\Gamma_m Q_{z,m}^{-1/2}Q_{z,m}^{-1/2}\Gamma_m' Q_{x,m}^{-1/2})^{-}Q_{x,m}^{-1/2}\Gamma_m Q_{z,m}^{-1/2}\| = O_p(\rho_{mn}^{-1})$ and Lemma 1. Then, by the properties of the matrix spectral norm, we have

$$nR_n(m) \geq \sigma^2 tr(D_m'D_m) \geq \sigma^2\lambda_{\max}(D_m'D_m) = \sigma^2\|D_m\|^2 = O_p(\rho_{mn}^{-2}J_mK_m).$$

Thus, we have $|\sigma^2\rho_{mn}^{-1}\sqrt{J_mK_m} - \sigma^2 tr(D_m)|/(nR_n(m)) \lesssim n\rho_{mn}(J_mK_m)^{-1/2}$ for all $m \in \mathcal{M}_n$. The supremum of the first term of (A.5) is $o_p(1)$ by Assumption 3 (ii) and Assumption 3 (v).

For the second term of (A.5), by Boole's inequality, Markov's inequality, Theorem 2 of Whittle (1960), $nR_n(m) \geq \sigma^2 tr(D_m'D_m)$, and Assumption 6, we have

$$\begin{aligned}&P\left(\sup_{m\in\mathcal{M}_n}|e'D_m e - \mathrm{E}(e'D_m e|Z)|/(nR_n(m)) > \delta|Z\right) \\ &\leq \sum_{m\in\mathcal{M}_n} P\left(|e'D_m e - \mathrm{E}(e'D_m e|Z)|/(nR_n(m)) > \delta|Z\right) \\ &\leq \sum_{m\in\mathcal{M}_n}\frac{\mathrm{E}\left(|e'D_m e - \mathrm{E}(e'D_m e|Z)|^{2(N+1)}|Z\right)}{\delta^{2(N+1)}(nR_n(m))^{2(N+1)}} \\ &\leq \frac{C}{\delta^{2(N+1)}}\sum_{m\in\mathcal{M}_n}\frac{\left(\sigma^2 tr(D_m'D_m)\right)^{N+1}}{(nR_n(m))^{2(N+1)}} \\ &\leq \frac{C}{\delta^{2(N+1)}}\sum_{m\in\mathcal{M}_n}(nR_n(m))^{-(N+1)} \to 0.\end{aligned}$$

In Lemma 2 (i), we show that the supremum of the third term of (A.5) is also $o_p(1)$. Thus, we obtain (A.3).

We now consider (A.4). Note that $g - \hat{g}_m = (I - \hat{D}_m)g - \hat{D}_m e = (I - \hat{D}_m)r_m - \hat{D}_m e$ since $\hat{D}_m \mathcal{X}_m = \mathcal{X}_m$. Thus, the loss function can be written as

$$L_n(m) = \frac{1}{n} r_m'(I - \hat{D}_m)'(I - \hat{D}_m) r_m - \frac{2}{n} r_m'(I - \hat{D}_m)' \hat{D}_m e + \frac{1}{n} e' \hat{D}_m' \hat{D}_m e.$$

Therefore, we have

$$\begin{aligned} L_n(m) - R_n(m) = \frac{1}{n} \left(r_m'(I - \hat{D}_m)'(I - \hat{D}_m) r_m - \phi_m^2 tr((I - D_m)'(I - D_m)) \right) \\ - \frac{2}{n} e' \hat{D}_m'(I - \hat{D}_m) r_m + \frac{1}{n} \left(e' \hat{D}_m' \hat{D}_m e - \sigma^2 tr(D_m' D_m) - \mathrm{E}(e \tilde{D}_m' u_m' u_m \tilde{D}_m e | Z) \right). \end{aligned}$$

Recall that $\mathrm{E}(e' D_m' D_m e | Z) = \sigma^2 tr(D_m' D_m)$ and $\mathrm{E}(r_m'(I - D_m)'(I - D_m) r_m | Z) = \phi_m^2 tr((I - D_m)'(I - D_m))$. Thus, (A.4) is valid if the following hold:

$$\sup_{m \in \mathcal{M}_n} |r_m'(I - \hat{D}_m)'(I - \hat{D}_m) r_m - r_m'(I - D_m)'(I - D_m) r_m| / (n R_n(m)) \xrightarrow{p} 0. \tag{A.6}$$

$$\sup_{m \in \mathcal{M}_n} |r_m'(I - D_m)'(I - D_m) r_m - \mathrm{E}(r_m'(I - D_m)'(I - D_m) r_m | Z)| / (n R_n(m)) \xrightarrow{p} 0. \tag{A.7}$$

$$\sup_{m \in \mathcal{M}_n} |e' \hat{D}_m'(I - \hat{D}_m) r_m| / (n R_n(m)) \xrightarrow{p} 0. \tag{A.8}$$

$$\sup_{m \in \mathcal{M}_n} |e' \hat{D}_m' \hat{D}_m e - e' D_m' D_m e - \mathrm{E}(e' \tilde{D}_m' u_m' u_m \tilde{D}_m e | Z)| / (n R_n(m)) \xrightarrow{p} 0. \tag{A.9}$$

$$\sup_{m \in \mathcal{M}_n} |e' D_m' D_m e - \mathrm{E}(e' D_m' D_m e | Z)| / (n R_n(m)) \xrightarrow{p} 0. \tag{A.10}$$

We first show (A.6). Using the triangle inequality, we have

$$\begin{aligned} &|r_m'(I - \hat{D}_m)'(I - \hat{D}_m) r_m - r_m'(I - D_m)'(I - D_m) r_m| / (n R_n(m)) \\ &\quad \leq 2 |r_m'(\hat{D}_m - D_m) r_m| / (n R_n(m)) + |r_m'(\hat{D}_m' \hat{D}_m - D_m' D_m) r_m| / (n R_n(m)) \end{aligned}$$

In Lemma 2 (ii) and (v), we show that the supremum of both terms are $o_p(1)$. Thus, we obtain (A.6).

We next show (A.7). Note that $tr((I - D_m)'(I - D_m)(I - D_m)'(I - D_m)) \leq \lambda_{\max}((I - D_m)'(I - D_m)) tr((I - D_m)'(I - D_m)) \lesssim n R_n(m)$ by Assumption 5 (ii). Thus, by Boole's inequality, Markov's inequality, Theorem 2 of Whittle (1960), and Assumption 6, we have

$$\begin{aligned} &P\left(\sup_{m \in \mathcal{M}_n} |r_m'(I - D_m)'(I - D_m) r_m - \mathrm{E}(r_m'(I - D_m)'(I - D_m) r_m | Z)| / (n R_n(m)) > \delta | Z \right) \\ &\leq \sum_{m \in \mathcal{M}_n} P\left(|r_m'(I - D_m)'(I - D_m) r_m - \mathrm{E}(r_m'(I - D_m)'(I - D_m) r_m | Z)| / (n R_n(m)) > \delta | Z \right) \\ &\leq \sum_{m \in \mathcal{M}_n} \frac{\mathrm{E}\left(|r_m'(I - D_m)'(I - D_m) r_m - \mathrm{E}(r_m'(I - D_m)'(I - D_m) r_m | Z)|^{2(N+1)} | Z \right)}{\delta^{2(N+1)} (n R_n(m))^{2(N+1)}} \\ &\leq \frac{C}{\delta^{2(N+1)}} \sum_{m \in \mathcal{M}_n} \frac{(tr((I - D_m)'(I - D_m)(I - D_m)'(I - D_m)))^{N+1}}{(n R_n(m))^{2(N+1)}} \\ &\leq \frac{C}{\delta^{2(N+1)}} \sum_{m \in \mathcal{M}_n} (n R_n(m))^{-(N+1)} \to 0. \end{aligned}$$

We now consider (A.8). Observe that

$$|e'\hat{D}_m'(I-\hat{D}_m)r_m|/(nR_n(m)) \leq |e'\hat{D}_m'r_m - e'D_m'r_m|/(nR_n(m)) + |e'D_m'(I-D_m)r_m|/(nR_n(m)) + |e'D_m'D_mr_m - e'\hat{D}_m'\hat{D}_mr_m|/(nR_n(m)).$$

The supremum of the first term is $o_p(1)$ by Lemma 2 (iii). For the second term, note that $\mathrm{E}(r_m'(I-D_m)'D_mD_m'(I-D_m)r_m|Z) = \phi_m^2 tr((I-D_m)'D_mD_m'(I-D_m)) \leq \phi_m^2\lambda_{\max}(D_m'D_m)tr((I-D_m)'(I-D_m)) \lesssim nR_n(m)$ by Assumption 5 (i). Thus, by Boole's inequality, Markov's inequality, Theorem 2 of Whittle (1960), and Assumption 6, we have

$$P\left(\sup_{m\in\mathcal{M}_n}|e'D_m'(I-D_m)r_m|/(nR_n(m)) > \delta|Z\right) \leq \frac{C}{\delta^{2(N+1)}}\sum_{m\in\mathcal{M}_n}\frac{\left(\phi_m^2\lambda_{\max}(D_m'D_m)tr((I-D_m)'(I-D_m))\right)^{N+1}}{(nR_n(m))^{2(N+1)}} \to 0.$$

Also, the supremum of the third term is $o_p(1)$ by Lemma 2 (vi). Thus, we obtain (A.8).

We now consider (A.9) and (A.10). In Lemma 2 (iv), we show (A.9). For (A.10), note that $tr(D_m'D_mD_m'D_m) \leq \lambda_{\max}(D_m'D_m)tr(D_m'D_m) \lesssim nR_n(m)$ by Assumption 5 (i). Thus, by Boole's inequality, Markov's inequality, Theorem 2 of Whittle (1960), and Assumption 6, we have

$$P\left(\sup_{m\in\mathcal{M}_n}|e'D_m'D_me - \mathrm{E}(e'D_m'D_me|Z)|/(nR_n(m)) > \delta|Z\right) \leq \frac{C}{\delta^{2(N+1)}}\sum_{m\in\mathcal{M}_n}\frac{\left(\sigma^2\lambda_{\max}(D_m'D_m)tr(D_m'D_m)\right)^{N+1}}{(nR_n(m))^{2(N+1)}} \to 0.$$

This completes the proof. ■

Proof of Theorem 2: Observe that

$$C_n(w) - L_n(w) = \frac{1}{n}e'e + \frac{2}{n}e'\hat{b}(w) + \frac{2\sigma^2}{n}\sum_{m=1}^{M}w_m\rho_{mn}^{-1}\sqrt{J_mK_m} - \frac{2}{n}e'\hat{D}(w)e. \tag{A.11}$$

Note that the term $e'e$ doesn't depend on w. Thus, Theorem 2 is valid if the following hold:

$$\sup_{w\in\mathcal{H}_n^*}|e'\hat{b}(w)|/(nR_n(w)) \xrightarrow{p} 0. \tag{A.12}$$

$$\sup_{w\in\mathcal{H}_n^*}|\sum_{m=1}^{M}w_m\sigma^2\rho_{mn}^{-1}\sqrt{J_mK_m} - e'\hat{D}(w)e|/(nR_n(w)) \xrightarrow{p} 0. \tag{A.13}$$

$$\sup_{w\in\mathcal{H}_n^*}|L_n(w)/R_n(w) - 1| \xrightarrow{p} 0. \tag{A.14}$$

Note that Equations (A.12), (A.13), and (A.14) correspond to (A.2), (A.3), and (A.4).

We first consider (A.12). We show (A.12) by a similar argument to the proof of (A.2) with Assumption 6 replaced by Assumption 8. Recall that $\hat{b}(w) = \sum_{m=1}^{M}w_m\hat{b}_m$ and $\hat{b}_m = (I-\hat{D}_m)r_m$.

Define $b_m = (I - D_m)r_m$. Using the triangle inequality, we have

$$\begin{aligned}|e'\hat{b}(w)|/(nR_n(w)) &\leq |e'b(w) - e'\hat{b}(w)|/(nR_n(w)) + |e'b(w)|/(nR_n(w)) \\ &\leq \sum_{m=1}^{M} w_m |e'(D_m - \hat{D}_m)r_m|/(nR_n(w)) + |e'b(w)|/(nR_n(w)).\end{aligned}$$

The supremum of the first term is $o_p(1)$ by Lemma 2 (iii) and Assumption 8. For the second term, we have

$$\begin{aligned}P\left(\sup_{w\in\mathcal{H}_n^*} |e'b(w)|/(nR_n(w)) > \delta|Z\right) &\leq \sum_{w\in\mathcal{H}_n^*} P\left(|e'b(w)|/(nR_n(w)) > \delta|Z\right) \\ &\leq \sum_{w\in\mathcal{H}_n^*} \frac{\mathrm{E}\left(|e'b(w)|^{2(N+1)}|Z\right)}{\delta^{2(N+1)}(nR_n(w))^{2(N+1)}} \\ &\leq \frac{C}{\delta^{2(N+1)}} \sum_{w\in\mathcal{H}_n^*} (nR_n(w))^{-(N+1)}\end{aligned}$$

where the first inequality holds by Boole's inequality, the second inequality holds by Markov's inequality, the third inequality holds by Theorem 2 of Whittle (1960) and the fact that $nR_n(w) \geq b(w)'b(w)$. Then by Assumption 8, the supremum of the second term is $o_p(1)$. Thus, we obtain (A.12).

We next consider (A.13). Using the triangle inequality, we have

$$\left|\sum_{m=1}^{M} w_m \sigma^2 \rho_{mn}^{-1} \sqrt{J_m K_m} - e'\hat{D}(w)e\right| \leq \sum_{m=1}^{M} w_m \left|\sigma^2 \rho_{mn}^{-1} \sqrt{J_m K_m} - e'\hat{D}_m e\right|.$$

Thus, we have (A.13) by Equation (A.3) and Assumption 8.

We now consider (A.14). Define $D(w) = \sum_{m=1}^{M} w_m D_m$. By the inequality $\sigma_{\max}(A+B) \leq \sigma_{\max}(A) + \sigma_{\max}(B)$ and Assumption 5, we have

$$\sigma_{\max}(D(w)) \leq \sum_{m=1}^{M} w_m \sigma_{\max}(D_m) < \infty$$

uniformly in $w \in \mathcal{H}_n^*$, almost surely as $n \to \infty$. This shows

$$\limsup_{n\to\infty} \sup_{w\in\mathcal{H}_n^*} \sigma_{\max}(D(w)) < \infty, \tag{A.15}$$

$$\limsup_{n\to\infty} \sup_{w\in\mathcal{H}_n^*} \sigma_{\max}(I - D(w)) < \infty, \tag{A.16}$$

which correspond to Assumption 5 (i) and (ii).

Note that $g - \hat{g}(w) = g - \hat{D}(w)y = (I - \hat{D}(w))g - \hat{D}(w)e = \hat{b}(w) - \hat{D}(w)e$. Thus, the loss

function can be written as

$$\begin{aligned}L_n(w) &= \frac{1}{n}\hat{b}(w)'\hat{b}(w) - \frac{2}{n}\hat{b}(w)'\hat{D}(w)e + \frac{1}{n}e'\hat{D}(w)'\hat{D}(w)e. \\ &= \frac{1}{n}\sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell r_m'(I-\hat{D}_m)'(I-\hat{D}_\ell)r_\ell - \frac{2}{n}\sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell r_m'(I-\hat{D}_m)'\hat{D}_\ell e \\ &\quad + \frac{1}{n}\sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell e'\hat{D}_m'\hat{D}_\ell e.\end{aligned}$$

Therefore, we have

$$\begin{aligned}L_n(w) - R_n(w) = \frac{1}{n}\sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell \big(r_m'(I-\hat{D}_m)'(I-\hat{D}_\ell)r_\ell - \phi_{m,\ell} tr((I-D_m)'(I-D_\ell))\big) \\ - \frac{2}{n}\sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell r_m'(I-\hat{D}_m)'\hat{D}_\ell e \\ + \frac{\sigma^2}{n}\sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell \big(e'\hat{D}_m'\hat{D}_\ell e - tr(D_m'D_\ell) - \mathrm{E}(e'\tilde{D}_m'u_m'u_\ell\tilde{D}_\ell e|Z)\big).\end{aligned}$$

By Equations (A.15), (A.16), Assumption 8, and a similar argument to the proof of (A.6)–(A.10), we have (A.14). This completes the proof. ■

Proof of Theorem 3: The proof is similar to that of Theorem 2 of Hansen and Racine (2012). We will verify that Assumption 8 holds almost surely, conditional on Z. Let $P_m = J_m + K_m$ be the sum of the number of explanatory variables and instruments for the mth model. Without loss of generality, arrange the models so that they are weakly ordered by P_m, i.e., $P_1 \leq P_2 \leq \cdots \leq P_M$. As in the proof of Theorem 1 of Hansen (2007), for integers $1 \leq l_1 \leq l_2 \leq \cdots \leq l_N \leq M$, let $w_{l_1,l_2,...,l_N}$ be the weight vector that sets $w_{lk} = 1/N$ for $k = 1, ..., N$, and the remainder zero.

Recall the definition of ξ_n and $\bar{q}_n$. Pick a sequence $\psi_n \to \infty$ that satisfies $\psi_n = o\left(\xi_n^{1+1/N}\right)$ yet $\bar{q}_n^{1+N} = o(\psi_n)$, which is feasible since $\xi_n \to \infty$ and $\bar{q}_n^{1+N} = o\left(\xi_n^{1+1/N}\right)$ under Assumption 9. We then have

$$\sum_{w\in\mathcal{H}_n^*}(nR_n(w))^{-(N+1)} = \sum_{l_N=1}^{M}\sum_{l_{N-1}=1}^{l_N}\cdots\sum_{l_1=1}^{l_2}(nR_n(w_{l_1,l_2,...,l_N}))^{-(N+1)} \leq S_{1n} + S_{2n}$$

where

$$S_{1n} = \sum_{l_N=1}^{\psi_n}\sum_{l_{N-1}=1}^{l_N}\cdots\sum_{l_1=1}^{l_2}(nR_n(w_{l_1,l_2,...,l_N}))^{-(N+1)} \tag{A.17}$$

$$S_{2n} = \sum_{l_N=\psi_n+1}^{\infty}\sum_{l_{N-1}=1}^{l_N}\cdots\sum_{l_1=1}^{l_2}(nR_n(w_{l_1,l_2,...,l_N}))^{-(N+1)}. \tag{A.18}$$

We first consider S_{1n}. Since S_{1n} has fewer than ψ_n^N elements, we use the bound $nR_n(w) \geq \xi_n$ from Assumption 9 (i) and find that $S_{1n} \leq \psi_n^N \xi_n^{-(N+1)} \to 0$ as $n \to \infty$.

Next consider S_{2n}. We use the simple bound

$$\begin{aligned} nR_n(w_{l_1,l_2,\ldots,l_N}) &\geq \sigma^2 tr(D(w_{l_1,l_2,\ldots,l_N})'D(w_{l_1,l_2,\ldots,l_N})) \\ &= \frac{\sigma^2}{N^2}\left(\sum_{m=1}^{N}\sum_{n=1}^{N} tr(D_{l_m}'D_{l_n})\right) \\ &\geq \frac{\sigma^2}{N^2}\sum_{m=1}^{N} tr(D_{l_m}'D_{l_m}) \\ &\geq \frac{\sigma^2}{N^2}\sum_{m=1}^{N} J_{l_m}K_{l_m} \\ &\geq \frac{\sigma^2}{N^2}J_{l_N}^2 \\ &\geq \frac{\sigma^2}{N^2}\frac{l_N^2}{\bar{q}_n^2} \end{aligned}$$

where the first inequality uses $nR_n(w) \geq \sigma^2 tr(D(w)'D(w))$, the following equality uses the definitions of $D(w)$ and $w_{l_1,l_2,\ldots,l_N}$, the second inequality uses $tr(D_{l_m}'D_{l_n}) \geq 0$, the third inequality uses $tr(D_m'D_m) = O_p(\rho_{mn}^{-2}J_mK_m)$ and $\rho_{mn}^{-1} \geq 1$, the fourth inequality uses $J_m \leq K_m$, and the final inequality follows from the definition of $\bar{q}_n$ and the ordering of the models by the number of parameters P_m. Therefore, we have

$$\begin{aligned} S_{2n} &\leq \sum_{l_N=\psi_n+1}^{\infty}\sum_{l_{N-1}=1}^{l_N}\cdots\sum_{l_1=1}^{l_2}\left(\frac{\sigma^2}{N^2}\frac{l_N^2}{\bar{q}_n^2}\right)^{-(N+1)} \\ &\leq \frac{N^{2(N+1)}}{\sigma^{2(N+1)}}\bar{q}_n^{2(N+1)}\sum_{l_N=\psi_n+1}^{\infty} l_N^{-4} \\ &\leq \frac{N^{2(N+1)}}{\sigma^{2(N+1)}}\bar{q}_n^{2(N+1)}\psi_n^{-2} \\ &\leq o(1). \end{aligned}$$

This completes the proof. ■

B Supplementary lemmas and their proofs

Lemma 1. *Suppose Assumptions 1–3 hold. Define* $\hat{\Xi}_m = Q_{x,m}^{1/2}(\hat{\Gamma}_m\hat{Q}_{z,m}^{-}\hat{\Gamma}_m')^{-}\hat{\Gamma}_m\hat{Q}_{z,m}^{-}Q_{z,m}^{1/2}$ *and* $\Xi_m = Q_{x,m}^{1/2}(\Gamma_m Q_{z,m}^{-}\Gamma_m')^{-}\Gamma_m Q_{z,m}^{-1/2}$. *Let* $\Psi_{mn} = \zeta_m\sqrt{(\log K_m)/n}$ *and* $\rho_{mn}^{-1}\Psi_{mn} = o(1)$ *for all* m. *Then we have*

(i) $\|(F_m'F_m)^{-1/2}F_m'e/n\| = O_p(\sqrt{J_m/n})$.

(ii) $\|(\mathcal{Z}_m'\mathcal{Z}_m)^{-1/2}\mathcal{Z}_m'e/n\| = O_p(\sqrt{K_m/n})$.

(iii) $\|\hat{\Xi}_m - \Xi_m\| = O_p\left(\rho_{mn}^{-2}\Psi_{mn}\right)$.

Proof of Lemma 1: Part (i) and (ii) are implied by Markov's inequality. Define $\hat{Q}^o_{x,m} = Q^{-1/2}_{x,m}\hat{Q}_{x,m}Q^{-1/2}_{x,m}$, $\hat{Q}^o_{z,m} = Q^{-1/2}_{z,m}\hat{Q}_{z,m}Q^{-1/2}_{z,m}$, and $\hat{\Gamma}^o_m = Q^{-1/2}_{x,m}\hat{\Gamma}_m Q^{-1/2}_{z,m}$. Let $Q^o_{x,m} = I_{J_m}$, $Q^o_{z,m} = I_{K_m}$, and Γ^o_m be their respective expected values. For part (iii), we pre and post multiply the terms by $Q^{-1/2}_{z,m}Q^{1/2}_{z,m}$ and $Q^{-1/2}_{x,m}Q^{1/2}_{x,m}$ to obtain

$$
\begin{aligned}
&\|\hat{\Xi}_m - \Xi_m\| \\
&= \|Q^{1/2}_{x,m}(\hat{\Gamma}_m\hat{Q}^-_{z,m}\hat{\Gamma}'_m)^-\hat{\Gamma}_m\hat{Q}^-_{z,m}Q^{1/2}_{z,m} - Q^{1/2}_{x,m}(\Gamma_m Q^-_{z,m}\Gamma'_m)^-\Gamma_m Q^{-1/2}_{z,m}\| \\
&= \|(Q^{-1/2}_{x,m}\hat{\Gamma}_m Q^{-1/2}_{z,m}(Q^{-1/2}_{z,m}\hat{Q}_{z,m}Q^{-1/2}_{z,m})^{-1}Q^{-1/2}_{z,m}\hat{\Gamma}'_m Q^{-1/2}_{x,m})^{-1}Q^{-1/2}_{x,m}\hat{\Gamma}_m Q^{-1/2}_{z,m}(Q^{-1/2}_{z,m}\hat{Q}_{z,m}Q^{-1/2}_{z,m})^{-1} \\
&\quad - (Q^{-1/2}_{x,m}\Gamma_m Q^{-1/2}_{z,m}Q^{-1/2}_{z,m}\Gamma'_m Q^{-1/2}_{x,m})^{-1}Q^{-1/2}_{x,m}\Gamma_m Q^{-1/2}_{z,m}\| \\
&= \|(\hat{\Gamma}^o_m\hat{Q}^{o-1}_{z,m}\hat{\Gamma}^{o\prime}_m)^{-1}\hat{\Gamma}^o_m\hat{Q}^{o-1}_{z,m} - (\Gamma^o_m\Gamma^{o\prime}_m)^{-1}\Gamma^o_m\| \\
&\leq \|(\hat{\Gamma}^o_m\hat{Q}^{o-1}_{z,m}\hat{\Gamma}^{o\prime}_m)^{-1}\hat{\Gamma}^o_m\hat{Q}^{o-1/2}_{z,m}(\hat{Q}^{o-1/2}_{z,m} - I_{K_m})\| + \|(\hat{\Gamma}^o_m\hat{Q}^{o-1}_{z,m}\hat{\Gamma}^{o\prime}_m)^{-1}\hat{\Gamma}^o_m\hat{Q}^{o-1/2}_{z,m} - (\Gamma^o_m\Gamma^{o\prime}_m)^{-1}\Gamma^o_m\| \\
&\leq \|(\hat{\Gamma}^o_m\hat{Q}^{o-1}_{z,m}\hat{\Gamma}^{o\prime}_m)^{-1}\hat{\Gamma}^o_m\hat{Q}^{o-1/2}_{z,m} - (\Gamma^o_m\Gamma^{o\prime}_m)^{-1}\Gamma^o_m\|\|\hat{Q}^{o-1/2}_{z,m} - I_{K_m}\| \\
&\quad + \|(\Gamma^o_m\Gamma^{o\prime}_m)^{-1}\Gamma^o_m\|\|\hat{Q}^{o-1/2}_{z,m} - I_{K_m}\| + \|(\hat{\Gamma}^o_m\hat{Q}^{o-1}_{z,m}\hat{\Gamma}^{o\prime}_m)^{-1}\hat{\Gamma}^o_m\hat{Q}^{o-1/2}_{z,m} - (\Gamma^o_m\Gamma^{o\prime}_m)^{-1}\Gamma^o_m\| \qquad \text{(B.1)}
\end{aligned}
$$

Following a similar argument to the proof of Lemma E.10 in Chen and Christensen (2015), we have

$$
\begin{aligned}
&\|\Xi_m\| = \|(\Gamma^o_m\Gamma^{o\prime}_m)^{-1}\Gamma^o_m\| = O_p(\rho^{-1}_{mn}) && \text{(B.2)} \\
&\|\hat{Q}^{o-1/2}_{z,m} - I_{K_m}\| = O_p(\zeta_{z,m}\sqrt{(\log K_m)/n}) && \text{(B.3)} \\
&\|(\hat{\Gamma}^o_m\hat{Q}^{o-1}_{z,m}\hat{\Gamma}^{o\prime}_m)^{-1}\hat{\Gamma}^o_m\hat{Q}^{o-1/2}_{z,m} - (\Gamma^o_m\Gamma^{o\prime}_m)^{-1}\Gamma^o_m\| = O_p(\rho^{-2}_{mn}\zeta_m\sqrt{(\log K_m)/n}) && \text{(B.4)}
\end{aligned}
$$

where $\zeta_m = \max(\zeta_{x,m}, \zeta_{z,m})$. Thus, the result follows by substituting (B.2), (B.3), and (B.4) into (B.1). This completes the proof. ■

Lemma 2. *Suppose Assumptions 1–6 hold. Then we have*

(i) $\sup_{m\in\mathcal{M}_n} |e'(\hat{D}_m - D_m)e|/(nR_n(m)) \to 0.$

(ii) $\sup_{m\in\mathcal{M}_n} |r'_m(\hat{D}_m - D_m)r_m|/(nR_n(m)) \to 0.$

(iii) $\sup_{m\in\mathcal{M}_n} |e'(\hat{D}_m - D_m)r_m|/(nR_n(m)) \to 0.$

(iv) $\sup_{m\in\mathcal{M}_n} |e'\hat{D}'_m\hat{D}_m e - e'D'_m D_m e - \mathrm{E}(e'\tilde{D}'_m u'_m u_m \tilde{D}_m e|Z)|/(nR_n(m)) \to 0.$

(v) $\sup_{m\in\mathcal{M}_n} |r'_m(\hat{D}'_m\hat{D}_m - D'_m D_m)r_m|/(nR_n(m)) \to 0.$

(vi) $\sup_{m\in\mathcal{M}_n} |e'(\hat{D}'_m\hat{D}_m - D'_m D_m)r_m|/(nR_n(m)) \to 0.$

Proof of Lemma 2: Recall that $\mathcal{X}_m = F_m + u_m$, $\hat{D}_m = \mathcal{X}_m(\hat{\Gamma}_m\hat{Q}^-_{z,m}\hat{\Gamma}'_m)^-\hat{\Gamma}_m\hat{Q}^-_{z,m}\mathcal{Z}'_m/n$, $D_m = F_m(\Gamma_m Q^-_{z,m}\Gamma'_m)^-\Gamma_m Q^-_{z,m}Z'_m/n$, and $\tilde{D}_m = (\Gamma_m Q^-_{z,m}\Gamma'_m)^-\Gamma_m Q^-_{z,m}Z'_m/n$. Define $H_m = F'_m F_m/n$,

$G_m = (\Gamma_m Q^-_{z,m}\Gamma'_m)^-\Gamma_m Q^-_{z,m}$ and $\hat{G}_m = (\hat{\Gamma}_m \hat{Q}^-_{z,m}\hat{\Gamma}'_m)^-\hat{\Gamma}_m \hat{Q}^-_{z,m}$. We then rewrite $\hat{D}_m$, D_m, and $\tilde{D}_m$ as $\hat{D}_m = \mathcal{X}_m \hat{G}_m \mathcal{Z}'_m/n$, $D_m = F_m G_m \mathcal{Z}'_m/n$, and $\tilde{D}_m = G_m \mathcal{Z}'_m/n$, respectively.

Define $\hat{\Xi}_m = Q^{1/2}_{x,m}(\hat{\Gamma}_m \hat{Q}^-_{z,m}\hat{\Gamma}'_m)^-\hat{\Gamma}_m \hat{Q}^-_{z,m} Q^{1/2}_{z,m}$ and $\Xi_m = Q^{1/2}_{x,m}(\Gamma_m Q^-_{z,m}\Gamma'_m)^-\Gamma_m Q^{-1/2}_{z,m}$. Multiplying the terms by $Q^{-1/2}_{z,m}Q^{1/2}_{z,m}$ and $Q^{-1/2}_{x,m}Q^{1/2}_{x,m}$, we obtain

$$G_m = (\Gamma_m Q^-_{z,m}\Gamma'_m)^-\Gamma_m Q^-_{z,m} = Q^{-1/2}_{x,m}(Q^{1/2}_{x,m}(\Gamma_m Q^-_{z,m}\Gamma'_m)^-\Gamma_m Q^{-1/2}_{z,m})Q^{-1/2}_{z,m} = Q^{-1/2}_{x,m}\Xi_m Q^{-1/2}_{z,m}$$

and

$$\begin{aligned}
\hat{G}_m - G_m &= (\hat{\Gamma}_m \hat{Q}^-_{z,m}\hat{\Gamma}'_m)^-\hat{\Gamma}_m \hat{Q}^-_{z,m} - (\Gamma_m Q^-_{z,m}\Gamma'_m)^-\Gamma_m Q^-_{z,m} \\
&= \left((\hat{\Gamma}_m \hat{Q}^-_{z,m}\hat{\Gamma}'_m)^-\hat{\Gamma}_m \hat{Q}^-_{z,m} Q^{1/2}_{z,m} - (\Gamma_m Q^-_{z,m}\Gamma'_m)^-\Gamma_m Q^{-1/2}_{z,m}\right) Q^{-1/2}_{z,m} \\
&= Q^{-1/2}_{x,m}\left(Q^{1/2}_{x,m}(\hat{\Gamma}_m \hat{Q}^-_{z,m}\hat{\Gamma}'_m)^-\hat{\Gamma}_m \hat{Q}^-_{z,m} Q^{1/2}_{z,m} - Q^{1/2}_{x,m}(\Gamma_m Q^-_{z,m}\Gamma'_m)^-\Gamma_m Q^{-1/2}_{z,m}\right) Q^{-1/2}_{z,m} \\
&= Q^{-1/2}_{x,m}(\hat{\Xi}_m - \Xi_m)Q^{-1/2}_{z,m}.
\end{aligned}$$

For (i), note that

$$\begin{aligned}
|e'(\hat{D}_m - D_m)e| &= |e'\mathcal{X}_m \hat{G}_m \mathcal{Z}'_m e - e' F_m G_m \mathcal{Z}'_m e|/n \\
&\le |e' F_m(\hat{G}_m - G_m)\mathcal{Z}'_m e|/n + |e' u_m(\hat{G}_m - G_m)\mathcal{Z}'_m e|/n + |e' u_m G_m \mathcal{Z}'_m e|/n \\
&= |e' F_m Q^{-1/2}_{x,m}(\hat{\Xi}_m - \Xi_m)Q^{-1/2}_{z,m}\mathcal{Z}'_m e|/n + |e' u_m Q^{-1/2}_{x,m}(\hat{\Xi}_m - \Xi_m)Q^{-1/2}_{z,m}\mathcal{Z}'_m e|/n \\
&\quad + |e' u_m Q^{-1/2}_{x,m}\Xi_m Q^{-1/2}_{z,m}\mathcal{Z}'_m e|/n \\
&\equiv A_1 + A_2 + A_3.
\end{aligned}$$

By Assumption 2 and Lemma 1, we have

$$\begin{aligned}
A_1 &\le n\|e' F_m Q^{-1/2}_{x,m}/n\|\|\hat{\Xi}_m - \Xi_m\|\|Q^{-1/2}_{z,m}\mathcal{Z}'_m e/n\| = O_p(\rho^{-2}_{mn}\Psi_{mn}\sqrt{J_m K_m}), \\
A_2 &\le \|e' u_m Q^{-1/2}_{x,m}\|\|(\hat{\Xi}_m - \Xi_m)\|\|Q^{-1/2}_{z,m}\mathcal{Z}'_m e/n\| = O_p(\rho^{-2}_{mn}\Psi_{mn}\sqrt{J_m K^3_m/n})), \\
A_3 &\le \|e' u_m Q^{-1/2}_{x,m}\|\|\Xi_m\|\|Q^{-1/2}_{z,m}\mathcal{Z}'_m e/n\| = O_p(\rho^{-1}_{mn}\sqrt{J_m K^3_m/n}).
\end{aligned}$$

Recall that $nR_n(m) \ge \sigma^2 tr(D'_m D_m) \ge \sigma^2\|D_m\|^2 = O_p(\rho^{-2}_{mn} J_m K_m)$. Thus, by Assumption 3 (ii) and Assumption 4 (i), we have $|e'(\hat{D}_m - D_m)e|/(nR_n(m)) \lesssim \Psi_{mn}(J_m K_m)^{-1/2}$ for all $m \in \mathcal{M}_n$. Since $J_m, K_m \to \infty$, the result follows from Assumption 4 (i).

For (ii), note that $|r'_m(\hat{D}_m - D_m)r_m| = |tr((\hat{D}_m - D_m)(r_m r'_m))| \le |\lambda_{\max}(r_m r'_m)tr(\hat{D}_m - D_m)| \le |r'_m r_m||tr(\hat{D}_m - D_m)|$. Observe that

$$\begin{aligned}
\|\hat{D}_m - D_m\| &= \|\mathcal{X}_m \hat{G}_m \mathcal{Z}'_m - F_m G_m \mathcal{Z}'_m\|/n \\
&\le \|F_m(\hat{G}_m - G_m)\mathcal{Z}'_m\|/n + \|u_m(\hat{G}_m - G_m)\mathcal{Z}'_m\|/n + \|u_m G_m \mathcal{Z}'_m\|/n \\
&= \|F_m Q^{-1/2}_{x,m}(\hat{\Xi}_m - \Xi_m)Q^{-1/2}_{z,m}\mathcal{Z}'_m\|/n + \|u_m Q^{-1/2}_{x,m}(\hat{\Xi}_m - \Xi_m)Q^{-1/2}_{z,m}\mathcal{Z}'_m\|/n \\
&\quad + \|u_m Q^{-1/2}_{x,m}\Xi_m Q^{-1/2}_{z,m}\mathcal{Z}'_m\|/n \\
&\equiv B_1 + B_2 + B_3.
\end{aligned}$$

By Lemma 1, we have

$$B_1 \le n\|F_m Q_{x,m}^{-1/2}/n\|\|\hat{\Xi}_m - \Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m/n\| = O_p(\rho_{mn}^{-2}\Psi_{mn}\sqrt{J_m K_m}),$$
$$B_2 \le \|u_m Q_{x,m}^{-1/2}\|\|\hat{\Xi}_m - \Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| = O_p(\rho_{mn}^{-2}\Psi_{mn}\sqrt{J_m K_m^3/n}),$$
$$B_3 \le \|u_m Q_{x,m}^{-1/2}\|\|\Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| = O_p(\rho_{mn}^{-2}\sqrt{J_m K_m^3/n}).$$

Recall that $nR_n(m) \ge \sigma^2\|D_m\|^2 = O_p(\rho_{mn}^{-2}J_m K_m)$ and $\mathrm{E}\left(r_{mi}^2|z_i\right) = \phi_m^2 < \infty$ by Assumption 4 (ii). Therefore, we have $|r_m'(\hat{D}_m - D_m)r_m|/(nR_n(m)) \le |r_m'r_m||tr(\hat{D}_m - D_m)|/(nR_n(m)) \lesssim \Psi_{mn}(J_m K_m)^{-1/2}$ for all $m \in \mathcal{M}_n$. Thus, the result follows from Assumption 4 (i).

For (iii), by the Cauchy-Schwarz inequality and the trace inequality, we have

$$\begin{aligned}
|e'(\hat{D}_m - D_m)r_m| &\le (e'(\hat{D}_m - D_m)(\hat{D}_m - D_m)'e)^{1/2}(r_m'r_m)^{1/2} \\
&= tr((ee')(\hat{D}_m - D_m)(\hat{D}_m - D_m)')^{1/2}(r_m'r_m)^{1/2} \\
&\le \|ee'\|^{1/2}tr(\hat{D}_m - D_m)(\hat{D}_m - D_m)')^{1/2}(r_m'r_m)^{1/2} \\
&= \|ee'\|^{1/2}\|\hat{D}_m - D_m\|(r_m'r_m)^{1/2}.
\end{aligned}$$

Note that $nR_n(m) \ge \sigma^2\|D_m\|^2 = O_p(\rho_{mn}^{-2}J_m K_m)$, $\mathrm{E}\left(e_i^2|z_i\right) = \sigma^2 \le nR_n(m)$, $\mathrm{E}\left(r_{mi}^2|z_i\right) = \phi_m^2 \le nR_n(m)$, and $\|\hat{D}_m - D_m\| = O_p(\rho_{mn}^{-2}\Psi_{mn}\sqrt{J_m K_m})$ shown in (ii). Therefore, we have $|e'(\hat{D}_m - D_m)r_m|/(nR_n(m)) \lesssim \Psi_{mn}(J_m K_m)^{-1/2}$ for all $m \in \mathcal{M}_n$. Thus, the result follows from Assumption 4 (i).

For (iv), note that

$$\begin{aligned}
&|e'\hat{D}_m'\hat{D}_m e - e'D_m'D_m e - \mathrm{E}(e'\tilde{D}_m'u_m'u_m\tilde{D}_m e|Z)| \\
&\quad = |e'\mathcal{Z}_m\hat{G}_m'\mathcal{X}_m'\mathcal{X}_m\hat{G}_m\mathcal{Z}_m'e - e'\mathcal{Z}_m G_m'F_m'F_m G_m\mathcal{Z}_m'e - \mathrm{E}(e'\mathcal{Z}_m G_m'u_m'u_m G_m\mathcal{Z}_m'e|Z)|/n^2 \\
&\quad \le |e'\mathcal{Z}_m(\hat{G}_m'F_m'F_m\hat{G}_m - G_m'F_m'F_m G_m)\mathcal{Z}_m'e|/n^2 + 2|e'\mathcal{Z}_m\hat{G}_m'F_m'u_m\hat{G}_m\mathcal{Z}_m'e|/n^2 \\
&\qquad + |e'\mathcal{Z}_m\hat{G}_m'u_m'u_m\hat{G}_m\mathcal{Z}_m'e - \mathrm{E}(e'\mathcal{Z}_m G_m'u_m'u_m G_m\mathcal{Z}_m'e|Z)|/n^2 \\
&\quad \equiv D_1 + 2D_2 + D_3
\end{aligned}$$

Consider D_1 first. Observe that

$$\begin{aligned}
D_1 &= n|(e'\mathcal{Z}_m/n)(\hat{G}_m'F_m'F_m\hat{G}_m - G_m'F_m'F_m G_m)(\mathcal{Z}_m'e/n)| \\
&\le n|(e'\mathcal{Z}_m/n)(\hat{G}_m - G_m)'H_m(\hat{G}_m - G_m)(\mathcal{Z}_m'e/n)| \\
&\quad + n|(e'\mathcal{Z}_m/n)(\hat{G}_m - G_m)'H_m G_m(\mathcal{Z}_m'e/n)| \\
&\quad + n|(e'\mathcal{Z}_m/n)G_m'H_m(\hat{G}_m - G_m)(\mathcal{Z}_m'e/n)| \\
&\le n\|e'\mathcal{Z}_m Q_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}H_m Q_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'e/n\| \\
&\quad + n\|e'\mathcal{Z}_m Q_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}H_m Q_{x,m}^{-1/2}\Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'e/n\| \\
&\quad + n\|e'\mathcal{Z}_m Q_{z,m}^{-1/2}/n\|\|\Xi_m'Q_{x,m}^{-1/2}H_m Q_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'e/n\| \\
&\equiv D_{11} + D_{12} + D_{13}.
\end{aligned}$$

Recall that $\|Q_{z,m}^{-1/2}\mathcal{Z}_m'e/n\| = O_p(\sqrt{K_m/n})$, $\|\hat{\Xi}_m - \Xi_m\| = O_p\left(\rho_{mn}^{-2}\Psi_{mn}\right)$, $\|\Xi_m\| = O_p(\rho_{mn}^{-1})$, and $\|H_m\| = O(J_m)$. Therefore, we have $D_{11} = O_p(\rho_{mn}^{-4}\Psi_{mn}^2 J_m K_m)$, $D_{12} = O_p(\rho_{mn}^{-3}\Psi_{mn}J_m K_m)$,

and $D_{13} = O_p(\rho_{mn}^{-3}\Psi_{mn}J_mK_m)$. Thus, by Assumption 4 (i), we obtain $D_1 = O_p(\rho_{mn}^{-3}\Psi_{mn}J_mK_m)$. Note that $nR_n(m) \geq \sigma^2\|D_m\|^2 = O_p(\rho_{mn}^{-2}J_mK_m)$. Therefore, we have $D_1/(nR_n(m)) \lesssim \rho_{mn}^{-1}\Psi_{mn}$ for all $m \in \mathcal{M}_n$. Thus, the supremum of $D_1/(nR_n(m))$ is $o_p(1)$ by Assumption 4 (i).

Next consider D_2. Note that

$$
\begin{aligned}
D_2 &\leq n|(e'\mathcal{Z}_m/n)(\hat{G}_m'(F_m'u_m/n)\hat{G}_m - G_m'(F_m'u_m/n)G_m)(\mathcal{Z}_m'e/n)| \\
&\quad + n|(e'\mathcal{Z}_m/n)(G_m'(F_m'u_m/n)G_m)(\mathcal{Z}_m'e/n)| \\
&\leq n|(e'\mathcal{Z}_m/n)(\hat{G}_m - G_m)'(F_m'u_m/n)(\hat{G}_m - G_m)(\mathcal{Z}_m'e/n)| \\
&\quad + n|(e'\mathcal{Z}_m/n)(\hat{G}_m - G_m)'(F_m'u_m/n)G_m(\mathcal{Z}_m'e/n)| \\
&\quad + n|(e'\mathcal{Z}_m/n)G_m'(F_m'u_m/n)(\hat{G}_m - G_m)(\mathcal{Z}_m'e/n)| \\
&\quad + n|(e'\mathcal{Z}_m/n)(G_m'(F_m'u_m/n)G_m)(\mathcal{Z}_m'e/n)| \\
&\leq n\|e'\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}(F_m'u_m/n)Q_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'e/n\| \\
&\quad + n\|e'\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}(F_m'u_m/n)Q_{x,m}^{-1/2}\Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'e/n\| \\
&\quad + n\|e'\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|\Xi_m'Q_{x,m}^{-1/2}(F_m'u_m/n)Q_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'e/n\| \\
&\quad + n\|e'\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|\Xi_m'Q_{x,m}^{-1/2}(F_m'u_m/n)Q_{x,m}^{-1/2}\Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'e/n\| \\
&\equiv D_{21} + D_{22} + D_{23} + D_{24}.
\end{aligned}
$$

Therefore, by Lemma 1, we have $D_{21} = O_p(\rho_{mn}^{-4}\Psi_{mn}^2\sqrt{J_m^2K_m^4/n})$, $D_{22} = O_p(\rho_{mn}^{-3}\Psi_{mn}\sqrt{J_m^2K_m^4/n})$, $D_{23} = O_p(\rho_{mn}^{-3}\Psi_{mn}\sqrt{J_m^2K_m^4/n})$, and $D_{24} = O_p(\rho_{mn}^{-2}\sqrt{J_m^2K_m^4/n})$. Thus, by Assumption 4 (i), we obtain $D_2 = O_p(\rho_{mn}^{-2}\sqrt{J_m^2K_m^4/n})$. Note that $nR_n(m) \geq \sigma^2\|D_m\|^2 = O_p(\rho_{mn}^{-2}J_mK_m)$. Therefore, we have $D_2/(nR_n(m)) \lesssim \sqrt{K_m^2/n}$ for all $m \in \mathcal{M}_n$. By Assumption 3 (ii), the supremum of $D_2/(nR_n(m))$ is $o_p(1)$.

We now consider D_3. Note that

$$
\begin{aligned}
D_3 &\leq n|(e'\mathcal{Z}_m/n)(\hat{G}_m'(u_m'u_m/n)\hat{G}_m - G_m'(u_m'u_m/n)G_m)(\mathcal{Z}_m'e/n)| \\
&\quad + |e'\mathcal{Z}_mG_m'u_m'u_mG_m\mathcal{Z}_m'e - \mathrm{E}(e'\mathcal{Z}_mG_m'u_m'u_mG_m\mathcal{Z}_m'e|Z)|/n^2 \\
&\leq n|(e'\mathcal{Z}_m/n)(\hat{G}_m - G_m)'(u_m'u_m/n)(\hat{G}_m - G_m)(\mathcal{Z}_m'e/n)| \\
&\quad + n|(e'\mathcal{Z}_m/n)(\hat{G}_m - G_m)'(u_m'u_m/n)G_m(\mathcal{Z}_m'e/n)| \\
&\quad + n|(e'\mathcal{Z}_m/n)G_m'(u_m'u_m/n)(\hat{G}_m - G_m)(\mathcal{Z}_m'e/n)| \\
&\quad + |e'\mathcal{Z}_mG_m'u_m'u_mG_m\mathcal{Z}_m'e - \mathrm{E}(e'\mathcal{Z}_mG_m'u_m'u_mG_m\mathcal{Z}_m'e|Z)|/n^2 \\
&\leq n\|e'\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}(u_m'u_m/n)Q_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'e/n\| \\
&\quad + n\|e'\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}(u_m'u_m/n)Q_{x,m}^{-1/2}\Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'e/n\| \\
&\quad + n\|e'\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|\Xi_m'Q_{x,m}^{-1/2}(u_m'u_m/n)Q_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'e/n\| \\
&\quad + |e'\mathcal{Z}_mG_m'u_m'u_mG_m\mathcal{Z}_m'e - \mathrm{E}(e'\mathcal{Z}_mG_m'u_m'u_mG_m\mathcal{Z}_m'e|Z)|/n^2 \\
&\equiv D_{31} + D_{32} + D_{33} + D_{34}.
\end{aligned}
$$

Therefore, $D_{31} = O_p(\rho_{mn}^{-4}\Psi_{mn}^2J_mK_m^3/n)$, $D_{32} = O_p(\rho_{mn}^{-3}\Psi_{mn}J_mK_m^3/n)$, and $D_{33} = O_p(\rho_{mn}^{-3}\Psi_{mn}J_mK_m^3/n)$.

For D_{34}, note that

$$\begin{aligned}
&tr(\mathcal{Z}_m G_m' u_m' u_m G_m \mathcal{Z}_m' \mathcal{Z}_m G_m' u_m' u_m G_m \mathcal{Z}_m'/n^4) \\
&\quad = tr(\tilde{D}_m' u_m' u_m \tilde{D}_m \tilde{D}_m' u_m' u_m \tilde{D}_m) \\
&\quad \leq \lambda_{\max}(\tilde{D}_m' u_m' u_m \tilde{D}_m) tr(\tilde{D}_m' u_m' u_m \tilde{D}_m) \\
&\quad \leq \lambda_{\max}(D_m' D_m) tr(D_m' D_m) \\
&\quad \lesssim n R_n(m)
\end{aligned}$$

where the second inequality holds by the definitions of D_m and $\tilde{D}_m$ and the last inequality holds by the fact that $nR_n(m) \geq \sigma^2 tr(D_m' D_m)$ and Assumption 5 (i). Then, by Boole's inequality, Markov's inequality, Theorem 2 of Whittle (1960), and Assumption 6, we have

$$\begin{aligned}
&P\left(\sup_{m \in \mathcal{M}_n} |e' \mathcal{Z}_m G_m' u_m' u_m G_m \mathcal{Z}_m' e/n^2 - \mathrm{E}(e' \mathcal{Z}_m G_m' u_m' u_m G_m \mathcal{Z}_m' e/n^2 | Z)|/(nR_n(m)) > \delta | Z \right) \\
&\leq \sum_{m \in \mathcal{M}_n} P\left(|e' \mathcal{Z}_m G_m' u_m' u_m G_m \mathcal{Z}_m' e/^2 - \mathrm{E}(e' \mathcal{Z}_m G_m' u_m' u_m G_m \mathcal{Z}_m' e/n^2 | Z)|/(nR_n(m)) > \delta | Z \right) \\
&\leq \sum_{m \in \mathcal{M}_n} \frac{\mathrm{E}\left(\left| e' \mathcal{Z}_m G_m' u_m' u_m G_m \mathcal{Z}_m' e/n^2 - \mathrm{E}(e' \mathcal{Z}_m G_m' u_m' u_m G_m \mathcal{Z}_m' e/n^2 | Z) \right|^{2(N+1)} | Z \right)}{\delta^{2(N+1)} (nR_n(m))^{2(N+1)}} \\
&\leq \frac{C}{\delta^{2(N+1)}} \sum_{m \in \mathcal{M}_n} \frac{\left(tr(\mathcal{Z}_m G_m' u_m' u_m G_m \mathcal{Z}_m' \mathcal{Z}_m G_m' u_m' u_m G_m \mathcal{Z}_m'/n^4) \right)^{N+1}}{(nR_n(m))^{2(N+1)}} \\
&\leq \frac{C}{\delta^{2(N+1)}} \sum_{m \in \mathcal{M}_n} (nR_n(m))^{-(N+1)} \to 0.
\end{aligned}$$

Thus, by Assumption 4 (i), we obtain $D_3 = O_p(\rho_{mn}^{-3} \Psi_{mn} J_m K_m^3/n)$. Note that $nR_n(m) \geq \sigma^2 \|D_m\|^2 = O_p(\rho_{mn}^{-2} J_m K_m)$. Therefore, we have $D_3/(nR_n(m)) \lesssim \rho_{mn}^{-1} \Psi_{mn} K_m^2/n$ for all $m \in \mathcal{M}_n$. By Assumption 3 (ii) and 4 (i), the supremum of $D_3/(nR_n(m))$ is $o_p(1)$.

For (v), note that $|r_m'(\hat{D}_m' \hat{D}_m - D_m' D_m) r_m| = |tr((\hat{D}_m' \hat{D}_m - D_m' D_m)(r_m r_m'))| \leq |\lambda_{\max}(r_m r_m') \times tr(\hat{D}_m' \hat{D}_m - D_m' D_m)| \leq |r_m' r_m| |tr(\hat{D}_m' \hat{D}_m - D_m' D_m)|$. Observe that

$$\begin{aligned}
\|\hat{D}_m' \hat{D}_m - D_m' D_m\| &= \|\mathcal{Z}_m \hat{G}_m' \mathcal{X}_m' \mathcal{X}_m \hat{G}_m \mathcal{Z}_m' - \mathcal{Z}_m G_m' F_m' F_m G_m \mathcal{Z}_m'\|/n^2 \\
&\leq \|\mathcal{Z}_m (\hat{G}_m' F_m' F_m \hat{G} - G_m' F_m' F_m G_m) \mathcal{Z}_m'\|/n^2 + \|\mathcal{Z}_m \hat{G}_m' F_m' u_m \hat{G}_m \mathcal{Z}_m'\|/n^2 \\
&\quad + \|\mathcal{Z}_m \hat{G}_m' u_m' F_m \hat{G}_m \mathcal{Z}_m'\|/n^2 + \|\mathcal{Z}_m \hat{G}_m' u_m' u_m \hat{G}_m \mathcal{Z}_m\|/n^2 \\
&\equiv E_1 + E_2 + E_3 + E_4.
\end{aligned}$$

Consider E_1 first. Observe that

$$\begin{aligned}
E_1 &= \|\mathcal{Z}_m(\hat{G}_m'(F_m'F_m/n)\hat{G}_m - G_m'(F_m'F_m/n)G_m)\mathcal{Z}_m'|/n \\
&\leq \|\mathcal{Z}_m(\hat{G}_m - G_m)'H_m(\hat{G}_m - G_m)\mathcal{Z}_m'\|/n + \|\mathcal{Z}_m(\hat{G}_m - G_m)'H_mG_m\mathcal{Z}_m'\|/n \\
&\quad + \|\mathcal{Z}_mG_m'H_m(\hat{G}_m - G_m)\mathcal{Z}_m'\|/n \\
&\leq n\|\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}H_mQ_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| \\
&\quad + n\|\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}H_mQ_{x,m}^{-1/2}\Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| \\
&\quad + n\|\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|\Xi_m'Q_{x,m}^{-1/2}H_mQ_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| \\
&\equiv E_{11} + E_{12} + E_{13}.
\end{aligned}$$

Therefore, we have $E_{11} = O_p(\rho_{mn}^{-4}\Psi_{mn}^2J_mK_m)$, $E_{12} = O_p(\rho_{mn}^{-3}\Psi_{mn}J_mK_m)$, and $E_{13} = O_p(\rho_{mn}^{-3}\Psi_{mn}J_mK_m)$. Thus, we obtain $E_1 = O_p(\rho_{mn}^{-3}\Psi_{mn}J_mK_m)$. Recall that $nR_n(m) \geq \sigma^2\|D_m\|^2 = O_p(\rho_{mn}^{-2}J_mK_m)$. Therefore, we have $E_1/(nR_n(m)) \lesssim \rho_{mn}^{-1}\Psi_{mn}$ for all $m \in \mathcal{M}_n$. Thus, the supremum of $E_1/(nR_n(m))$ is $o_p(1)$ by Assumption 4 (i).

Next consider E_2. Note that

$$\begin{aligned}
E_2 &\leq \|\mathcal{Z}_m(\hat{G}_m'(F_m'u_m/n)\hat{G}_m - G_m'(F_m'u_m/n)G_m)\mathcal{Z}_m'\|/n + \|\mathcal{Z}_mG_m'(F_m'u_m/n)G_m\mathcal{Z}_m'\|/n \\
&\leq \|\mathcal{Z}_m(\hat{G}_m - G_m)'(F_m'u_m/n)(\hat{G}_m - G_m)\mathcal{Z}_m'\|/n + \|\mathcal{Z}_m(\hat{G}_m - G_m)'(F_m'u_m/n)G_m\mathcal{Z}_m'\|/n \\
&\quad + \|\mathcal{Z}_mG_m'(F_m'u_m/n)(\hat{G}_m - G_m)\mathcal{Z}_m'\|/n + \|\mathcal{Z}_mG_m'(F_m'u_m/n)G_m\mathcal{Z}_m'\|/n \\
&\leq n\|\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}(F_m'u_m/n)Q_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| \\
&\quad + n\|\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}(F_m'u_m/n)Q_{x,m}^{-1/2}\Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| \\
&\quad + n\|\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|\Xi_m'Q_{x,m}^{-1/2}(F_m'u_m/n)Q_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| \\
&\quad + n\|\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|\Xi_m'Q_{x,m}^{-1/2}(F_m'u_m/n)Q_{x,m}^{-1/2}\Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| \\
&\equiv E_{21} + E_{22} + E_{23} + E_{24}.
\end{aligned}$$

Therefore, it follows that $E_{21} = O_p(\rho_{mn}^{-4}\Psi_{mn}^2\sqrt{J_m^2K_m^4/n})$, $E_{22} = O_p(\rho_{mn}^{-3}\Psi_{mn}\sqrt{J_m^2K_m^4/n})$, $E_{23} = O_p(\rho_{mn}^{-3}\Psi_{mn}\sqrt{J_m^2K_m^4/n})$, and $E_{24} = O_p(\rho_{mn}^{-2}\sqrt{J_m^2K_m^4/n})$. Thus, by Assumption 4 (i), we obtain $E_2 = O_p(\rho_{mn}^{-2}\sqrt{J_m^2K_m^4/n})$. Note that $nR_n(m) \geq \sigma^2\|D_m\|^2 = O_p(\rho_{mn}^{-2}J_mK_m)$. Therefore, we have $E_2/(nR_n(m)) \lesssim \sqrt{K_m^2/n}$ for all $m \in \mathcal{M}_n$. By Assumption 3 (ii), the supremum of $E_2/(nR_n(m))$ is $o_p(1)$. By a similar argument, we have $E_3 = O_p(\rho_{mn}^{-2}\sqrt{J_m^2K_m^4/n})$ and the supremum of $E_3/(nR_n(m))$ is $o_p(1)$.

We now consider E_4. Note that

$$\begin{aligned}
E_4 &\leq \|\mathcal{Z}_m\hat{G}_m'u_m'u_m\hat{G}_m\mathcal{Z}_m - \mathcal{Z}_mG_m'u_m'u_mG_m\mathcal{Z}_m'\|/n^2 + \|\mathcal{Z}_mG_m'u_m'u_mG_m\mathcal{Z}_m'\|/n^2 \\
&\leq \|\mathcal{Z}_m(\hat{G}_m - G_m)'(u_m'u_m/n)(\hat{G}_m - G_m)\mathcal{Z}_m'\|/n + \|\mathcal{Z}_m(\hat{G}_m - G_m)'(u_m'u_m/n)G_m\mathcal{Z}_m'\|/n \\
&\quad + \|\mathcal{Z}_mG_m'(u_m'u_m/n)(\hat{G}_m - G_m)\mathcal{Z}_m'\|/n + \|\mathcal{Z}_mG_m'(u_m'u_m/n)G_m\mathcal{Z}_m'\|/n \\
&\leq n\|\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}(u_m'u_m/n)Q_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| \\
&\quad + n\|\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|(\hat{\Xi}_m - \Xi_m)'Q_{x,m}^{-1/2}(u_m'u_m/n)Q_{x,m}^{-1/2}\Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| \\
&\quad + n\|\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|\Xi_m'Q_{x,m}^{-1/2}(u_m'u_m/n)Q_{x,m}^{-1/2}(\hat{\Xi}_m - \Xi_m)\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| \\
&\quad + n\|\mathcal{Z}_mQ_{z,m}^{-1/2}/n\|\|\Xi_m'Q_{x,m}^{-1/2}(u_m'u_m/n)Q_{x,m}^{-1/2}\Xi_m\|\|Q_{z,m}^{-1/2}\mathcal{Z}_m'/n\| \\
&\equiv E_{41} + E_{42} + E_{43} + E_{44}.
\end{aligned}$$

Then we have $E_{41} = O_p(\rho_{mn}^{-4}\Psi_{mn}^2J_mK_m^3/n)$, $E_{42} = O_p(\rho_{mn}^{-3}\Psi_{mn}J_mK_m^3/n)$, $E_{43} = O_p(\rho_{mn}^{-3}\Psi_{mn}J_mK_m^3/n)$, and $E_{44} = O_p(\rho_{mn}^{-2}J_mK_m^3/n)$. Thus, by Assumption 4 (i), we obtain $E_4 = O_p(\rho_{mn}^{-2}J_mK_m^3/n)$. Note that $nR_n(m) \geq \sigma^2\|D_m\|^2 = O_p(\rho_{mn}^{-2}J_mK_m)$. Therefore, we have $E_4/(nR_n(m)) \lesssim K_m^2/n$ for all $m \in \mathcal{M}_n$. By Assumption 3 (ii), the supremum of $E_4/(nR_n(m))$ is $o_p(1)$. Therefore, we have $|r_m'(\hat{D}_m'\hat{D}_m - D_m'D_m)r_m|/(nR_n(m)) \leq |r_m'r_m||tr(\hat{D}_m'\hat{D}_m - D_m'D_m)/(nR_n(m))| \lesssim \rho_{mn}^{-1}\Psi_{mn} + K_m^2/n$ for all $m \in \mathcal{M}_n$. Thus, the result follows from Assumption 3 (ii) and Assumption 4 (i).

For (vi), by the Cauchy-Schwarz inequality and the trace inequality, we have

$$\begin{aligned}
|e'(\hat{D}_m'\hat{D}_m - D_m'D_m)r_m| &\leq (e'(\hat{D}_m'\hat{D}_m - D_m'D_m)(\hat{D}_m'\hat{D}_m - D_m'D_m)'e)^{1/2}(r_m'r_m)^{1/2} \\
&= tr((ee')(\hat{D}_m'\hat{D}_m - D_m'D_m)(\hat{D}_m'\hat{D}_m - D_m'D_m)')^{1/2}(r_m'r_m)^{1/2} \\
&\leq \|ee'\|^{1/2}tr(\hat{D}_m'\hat{D}_m - D_m'D_m)(\hat{D}_m'\hat{D}_m - D_m'D_m)')^{1/2}(r_m'r_m)^{1/2} \\
&= \|ee'\|^{1/2}\|\hat{D}_m'\hat{D}_m - D_m'D_m\|(r_m'r_m)^{1/2}.
\end{aligned}$$

Note that $nR_n(m) \geq \sigma^2\|D_m\|^2 = O_p(\rho_{mn}^{-2}J_mK_m)$, $\mathrm{E}\left(e_i^2|z_i\right) = \sigma^2 \leq \infty$, $\mathrm{E}\left(r_{mi}^2|z_i\right) = \phi_m^2 \leq \infty$, and $\|\hat{D}_m'\hat{D}_m - D_m'D_m\| = O_p(\rho_{mn}^{-3}\Psi_{mn}J_mK_m + \rho_{mn}^{-2}J_mK_m^3/n)$ shown in (v). Therefore, we have $|e'(\hat{D}_m - D_m)r_m|/(nR_n(m)) \lesssim \rho_{mn}^{-1}\Psi_{mn} + K_m^2/n$ for all $m \in \mathcal{M}_n$. Thus, the result follows from Assumption 3 (ii) and Assumption 4 (i). This completes the proof. ■

C Figures

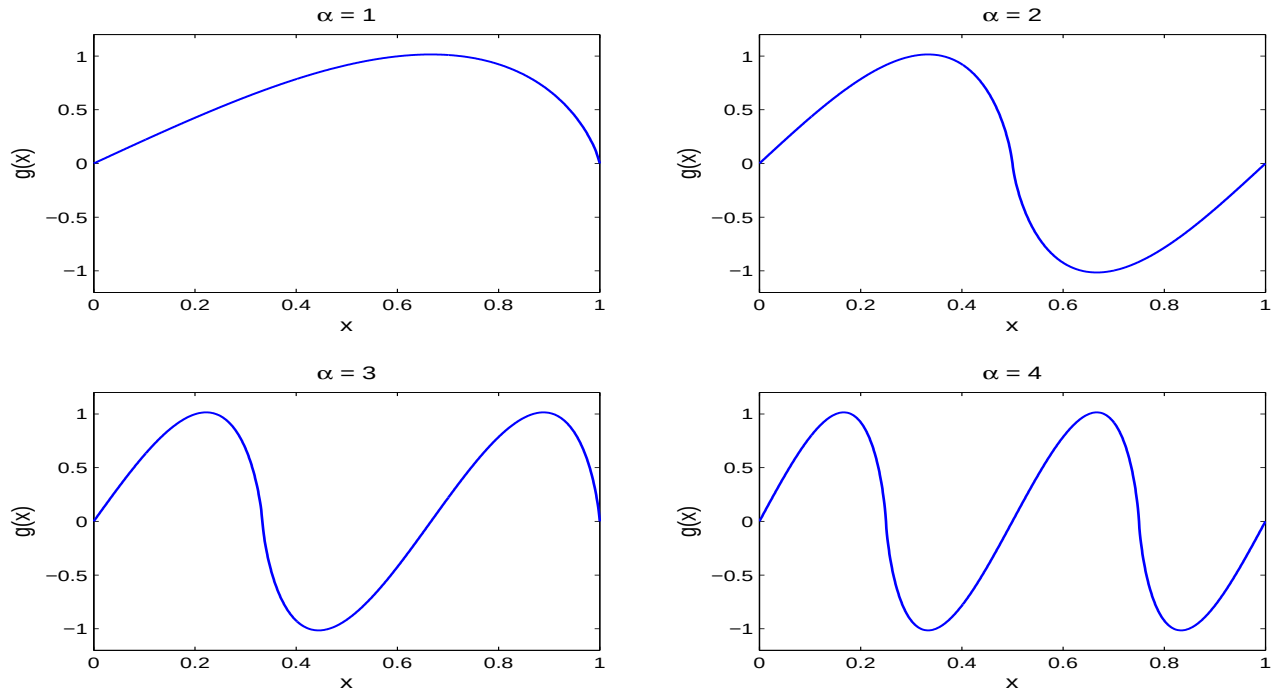

Figure 1: Graph of $g(x)$.

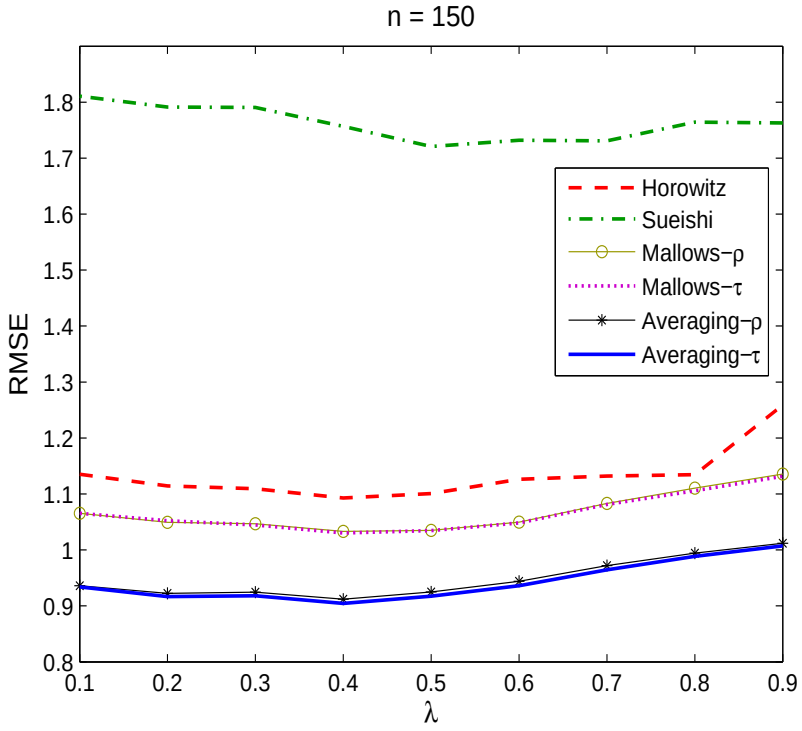

Figure 2: Normalized RMSE for Legendre polynomials.

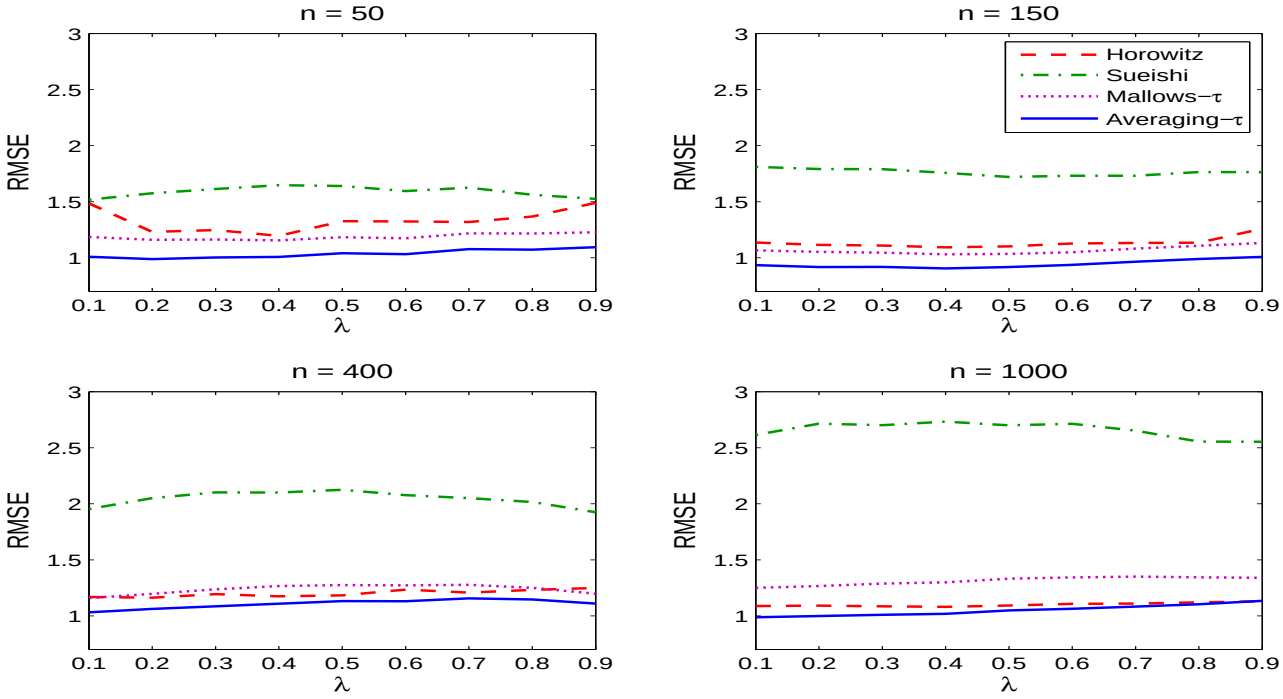

Figure 3: Normalized RMSE for Legendre polynomials.

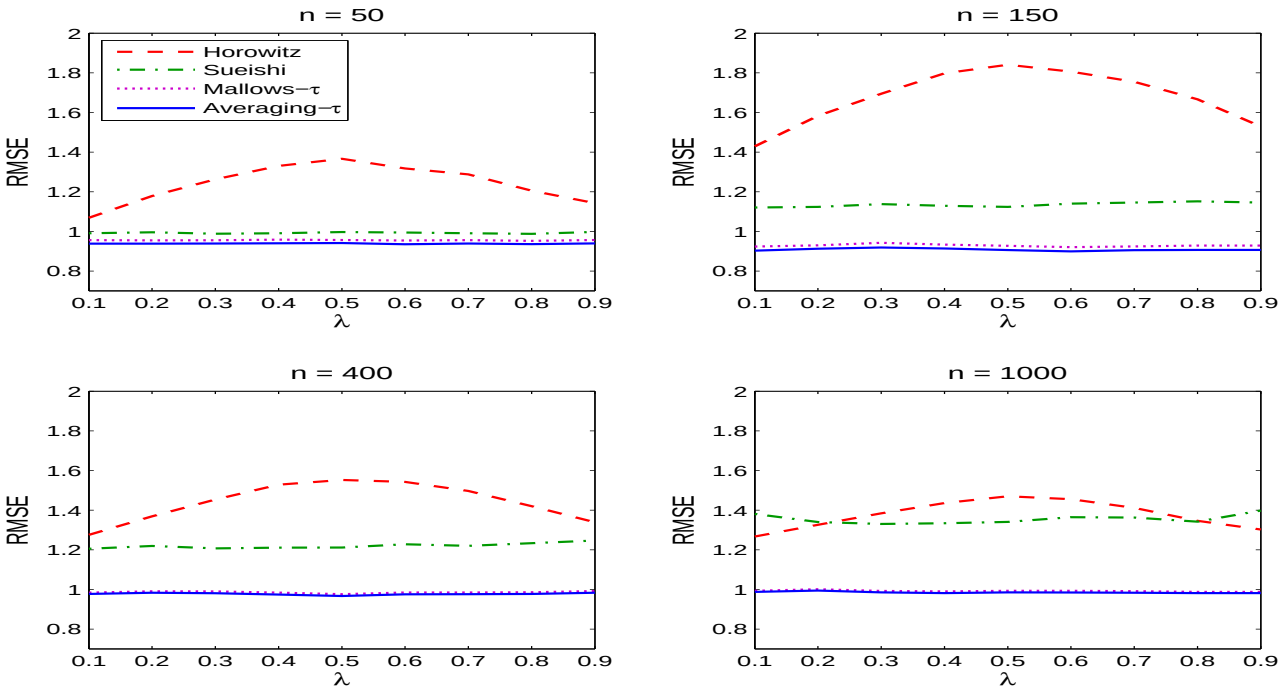

Figure 4: Normalized RMSE for B-splines.

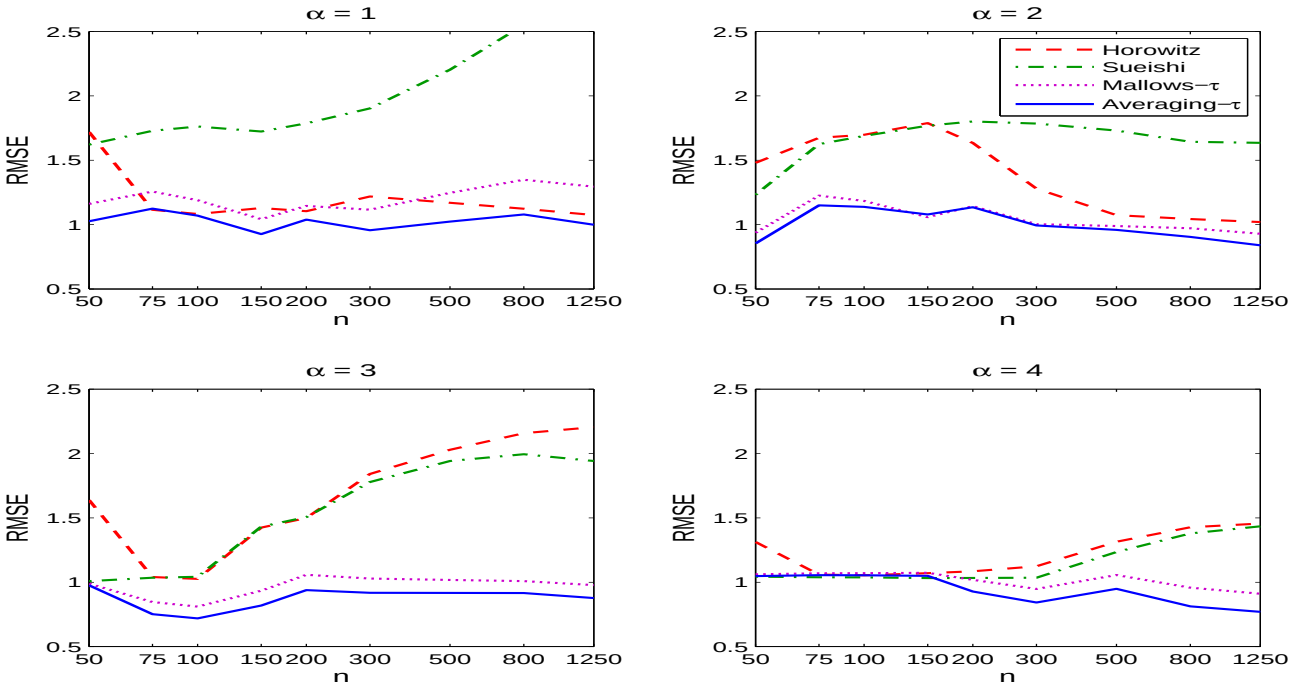

Figure 5: Normalized RMSE for Legendre polynomials.

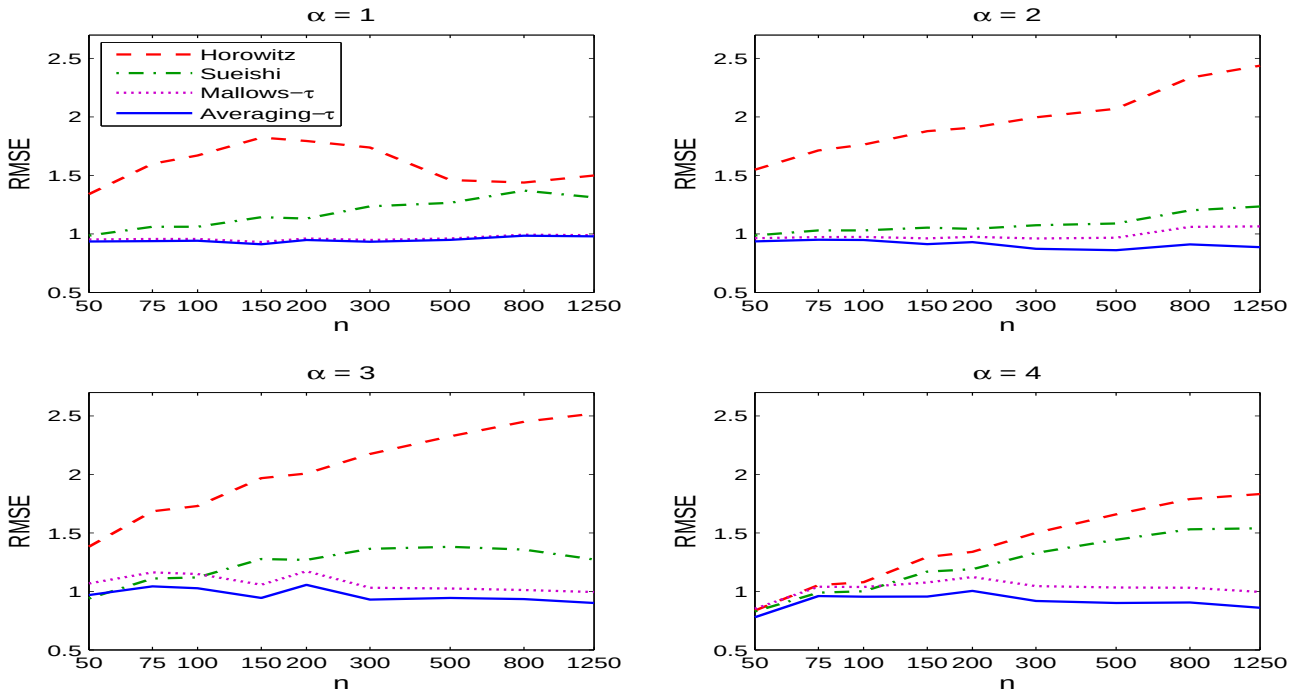

Figure 6: Normalized RMSE for B-splines.

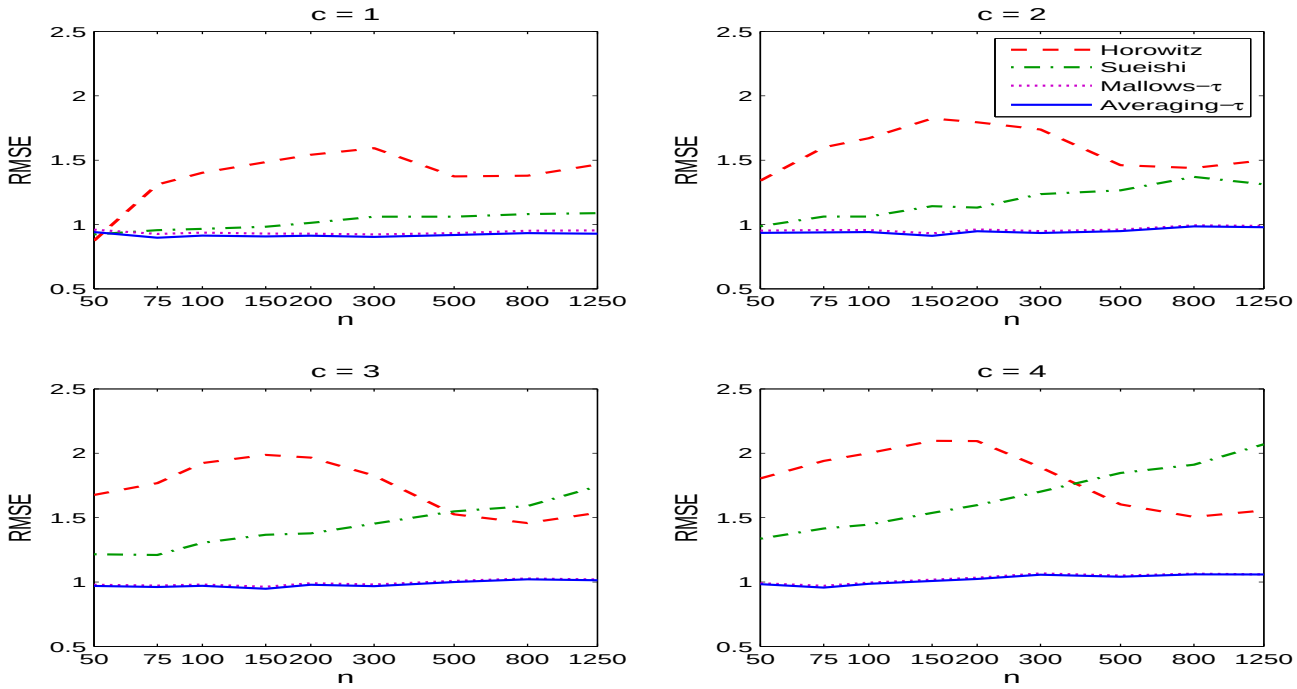

Figure 7: Normalized RMSE for B-splines with $\alpha = 1$.

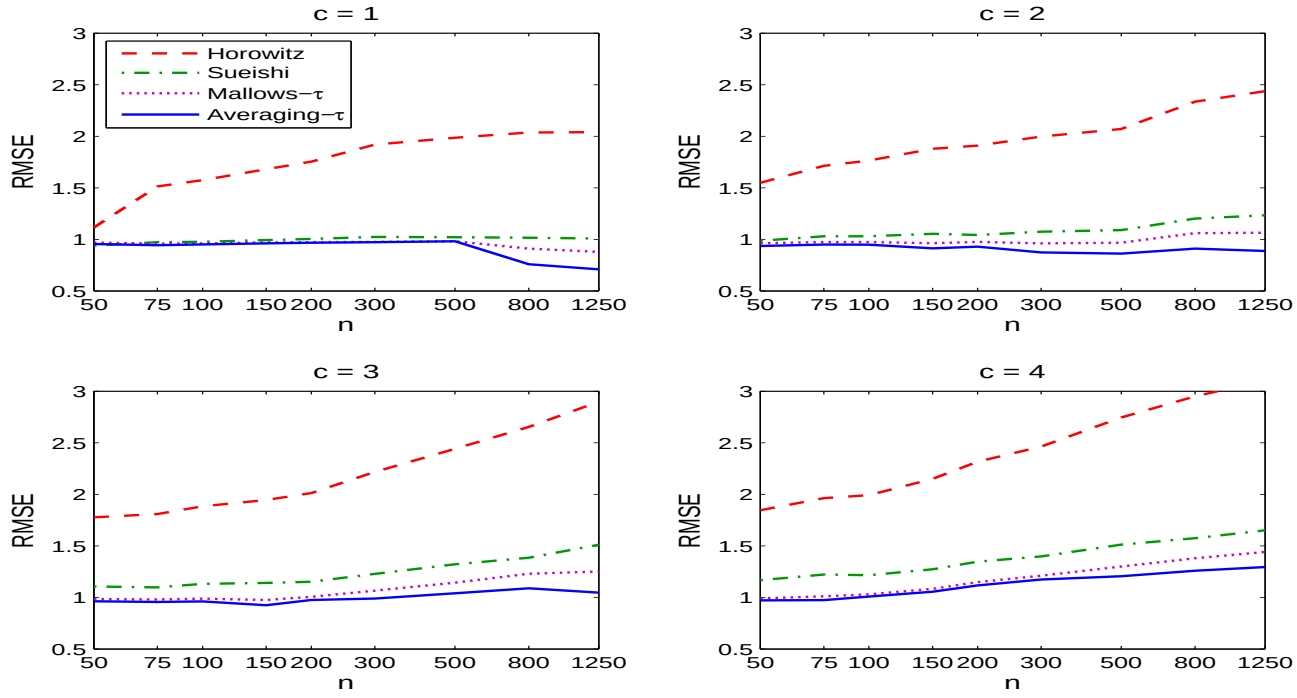

Figure 8: Normalized RMSE for B-splines with $\alpha = 2$.

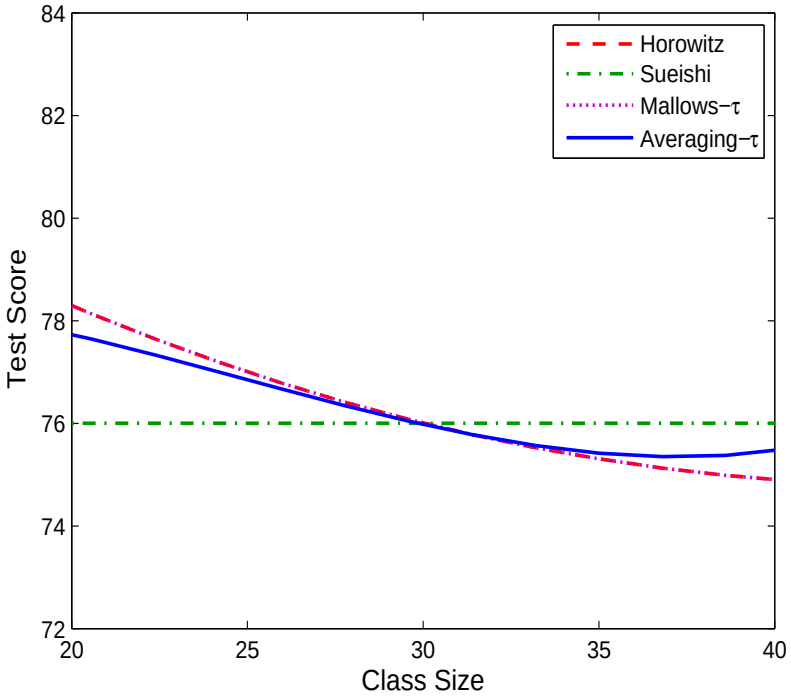

Figure 9: Estimate of test score as a function of class size by Legendre polynomials.

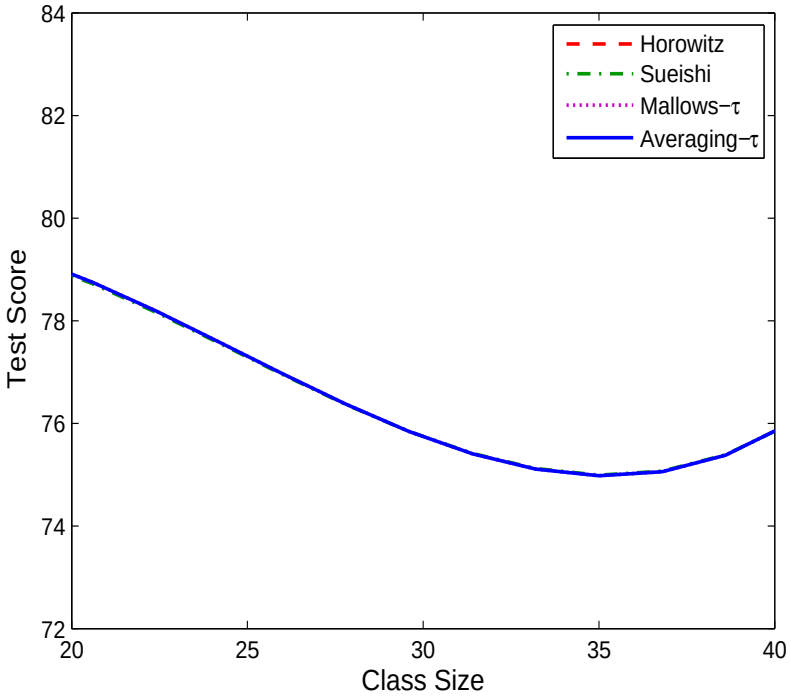

Figure 10: Estimate of test score as a function of class size by B-splines.

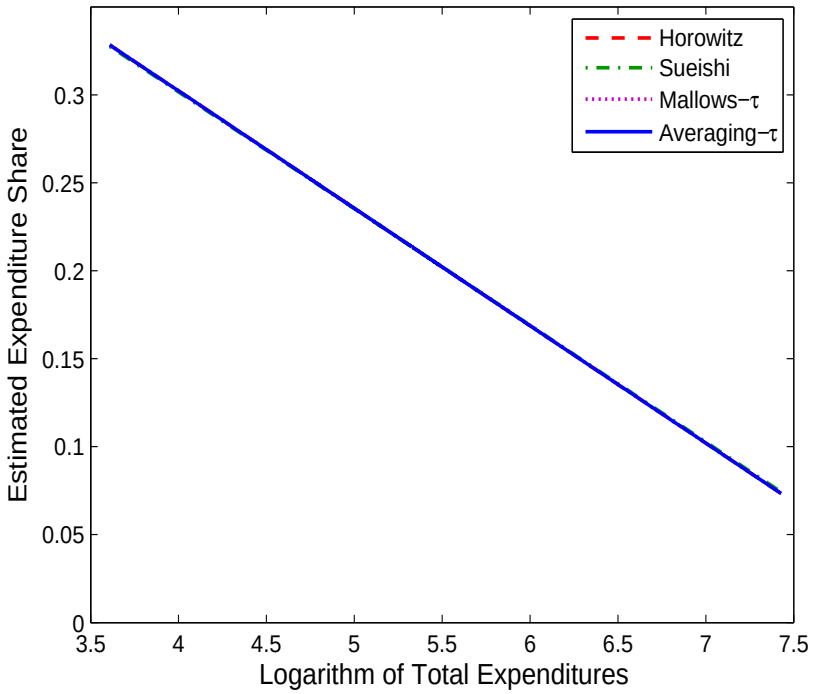

Figure 11: Estimate of Engel curve by Legendre polynomials.

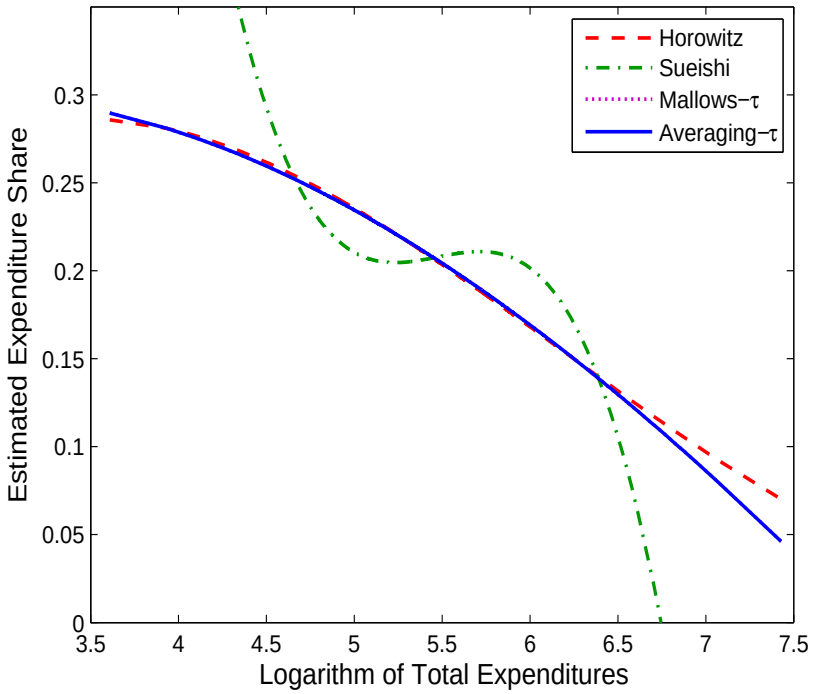

Figure 12: Estimate of Engel curve by B-splines.

References

Ai, C. and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.

Andrews, D. W. K. (1991): "Asymptotic Optimality of Generalized C_L, Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359–377.

——— (1999): "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," *Econometrica*, 67, 543–563.

——— (2011): "Examples of L^2-Complete and Boundedly-Complete Distributions," *Cowles Foundation for Research in Economics*.

Andrews, D. W. K. and B. Lu (2001): "Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models," *Journal of Econometrics*, 101, 123–164.

Angrist, J. D. and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *The Quarterly Journal of Economics*, 114, 533–575.

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, 2369–2429.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75, 1613–1669.

Breunig, C. and J. Johannes (2015): "Adaptive Estimation of Functionals in Nonparametric Instrumental Regression," *Econometric Theory*, FirstView, 1–43.

Carrasco, M. (2012): "A Regularization Approach to the Many Instruments Problem," *Journal of Econometrics*, 170, 383–398.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 5633–5751.

Centorrino, S. (2015): "Data driven selection of the regularization parameter in additive nonparametric instrumental regressions," Working Paper.

Centorrino, S., F. Feve, and J.-P. Florens (2015): "Additive Nonparametric Instrumental Regressions: A Guide to Implementation," Forthcoming. *Journal of Econometric Methods*.

Chen, X., V. Chernozhukov, S. Lee, and W. K. Newey (2014): "Local Identification of Nonparametric and Semiparametric Models," *Econometrica*, 82, 785–809.

Chen, X. and T. Christensen (2013): "Optimal Uniform Convergence Rates for Sieve Nonparametric Instrumental Variables Regression," Cemmap Working Paper CWP 56/13.

——— (2015): "Optimal Sup-norm Rates, Adaptivity and Inference in Nonparametric Instrumental Variables Estimation," Cemmap Working Paper CWP 32/15.

Chen, X. and D. Pouzo (2012): "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals," *Econometrica*, 80, 277–321.

Chen, X. and M. Reiss (2011): "On Rate Optimality for Ill-Posed Inverse Problems in Econometrics," *Econometric Theory*, 27, 497–521.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139, 4–14.

Claeskens, G. and N. L. Hjort (2008): *Model Selection and Model Averaging*, Cambridge: Cambridge University Press.

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79, 1541–1565.

D'Haultfoeuille, X. (2011): "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, 27, 460–471.

Donald, S. G., G. W. Imbens, and W. K. Newey (2009): "Choosing Instrumental Variables in Conditional Moment Restriction Models," *Journal of Econometrics*, 152, 28–36.

Donald, S. G. and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69, 1161–1191.

Gagliardini, P. and O. Scaillet (2012a): "Nonparametric Instrumental Variable Estimation of Structural Quantile Effects," *Econometrica*, 80, 1533–1562.

——— (2012b): "Tikhonov Regularization for Nonparametric Instrumental Variable Estimators," *Journal of Econometrics*, 167, 61–75.

Hall, P. and J. L. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *The Annals of Statistics*, 33, 2904–2929.

Hansen, B. E. (2007): "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189.

——— (2015): "The Integrated Mean Squared Error of Series Regression and a Rosenthal Hilbert-Space Inequality," *Econometric Theory*, 31, 337–361.

Hansen, B. E. and J. Racine (2012): "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46.

Hanushek, E. A. (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24, 1141–1177.

Hjort, N. L. and G. Claeskens (2003): "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899.

Hong, H., B. Preston, and M. Shum (2003): "Generalized Empirical Likelihood-Based Model Selection Criteria for Moment Condition Models," *Econometric Theory*, 19, 923–943.

Horowitz, J. L. (2011): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79, 347–394.

——— (2012): "Specification Testing in Nonparametric Instrumental Variable Estimation," *Journal of Econometrics*, 167, 383–396.

——— (2014): "Adaptive Nonparametric Instrumental Variables Estimation: Empirical Choice of the Regularization Parameter," *Journal of Econometrics*, 180, 158–173.

Horowitz, J. L. and S. Lee (2007): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," *Econometrica*, 75, 1191–1208.

Ing, C.-K. and C.-Z. Wei (2003): "On Same-Realization Prediction in an Infinite-Order Autoregressive Process," *Journal of Multivariate Analysis*, 85, 130–155.

——— (2005): "Order Selection for Same-Realization Predictions in Autoregressive Processes," *The Annals of Statistics*, 33, 2423–2474.

Kress, R. (1999): *Linear Integral Equations*, Springer-Verlag.

Kuersteiner, G. and R. Okui (2010): "Constructing Optimal Instruments by First-Stage Prediction Averaging," *Econometrica*, 78, 697–718.

Leeb, H. and B. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

Li, K.-C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975.

Liu, Q. and R. Okui (2013): "Heteroscedasticity-Robust C_p Model Averaging," *The Econometrics Journal*, 16, 463–472.

Newey, W. K. (2013): "Nonparametric Instrumental Variables Estimation," *The American Economic Review*, 103, 550–556.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Okui, R. (2011): "Instrumental Variable Estimation in the Presence of Many Moment Conditions," *Journal of Econometrics*, 165, 70–86.

Shao, J. (1997): "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, 7, 221–242.

Shibata, R. (1980): "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process," *The Annals of Statistics*, 8, 147–164.

——— (1981): "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45–54.

Sueishi, N. (2012): "Model Selection Criterion for Instrumental Variable Models," Working Paper, Kobe University.

Wan, A., X. Zhang, and G. Zou (2010): "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283.

Whittle, P. (1960): "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability and Its Applications*, 5, 302–305.

Zhang, X., A. T. Wan, and G. Zou (2013): "Model Averaging by Jackknife Criterion in Models with Dependent Data," *Journal of Econometrics*, 174, 82–94.