Nonnested model selection based on empirical likelihood

Jiancheng Jiang

Department of Mathematics and Statistics & School of Data Science,
University of North Carolina at Charlotte, NC 28223, USA.

Xuejun Jiang *

Department of Statistics and Data Science,
Southern University of Science and Technology, Shenzhen, 518055, China.

Haofeng Wang

Department of Mathematics,
Harbin Institute of Technology, Harbin, 150001, China.

January 21, 2022

Abstract

We propose an empirical likelihood ratio test for nonparametric model selection, where the competing models may be nested, nonnested, overlapping, misspecified, or correctly specified. It compares the squared prediction errors of models based on the cross-validation and allows for heteroscedasticity of the errors of models. We develop its asymptotic distributions for comparing additive models and varying-coefficient models and extend it to test significance of variables in additive models with massive data. The method is applicable to other model comparison problems. To facilitate implementation of the test, we provide a fast calculation procedure. Simulations show that the proposed tests work well and have favorable finite sample performance over some existing approaches. The methodology is validated on an empirical application.

Keywords: Empirical likelihood ratio; Distributed computation; Model selection; Nonparametric smoothing; Prediction error.

---

*Correspondence: Xuejun Jiang, Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, 518055, China. Email: *jiangxj@sustech.edu.cn*

1 Introduction

In application, one often needs to decide which model works better for a given dataset among a set of misspecified models since no model is right. This motivates us to introduce a novel empirical likelihood ratio (ELR) test to model selection. The proposed method is applicable to model selection between any two supervising learning models, which may be nested, nonnested, overlapping, misspecified, or correctly specified.

Most existing model selection methods use likelihood or information criteria, such as AIC, BIC, LASSO or SCAD, etc. They are widely used in statistical theory and have made great success in practice, but cannot be directly applied to nonnested model selection. Consider, for example, selecting important genes in the non-Hodgkin's lymphoma data in Dave et al. (2004) using the famous Cox's model and the additive hazard model, based on the LASSO. Each model may lead to a different group of important genes, but there is no general tool to judge which model is better. Since the two models are nonnested, the likelihood comparison does not make sense, and hence the AIC and BIC criteria cannot be used. Comparison of nonnested models also arises in time series modeling, for instance, assessment of an ARCH(7) model versus a GARCH(1,2) model. In other situations, even if models are nested, one may have difficulty in making a decision on selecting a better model. For example, suppose there are two candidate models with AIC values equal to 100 and 102. Then the model with an AIC value of 100 is preferred according to this criterion. However, one cannot conclude that it is definitely better, because one AIC value is smaller than the other possibly due to randomness of the sample. In other words, one does not have a clear cutoff for the difference of AIC values to judge which model is significantly better. Therefore, there is a genuine need to develop a formal test that furnishes a critical value for nonnested model selection.

There exist many works on hypothesis testing for nonnested model selection. Cox (1961, 1962) pioneered a likelihood ratio (LR) test for two separate families of hypotheses and heuristically argued its asymptotic normality, which was rigorously proven by White (1982a) under regularity conditions. In a seminal article, Vuong (1989) used the Kullback-Leibler information criterion (KLIC) to measure the closeness of a model to the true data generating process (DGP). He introduced an LR test for competing models, which may be nested or non-nested and correctly specified or misspecified. However, it works only for parametric models with known distributions. Rivers and Voung (2002) extended this approach by replacing the likelihood with general lack of fit criteria, which allows for more estimation approaches, but it still works only for parametric dynamic models. Chen, Hong and Shum (2007) advanced a nonparametric LR test for comparing a parametric likelihood model with a parametric moment condition model, based on the KLIC criterion, which can be regarded as extensions to the LR test of Vuong (1989). McElroy (2016) proposed a Whittle LR test for nonnested model selection. This approach also employs the KLIC criterion and is designed for comparing two parametric time series models with spectral densities.

However, the above KLIC-criterion based tests have different limiting distributions, depending on whether the two models are overlapping or not, and whether one of the model is correctly specified or not. They require one to pretest which distribution to use before applying them. As a consequence, they are basically two-step test procedures, which may induce the nonuniformity phenomenon of tests (Leeb and Pötscher, 2005) and result in size distortions (Shi, 2015; Schennach and Wilhelm, 2017). In order to deal with this problem, Shi (2015) proposed a one-step nondegenerated test for nonnested models, which is a modification of the Vuong test, and Schennach and Wilhelm (2017) suggested a reweighted LR test for nonnested model selection. Both of the tests achieve uniform size control, but they are tailored for parametric models with densities.

Some authors introduced nonparametric extensions to the LR test. Fan et al. (2001) proposed generalized likelihood ratio (GLR) tests and showed that the Wilks type of results hold for a variety of useful models, including univariate non-parametric models, varying-coefficient models, and their extensions. Fan and Jiang (2005) developed the GLR test for additive models based on the local polynomial fitting and the backfitting algorithm. Fan et al. (2001), Fan and Huang (2005), and Fan and Jiang (2005, 2007) showed the generality of the Wilks phenomenon and enriched the applicability of the GLR tests. However, the GLR tests work only for nested models, require the working models contains the DGP, and generally assume homogeneity of variance. Moreover, the asymptotic distributionbs of GLR tests explicitly depend on the bandwidth. It remains unknown if the GLR test can be modified for nonnested model selection. Liao and Shi (2020) proposed a nondegenerate Vuong test to comparison of nonnested nonparametric models, which employs sieve approximations for M-estimation of the models, but the test requires correcting two bias terms and estimating the complicate variance, which explicitly depends on the tuning parameter in the sieve approximation. In addition, it cannot deal with heteroscedastic errors, because their assumption 4.1(a) and the natrure of their M-estimate in eq.(3.3) assume the error has a constant variance (see also Example 2).

Last but not the least, it is worthy of mentioning that there are various metrics for model comparison within the Bayesian framework. Two popular approaches among them are Bayes factors (Lewis and Raftery, 1997) and the deviance information criterion (Spiegelhalter et al., 2002). However, these methods are designed only for comparison of parametric models.

In this paper, we propose a general nonparametric test approach to model selection. It is known that the prediction error criterion allows one to compare any two supervised statistical learning methods (parametric or nonparametric). In practice, a statistical learning procedure with a smaller average (absolute or squared) prediction error (APE) is usually

preferred. However, if the APEs are close among competing models, one does not know if the APEs are significantly different. Furthermore, a model with a smaller APE may be caused by randomness of the sample, but not because of a better model. These problems have hovered around statisticians over decades. In an effort to solve them and to perform an accurate model selection, we will resort to the idea of Owen (1988, 1989) and propose some new ELR tests to compare prediction errors of competing models, based on the cross-validation method. The proposed tests possess the following appealing characteristics:

  (i) They are nonparametric tests without requiring a specific parametric structure or likelihood.

 (ii) The ELR tests allow for heteroscedasticity of the errors, and their asymptotic distributions and power do not depend on the smoothing parameters.

(iii) It allows one to fast implement the tests.

(iv) The tests have power to detect all $\sqrt{n}$ local alternatives.

 (v) The idea is applicable to comparison between any two supervising statistical learning models, nested, non-nested, overlapping, correctly specified, or misspecified.

Because of the above features, the proposed ELR test is robust against heteroskedasticity, a striking contrast to the GLR tests, and it can be applied to post model inference (Tibshirani et al. 2016), for example, comparison between two post LASSO models, nested or nonnested. The ELR test targets at comparing forecast equivalence of nonnested models, so can it be used to measure importance of explanatory variables for forecast in big data settings where mere significance tests do not make much sense (see Section 3).

The empirical likelihood has been demonstrated as a powerful nonparametric tool for interval estimates (Owen, 2001). The method has many advantages over the normal approximation-based method and the bootstrap method for constructing confidence intervals, such as the transformation respecting, the range of parameter preserving, the Bartlett correctable property, no requirement for estimating scale and skewness, and no predetermined shape requirement (Hall and La Scala, 1990). There exists a vast literature devoted

5

to the empirical likelihood for parametric models, but relatively less work for nonparametric models. For interval estimation and hypothesis testing based on the empirical likelihood, they include but not limited to Hall (1990), Fan and Zhang (2004), Chen and Keilegom (2009), etc. However, all these works formulate the ELR with some moment constraints from the estimation equations for the parameter of interest, and no one acts for nonnested models. In the construction of the proposed ELR tests for nonnested models, we do not use moment constraints from the estimation equations. Existing techniques for the ELR tests work only for correctly specified models and cannot be used to derive the asymptotic distributions of the proposed ELR tests. These endow our work with challenges and intelligence. Since the ELR tests employ the leave-one-out cross-validation (LOOCV) to calculate the prediction errors, it is computationally expensive if one directly fits the models to the data with each observation held out. Due to the nature of global polynomial spline smoother used for competing models, we are able to introduce a fast computation procedure for implementation of the ELR tests. This procedure requires us to fit the models to the data only once. Furthermore, it is extended to test signficance of variables in additive models with massive or distributed data, and a distributed ELR test is developed and posesses the same performance as the ideal ELR test in the sense that there is no limited memory constraint and full data can be run on one machine.

This article is organized as follows. In Section 2 we describe the methodology. The asymptotic distributions of our ELR statistic are established whether or not the models are nested or misspecified, from which a decision rule of model selection is proposed. The fast implementation of the test is also considered. In Section 3, we develop the distributed ELR test for massive data. In Section 4 we investigate finite sample performance of ELR tests via simulation, and in Section 5 we provide an example of ELR test on a real dataset. Conditions and technical proofs are delegated to the Appendix.

2 Methods

Our main objective is to develop the ELR theory for model selection. To expose our idea, we consider model comparison between the additive model and the varying coefficient model. For other model comparison problems, our procedure can still applied but needs to be studied on a case-by-case basis.

2.1 Model comparison based on prediction errors

Nonlinearity relationship exists widely in statistical theory and practice. Suppose we have a random sample $\{y_i, \mathbf{X}_i, z_i\}_{i=1}^n$, where $\mathbf{X}_i = (x_{i,1}, \ldots, x_{i,p})^\top$, and we have found that there exists some in-sample significant evidence of "nonlinearity" between y_i and $\mathbf{X}_i$. We are interested in further investigating whether the documented "nonlinearity" is the true nonlinearity between y_i and $\mathbf{X}_i$, or is due to the functional coefficients in a linear regression model. To deal with this problem, we conduct model selection between the functional coefficient model (Hastie and Tibshirani, 1993; Fan and Zhang, 1999; Cai, Fan and Yao, 2000)

$$y_i = \beta_0(z_i) + \sum_{j=1}^p x_{i,j} \beta_j(z_i) + u_i, \tag{1}$$

and the nonparametric additive model (Hastie and Tibshirani, 1990)

$$y_i = \alpha + \sum_{j=1}^p m_j(x_{i,j}) + v_i, \tag{2}$$

in the framework that both models may be wrongly specified, where z_i may be a component of X_i or not, and for identifiability it is assumed that $E\{m_j(x_{i,j})\} = 0$. Obviously, models (1) and (2) are nonnested in general, but nested when $p = 1$ and $z_i = x_{i,1}$. They also overlap at the region where y_i and $\mathbf{X}_i$ are linearly related.

To get the prediction errors, we first need to estimate the unknown functions of models (1) and (2). Various estimation methods can be applied, such as the kernel smoother (Opsomer and Ruppert, 1997, 1998; Mammen and Park, 2006), the spline method (Stone,

1986; Zhou, Shen and Wolfe, 1998; Huang and Shen, 2004; Li and Liang, 2008), and even the boosting learning algorithms (Freund and Schapire, 1997; Friedman, 2001). In fact, one regards an estimation approach for a given supervising statistical model as a learning algorithm, and our ELR test can compare any two learning algorithms that provide predictions. Then we need a good measure for assessing the performance of models (1) and (2).

A natural one is the prediction error from the widely used K-fold cross validation (CV) (Hastie and Tibshirani, 1990), even though other measures may be used. This method randomly partitions the data into K roughly equal-sized parts. For the kth part, one uses the other $K - 1$ parts of the data for training and calculates the prediction error of each fitted model when predicting the kth part of the data. As in Hastie et al. (2009), we let $\theta : \{1, \ldots, n\} \to \{1, \ldots, K\}$ be an indexing function that indicates the partition to which observation i is allocated by the randomization, and let $\hat{\alpha}^{[-k]}$, $\hat{\beta}_j^{[-k]}(\cdot)$ and $\hat{m}_j^{[-k]}(\cdot)$ be fitted functions, computed with the kth part of the data removed. In particular, when $K = n$, $\theta(i) = i$, which corresponds to the leave-one-out CV (LOOCV). Denote by $\hat{\varepsilon}_{1,i} = y_i - \hat{\beta}_0^{[-\theta(i)]}(z_i) - \sum_{j=1}^p x_{i,j} \hat{\beta}_j^{[-\theta(i)]}(z_i)$ and $\hat{\varepsilon}_{2,i} = y_i - \hat{\alpha}^{[-\theta(i)]} - \sum_{j=1}^p \hat{m}_j^{[-\theta(i)]}(x_{i,j})$ the prediction errors for model (1) and (2), respectively. Then the average (squared) prediction errors (APE) are

$$APE_1 = n^{-1} \sum_{i=1}^n \hat{\varepsilon}_{1,i}^2 \ \text{ and } \ APE_2 = n^{-1} \sum_{i=1}^n \hat{\varepsilon}_{2,i}^2$$

for model (1) and (2), respectively. Let $\hat{\xi}_i = \hat{\varepsilon}_{1,i}^2 - \hat{\varepsilon}_{2,i}^2$. Then the difference

$$APE_1 - APE_2 = n^{-1} \sum_{i=1}^n \hat{\xi}_i$$

is an appropriate estimate of the difference of mean squared prediction errors

$$\mu_\xi = E(\hat{\xi}_i) = E(\hat{\varepsilon}_{1,i}^2) - E(\hat{\varepsilon}_{2,i}^2)$$

between the two models, and it can be used to compare the performance of the two models

in terms of prediction. When it is significantly different from zero, it signals that the two models are not competing. Otherwise, it is an indication of forecast equivalence.

2.2 The ELR test

Most existing model selection methods employ the likelihood or information criteria to measure the distance of a working model to the DGP. However, for nonparametric models, the generalized likelihood ratio method works only for nested models and has some disadvantages, and the information approach KLIC is not applicable, as discussed in Section 1. We here introduce an EL approach to evaluate probability of forecast equivalence of two competing models.

As a nonparametric method, the EL (Owen, 2001) has become a standard approach to construct interval estimates. To use the EL, one must specify estimating equations for the parameters of interest, but it is not necessary to estimate the variances of the estimators of parameters. The latter property endows the EL with ability of handling heteroscedastic and asymmetric errors. Note that the nonparametric likelihood for forecast equivalence of the competing models is characterized by

$$\sup\{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_i p_i = 1, \sum_i p_i \hat{\xi}_i = 0\}.$$

Following the idea of (Owen, 1988, 1990; Qin and Lawless, 1994), the above likelihood can be compared with nonparametric likelihood of a saturated model without any constraints, in which all p_i are equal to $1/n$. Hence, we define the logarithm of the ELR

$$R_{n,1} = -2 \log \sup\{\prod_{i=1}^{n} (n p_i) : p \in \mathcal{G}\}, \tag{3}$$

where $\mathcal{G} = \{p : p_i \geqslant 0, \sum_i p_i = 1, \sum_i p_i \hat{\xi}_i = 0\}$. Note that $\min \hat{\xi}_i \leq \sum_i p_i \hat{\xi}_i \leq \max \hat{\xi}_i$. If $0 \notin [\min \hat{\xi}_i, \max \hat{\xi}_i]$, then $\mathcal{G}$ is empty and we set $R_{n,1} = +\infty$.

In the above construction, we do not set any moment constraints from the estimation equations for both models (1) and (2). This is remarkably different from the classical ELR

9

statistics where some moment constraints from the estimation equations are placed. Using the Lagrange multiplier technique, when $\min \hat{\xi}_i \leq 0 \leq \max \hat{\xi}_i$, we obtain that $p_i = n^{-1}\frac{1}{1+\lambda\hat{\xi}_i}$, where λ satisfies that

$$f(\lambda) \equiv \sum_{i=1}^{n} \hat{\xi}_i/(1 + \lambda\hat{\xi}_i) = 0. \tag{4}$$

Let $\hat{\lambda}$ be the solution of equation (4). Then the logarithm of the ELR becomes

$$R_{n,1} = 2\sum_{i=1}^{n} \log(1 + \hat{\lambda}\hat{\xi}_i). \tag{5}$$

If $\mu_\xi = 0$, then the two models have equivalent performance in the sense that they have same prediction error on average, that is, forecast equivalence (McElroy, 2016). The empirical likelihood ratio $R_{n,1}$ can be used to assess which model is better in terms of APE. In fact, if the two models perform equivalently, $R_{n,1}$ will be like a chi-squared random variable; if one model is significantly better than the other, then $R_{n,1}$ will go to infinity (see Theorem 1 below).

Given the competing models (1) and (2), one usually selects the model with a smaller APE, but it is still unknown if the selected model is significantly better. In other words, we need to develop a test to distinguish if $APE_1 - APE_2$ is significantly different from zero. Therefore, we consider the following hypothesis testing problem:

$$H_0^{(1)} : \mu_\xi = 0 \text{ against } H_a^{(1)} : \mu_\xi \neq 0. \tag{6}$$

The null H_0 means that models (1) and (2) perform equivalently according to prediction, and the alternative represents one model is sufficiently better than the other.

2.3 Asymptotic distributions and the decision rule

The K-fold CV is easy to implement, but for a given sample one has to select a value of K. For ease of notations and for convenience of technical arguments, we only consider the LOOCV. Even though our results hold for a general K-fold CV, but it requires K

to depend on sample size n and involves complicated specification of the rate of K going to ∞ as $n \to \infty$, because theoretically the K-fold CV provides an asymptotic unbiased prediction only when $K \to \infty$.

For fitting models (1) and (2), we need a smoothing method. Different smoothers can be employed, and examples include the local linear smoother (Fan and Zhang, 1999; Fan and Jiang, 2005) and the global polynomial spline smoothing (Stone, 1986; Li and Liang, 2008; Jiang and Jiang, 2011), among other. For illustration, we consider only the global polynomial spline smoothing, which has stable performance and allows one to fast implement in the LOOCV (see the next section).

For model (2), we estimate α by $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$ and use B-spline basis functions to approximate each $m_j(\cdot)$. Without loss of generality, assume $\mathbf{X} = (x_1, \ldots, x_p)^\top$ takes values in $\mathcal{W} = [0, 1]^p$. For approximating function $m_j(\cdot)$, we need a knot sequence $\bar{\phi}_j = \{\phi_{j,k}\}_{k=0}^{q_j+1}$ such that $0 = \phi_{j,0} < \phi_{j,1} < \cdots < \phi_{j,q+1} = 1$. Denoted by $\mathcal{S}(\ell_j, \bar{\phi}_j)$ the space of polynomial splines of order ℓ_j and knot sequence $\bar{\phi}_j$. Since $\mathcal{S}(\ell_j, \bar{\phi}_j)$ is a κ_j-dimensional linear space with $\kappa_j = q_j + \ell_j$, for any $m_j \in \mathcal{S}(\ell_j, \bar{\phi}_j)$, there exists a local basis $\{B_{j,k}(\cdot)\}_{k=1}^{\kappa_j}$ for $\mathcal{S}(\ell_j, \bar{\phi}_j)$, such that $m_j(x_j) = \sum_{k=1}^{\kappa_j} b_{jk} B_{j,k}(x_j)$ for $j = 1, \ldots, p$ (Schumaker, 1981; Jiang and Jiang, 2011). The local basis $\{B_{j,k}(\cdot)\}_{k=1}^{\kappa_j}$ depends on the knot sequence $\bar{\phi}_j$ and order ℓ_j. Let $\mathbf{b}_j = (b_{j1}, \ldots, b_{j\kappa_j})^\top$, $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_p^\top)^\top$, $\mathbf{\Pi}_j(x_j) = (B_{j,1}(x_j), \ldots, B_{j,\kappa_j}(x_j))^\top$, and $\mathbf{\Pi}(\mathbf{X}) = (\mathbf{\Pi}_1^\top(x_1), \ldots, \mathbf{\Pi}_p^\top(x_p))^\top$. For simplicity, denoted by $\mathrm{y}_i = y_i - \bar{y}$. For any $1 \leq i \leq n$, we minimize the approximated sum of squared errors

$$\sum_{j=1(\neq i)}^{n} \{\mathrm{y}_j - \mathbf{\Pi}(\mathbf{X}_j)^\top \mathbf{b}\}^2 \tag{7}$$

over $\mathbf{b}$, which leads to the minimizer

$$\hat{\mathbf{b}}^{(-i)} = \{\mathbf{\Sigma}_n^{(-i)}\}^{-1} \{\frac{1}{n-1} \sum_{j=1(\neq i)}^{n} \mathbf{\Pi}(\mathbf{X}_j) \mathrm{y}_j\}, \tag{8}$$

where $\mathbf{\Sigma}_n^{(-i)} = \frac{1}{n-1} \sum_{j=1(\neq i)}^{n} \mathbf{\Pi}(\mathbf{X}_j) \mathbf{\Pi}(\mathbf{X}_j)^\top$. Let $\hat{m}(\mathbf{X}_i) = \mathbf{\Pi}(\mathbf{X}_i)^\top \hat{\mathbf{b}}^{(-i)}$. Then the prediction

error of model (2) is given by

$$\hat\varepsilon_{2,i} = \mathrm{y}_i - \hat m(\mathbf{X}_i). \tag{9}$$

For functional coefficient model (1), we also assume z_i takes values in $[0,1]$. Similarly, there exits a local basis $\{B_{j,k}(\cdot)\}_{k=1}^{\widetilde\kappa_j}$ such that $\beta_j(z) = \sum_{k=1}^{\widetilde\kappa_j} B_{j,k}(z)c_{jk}$ for $j = 0, 1, \ldots, p$. Let $\mathbf{c}_j = (c_{j1}, \ldots, c_{j\widetilde\kappa_j})^\top$, $\mathbf{c} = (\mathbf{c}_0^\top, \mathbf{c}_1^\top, \ldots, \mathbf{c}_p^\top)^\top$, $\mathbf{\Gamma}_j(z) = (B_{j,1}(z), \ldots, B_{j,\widetilde\kappa_j}(z))^\top$, and $\mathbf{\Gamma}(\mathbf{X}, z) = (\mathbf{\Gamma}_0^\top(z), x_1\mathbf{\Gamma}_1^\top(z), \ldots, x_p\mathbf{\Gamma}_p^\top(z))^\top$. For any $1 \le i \le n$, we minimize

$$\sum_{j=1(\ne i)}^{n} \{y_j - \mathbf{\Gamma}(\mathbf{X}_j, z_j)^\top \mathbf{c}\}^2$$

over $\mathbf{c}$ and get the minimizer $\hat{\mathbf{c}}^{(-i)} = \{\mathbf{G}_n^{(-i)}\}^{-1}\{\frac{1}{n-1}\sum_{j=1(\ne i)}^{n-1} \mathbf{\Gamma}(\mathbf{X}_j, z_j)y_j\}$, where $\mathbf{G}_n^{(-i)} = \frac{1}{n-1}\sum_{j=1(\ne i)}^{n} \mathbf{\Gamma}(\mathbf{X}_j, z_j)\mathbf{\Gamma}(\mathbf{X}_j, z_j)^\top$. Then the prediction error of model (1) is

$$\hat\varepsilon_{1,i} = y_i - \mathbf{\Gamma}(\mathbf{X}_i, z_i)^\top \hat{\mathbf{c}}^{(-i)}. \tag{10}$$

Given $\hat\xi_i = \hat\varepsilon_{1,i}^2 - \hat\varepsilon_{2,i}^2$, we can calculate the ELR statistic $R_{n,1}$ in (5). The following theorem describes its asymptotic null distribution.

Theorem 1. Assume that conditions A1 - A3 in Appendix A hold. Under $H_0^{(1)}$, $R_{n,1} \to \chi_1^2$ in distribution, where χ_1^2 is the chi-squared distribution with one degree of freedom.

Let $\chi_{1,1-\alpha}^2$ be the $(1-\alpha)$th percentile of χ_1^2. By Theorem 1, at significance level α, the rejection region of the ELR test is $W = \{R_{n,1} > \chi_{1,1-\alpha}^2\}$. To investigate the power of the proposed test, we consider the contiguous alternative of form:

$$H_{a,n}^{(1)} : \mu_\xi = a_n\sigma_\xi n^{-1/2}, \tag{11}$$

where $\mu_\xi = E(\hat\xi_1)$, σ_ξ is the standard deviation of $\hat\xi_1$ and greater than zero, and a_n is a sequence of real numbers such that $\lim_{n\to\infty} a_n = a$. Then the power of the ELR test can be approximated using the following theorem.

Theorem 2. Suppose conditions A1 - A3 hold. Under $H_{a,n}^{(1)}$, $R_{n,1} \to \chi_1^2(a^2)$ if $|a| < +\infty$, and $P(R_{n,1} \to +\infty) \to 1$ if $|a| = +\infty$, where $\chi_1^2(a^2)$ is the noncentral chi-squared distribution with one degree of freedom and noncentral parameter a^2.

Theorems 1-2 have an interesting implication. We can approximate the power of the test by

$$P_{H_{a,n}^{(1)}}(W) \approx P(\chi_1^2(a^2) > \chi_{1,1-\alpha}^2) = 1 - \{\Phi(|a| + \sqrt{\chi_{1,1-\alpha}^2}) - \Phi(|a| - \sqrt{\chi_{1,1-\alpha}^2})\},$$

where $\Phi(\cdot)$ is the distribution function of $\mathcal{N}(0,1)$. The power function is increasing in $|a|$ and shares the same formula as that of the likelihood ratio test for testing $H_0 : \mu = 0$ against $H_{1n} : \mu = a_n \sigma n^{-1/2}$, based on an iid sample of size n from the normal population $\mathcal{N}(\mu, \sigma^2)$. This suggests that the proposed test is powerful.

From Theorems 1-2, given a significant level α, we conduct a model selection procedure based on the following decision rule:

(i) If $R_{n,1} < \chi_{1,1-\alpha}^2$, then we cannot reject $H_0^{(1)} : \mu_\xi = 0$, and we say the two models are asymptotically equivalent.

(ii) If $R_{n,1} > \chi_{1,1-\alpha}^2$, one model is sufficiently better than the other. Furthermore,

    (a) if $APE_1 < APE_2$, model (1) is better than model (2);

    (b) if $APE_1 > APE_2$, model (2) is better than model (1).

Our ELR test is asymptotically chi-squared under H_0 that the two models are forecast equivalent, no matter if the models are nested, nonnested, overlapping, correctly specified, or misspecified. As shown in Theorem 2, it has nontrivial power against all local alternatives $H_{a,n}^{(1)}$ which converge to the null at rate $\sqrt{n}$ or faster ($|a| \le +\infty$). Unlike Vuong's type of tests, we do not pretest which distribution to use for calculating the critical value, and thus the ELR test can uniformly control the size of test as in Shi (2015) and Schennach and Wilhelm (2017). We test the forecast equivalence against non-equivalence of the two models. If the null is rejected, we retain the model with smaller APE; otherwise, we believe both models provide equal forecast performance. In any cases, we make this kind

of conclusions, no matter if the models are nested or not and misspecified or not, which is consistent with the framework of likelihood inference under model misspecification in White (1982b).

2.4 Fast implementation

The ELR test involves the LOOCV for calculating the prediction errors, which requires fitting the models to each subsample with one observation held out. In the following we introduce a fast algorithm for computing the prediction errors.

Define the projection matrices $\mathbf{P}_A = \mathbf{D}(\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top$ and $\mathbf{P}_V = \mathbf{E}(\mathbf{E}^\top \mathbf{E})^{-1} \mathbf{E}^\top$, where $\mathbf{D} = (\mathbf{\Pi}(\mathbf{X}_1), \ldots, \mathbf{\Pi}(\mathbf{X}_n))^\top$ and $\mathbf{E} = (\mathbf{\Gamma}(\mathbf{X}_1, z_1), \ldots, \mathbf{\Gamma}(\mathbf{X}_n, z_n))^\top$. Let the residual vectors be $\mathbf{e}_1 = (e_{1,1}, \ldots, e_{1,n})^\top = \mathbf{Y} - \mathbf{P}_V \mathbf{Y}$ and $\mathbf{e}_2 = (e_{2,1}, \ldots, e_{2,n})^\top = \mathbf{Y} - \mathbf{P}_A \mathbf{Y}$, where $\mathbf{Y} = (y_1, \ldots, y_n)$ and $\mathbf{Y} = (y_1, \ldots, y_n)^\top$. Then, using the classical technique in linear models, we obtain that

$$\hat{\mathbf{b}}^{(-i)} = \hat{\mathbf{b}} - (1 - p_{A,i})^{-1}(\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{\Pi}(\mathbf{X}_i) e_{2,i},$$

$$\hat{\mathbf{c}}^{(-i)} = \hat{\mathbf{c}} - (1 - p_{V,i})^{-1}(\mathbf{E}^\top \mathbf{E})^{-1} \mathbf{\Gamma}(\mathbf{X}_i, z_i) e_{1,i},$$

where $\hat{\mathbf{b}} = (\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top \mathbf{Y}$, $\hat{\mathbf{c}} = (\mathbf{E}^\top \mathbf{E})^{-1} \mathbf{E}^\top \mathbf{Y}$, and $p_{A,i}$ and $p_{V,i}$ are the ith diagonal entries of the matrices $\mathbf{P}_A$ and $\mathbf{P}_V$, respectively. Furthermore,

$$\hat{\varepsilon}_{1,i} = (1 - p_{V,i})^{-1} e_{1,i} \text{ and } \hat{\varepsilon}_{2,i} = (1 - p_{A,i})^{-1} e_{2,i}. \tag{12}$$

Hence, $\hat{\xi}_i = \hat{\varepsilon}_{1,i}^2 - \hat{\varepsilon}_{2,i}^2$ can be calculated by fitting the models to full data only. Since $f'(\lambda) = \sum_{i=1}^n \hat{\xi}_i^2 / (1 + \lambda \hat{\xi}_i)^2 < 0$, $f(\lambda)$ is strictly decreasing. Then evaluation of the ELR statistic $R_{n,1}$ is straightforward while solving equation (4) to obtain $\hat{\lambda}$ by the Newton-Raphson iterations with initial value $\lambda = 0$.

3 Application to big data

In most cases the sample size of big data is huge, and existing statistical inference methods face up to challenges. Consider fitting a linear model with big data, for example, the p value of a t-statistic for an individual coefficient is possibly less than 5%. Since no model is right, the p value goes to zero as sample size n goes to ∞, no matter how small the coefficient is. Even if the model is correct, the p value may be very small for a nonzero coefficient as sample size n gets large enough.

However, for a very small coefficient the corresponding covariate may not be of practical importance at all. That is, statistical significance may not imply practical importance in a big data setting. Naturally, one may ask how to measure practical importance of a covariate if there is a huge sample. In other words, we need some measures to calibrate the importance of explanatory variables (or their functional forms), rather than merely assessing their statistical significance. This is expected to be a challenge in statistical analysis for massive data where the memory of one machine cannot fit all the data.

3.1 Distributed ELR test

As we discussed before, no model is right, and any statistical model can be misspecified in practice. Therefore, it will make much more sense to make comparison between misspecified models than concentrating on statistical significance of covariates in big data settings. This is particularly relevant to economic modeling, because it is possible that more than one economic models, some of which can be even conflicting to each other, coexist in explaining the same economic phenomenon, and the existing econometric tools cannot distinguish them from each other for various reasons. Obviously, our ELR test can be used for this task, and in particular it can be used to compare the two models with or without an explanatory variable.

Consider modeling a massive dataset, for example, using the additive model (2). To evaluate importance of the ℓth variable for forecast, we compare model (2) with

$$y_i = \alpha + \sum_{j=1(\neq \ell)}^{p} m_j(x_{i,j}) + v_i, \quad i = 1, \ldots, n. \tag{13}$$

Let $\mathbf{\Pi}_{-\ell}(\mathbf{X}) = (\mathbf{\Pi}_1^\top(x_1), \ldots, \mathbf{\Pi}_{\ell-1}^\top(x_{\ell-1}), \mathbf{\Pi}_{\ell+1}^\top(x_{\ell+1}), \ldots, \mathbf{\Pi}_p^\top(x_p))^\top$ and

$$\mathbf{\Sigma}_{n,-\ell}^{(-i)} = \frac{1}{n-1} \sum_{j=1(\neq i)}^{n} \mathbf{\Pi}_{-\ell}(\mathbf{X}_j)\mathbf{\Pi}_{-\ell}(\mathbf{X}_j)^\top.$$

Then, similar to (8), the spline coefficient $\mathbf{b}$ is estimated by

$$\hat{\mathbf{b}}_{-\ell}^{(-i)} = \{\mathbf{\Sigma}_{n,-\ell}^{(-i)}\}^{-1}\{\frac{1}{n-1} \sum_{j=1(\neq i)}^{n} \mathbf{\Pi}_{-\ell}(\mathbf{X}_j)\mathbf{y}_j\}. \tag{14}$$

Similar to (9), we obtain the prediction errors from model (13):

$$\hat{\varepsilon}_{3,i} = \mathbf{y}_i - \mathbf{\Pi}_{-\ell}(\mathbf{X}_i)^\top \hat{\mathbf{b}}_{-\ell}^{(-i)}. \tag{15}$$

Then the difference of squared prediction errors between model (13) and (2) are given by $\hat{\eta}_i = \hat{\varepsilon}_{3,i}^2 - \hat{\varepsilon}_{2,i}^2$, which can be calculated quickly using the same technique as in (12). Similar to (6), comparing model (2) to model (13) reduces to testing

$$H_0^{(2)} : \mu_\eta = 0 \text{ against } H_a^{(2)} : \mu_\eta \neq 0, \tag{16}$$

where $\mu_\eta = E(\hat{\eta}_1)$. Using the same argument as for $R_{n,1}$, we obtain the ELR statistic

$$R_{n,2} = 2\sum_{i=1}^{n} \log(1 + \hat{\nu}\hat{\eta}_i), \tag{17}$$

where $\hat{\nu}$ satisfies that $\sum_{i=1}^{n} \hat{\eta}_i/(1 + \hat{\nu}\hat{\eta}_i) = 0$. As argued for $R_{n,1}$, if $0 \notin [\min \hat{\eta}_i, \max \hat{\eta}_i]$, we set $R_{n,2} = +\infty$.

Large values suggest rejection of $H_0^{(2)}$. If $H_0^{(2)}$ is rejected, then it suggests that the ℓth covariate is practically important for forecasting the response. This is a variable selection problem in which both models are nested but may be misspecified. Existing approaches deal with it by assuming the larger model is correctly specified. However, our ELR test does not require this condition.

Some challenges arise when we use $R_{n,2}$ for massive or distributed data. Practically, we need to solve the computation problem since the classical computation methods for estimating m_j's and for empirical likelihood ratio (Hall and La Scala, 1990) are computationally infeasible for massive data. We need to develop some distributed computing methods to solve this problem. The existing divide-and-conquer method (Zhang et al., 2013; Chen and Xie, 2014; Chen et al., 2019; Battey et al., 2018) can be employed in general, but the resulting limiting distribution of the test statistic should be consistent with that of the original test with full data, or some other inference methods such as the bootstrap procedure adaptive massive data (Chen and Peng, 2018) are to be advanced. In the following we will work on these problems and provide a distributed ELR test for nonnested model selection with massive data. Remarkably, our distributed test will perform the same as the original test.

Suppose we have a massive sample of size $n = Nm$. Then we randomly split the entire dataset $\{y_i, \mathbf{X}_i, 1 \le i \le n\}$ into N subsamples $\mathcal{D}_1, \ldots, \mathcal{D}_N$, each of which has the same size $m = n/N$. For distributed data, the full sample consists of these subsamples installed on N machines at different sites. If different machines have different subsample sizes, our procedure can be straightforwardly extended. Typically, using the divide-and-conquer algorithm one fits the models with each subsample on each machine and gets the prediction error for each subsample point on each machine, and integrates them to form the ELR test. Since each prediction error uses information from only a subsample, the resulting ELR test will not be as powerful as the original ELR test with full data. Even if one calculates the prediction error with full sample information, the resulting ELR test will not have the same finite sample performance as the original ELR test. Instead of fitting the models to the subsample on every machine, we calculate only some sufficient statistics from each subsample and use them to estimate the spline coefficients. Then the estimated coefficients are feedback to each slaver so that the LOOCV error can be calculated.

Specifically, let $\mathcal{D}_k = \{y_j^{(k)}, \mathbf{X}_j^{(k)}, j = 1, \ldots, m\}$ be the subsample distributed on the kth machine for $1 \le k \le N$. Then there exists a one to one mapping $\nu : \{1, \ldots, m\} \otimes \{1, \ldots, N\} \to \{1, \ldots, n\}$ such that

$$i = \nu(j, k) \quad \text{and} \quad (y_j^{(k)}, \mathbf{X}_j^{(k)}) = (y_i, \mathbf{X}_i) \quad \text{for} \quad i = 1, \ldots, n. \tag{18}$$

For example, $\nu(j, k) = j + (k - 1)m$ is such a mapping.

Note that the B-spline basis vector $\mathbf{\Pi}(\mathbf{X}_i)$ depends on the knot sequences $\{\bar{\phi}_j\}_{j=1}^p$ with $\bar{\phi}_j = \{\phi_{j,k}\}_{k=0}^{q_j+1}$. For each subset $\mathcal{D}_k$, we can compute $\{\mathbf{\Pi}(\mathbf{X}_j^{(k)}), j = 1, \ldots, m\}$ and $\{\mathbf{\Pi}_{-\ell}(\mathbf{X}_j^{(k)}), j = 1, \ldots, m\}$ with some given knot sequences $\{\bar{\phi}_j\}_{j=1}^p$ independent of k. Choice of such knot sequences for massive or distributed data will be discussed in Section 3.2. It follows from (18) that, for $\alpha = 0, \ell$,

$$\mathbf{\Pi}_{-\alpha}(\mathbf{X}_j^{(k)}) = \mathbf{\Pi}_{-\alpha}(\mathbf{X}_i),$$

where, with a little abuse of notations, we denote $\mathbf{\Pi}(\mathbf{X}_i)$ by $\mathbf{\Pi}_{-0}(\mathbf{X}_j^{(k)})$ for convenience. Note that the LOOCV estimators for models (2) and (13) involve only statistics:

$$\mathcal{X}_\alpha^{(-i)} \equiv \sum_{s=1(\ne i)}^n \mathbf{\Pi}_{-\alpha}(\mathbf{X}_s)\mathbf{\Pi}_{-\alpha}(\mathbf{X}_s)^\top \quad \text{and} \quad \mathfrak{F}_\alpha^{(-i)} \equiv \sum_{s=1(\ne i)}^n \mathbf{\Pi}_{-\alpha}(\mathbf{X}_s)y_s$$

for $i = 1, \ldots, n$. Let

$$\mathcal{A}_\alpha = \sum_{k=1}^N A_{-\alpha}^{(k)} \quad \text{and} \quad \mathcal{B}_\alpha = \sum_{k=1}^N B_{-\alpha}^{(k)},$$

where $A_{-\alpha}^{(k)} = \sum_{s=1}^m \mathbf{\Pi}_{-\alpha}(\mathbf{X}_s^{(k)})\mathbf{\Pi}_{-\alpha}(\mathbf{X}_s^{(k)})^\top$ and $B_{-\alpha}^{(k)} = \sum_{s=1}^m \mathbf{\Pi}_{-\alpha}(\mathbf{X}_s^{(k)})y_s^{(k)}$ are sufficient statistics for the subsample on the kth machine. These sufficient statistics can be calculated on individual machines. Then

$$\mathcal{X}_\alpha^{(-i)} = \mathcal{A}_\alpha - \mathbf{\Pi}_{-\alpha}(\mathbf{X}_j^{(k)})\mathbf{\Pi}_{-\alpha}(\mathbf{X}_j^{(k)})^\top \quad \text{and} \quad \mathfrak{F}_\alpha^{(-i)} = \mathcal{B}_\alpha - \mathbf{\Pi}_{-\alpha}(\mathbf{X}_j^{(k)})y_j^{(k)}.$$

Hence, the LOOCV estimators in (8) and (14) are rewritten as

$$\hat{\mathbf{b}}_{\alpha,j}^{(k)} = \{\mathcal{A}_\alpha - \mathbf{\Pi}_{-\alpha}(\mathbf{X}_j^{(k)})\mathbf{\Pi}_{-\alpha}(\mathbf{X}_j^{(k)})^\top\}^{-1}\{\mathcal{B}_\alpha - \mathbf{\Pi}_{-\alpha}(\mathbf{X}_j^{(k)})y_j^{(k)}\},$$

18

respectively for $\alpha = 0, \ell$. Then distributed prediction errors from model (2) and (13) are given by

$$\hat{\varepsilon}_{2,j}^{(k)} = \mathbf{y}_j^{(k)} - \mathbf{\Pi}_{-0}(\mathbf{X}_j^{(k)})^\top \hat{\mathbf{b}}_{0,j}^{(k)} \quad \text{and} \quad \hat{\varepsilon}_{3,j}^{(k)} = \mathbf{y}_j^{(k)} - \mathbf{\Pi}_{-\ell}(\mathbf{X}_j^{(k)})^\top \hat{\mathbf{b}}_{\ell,j}^{(k)},$$

respectively. Therefore, for $\mathbf{y}_j^{(k)}$ the difference of squared prediction errors from models (13) and (2) is given by $\hat{\eta}_j^{(k)} = |\hat{\varepsilon}_{3,j}^{(k)}|^2 - |\hat{\varepsilon}_{2,j}^{(k)}|^2$. Then, similar to (17), we obtain the distributed ELR statistic

$$R_{n,3} = 2 \sum_{k=1}^{N} \sum_{j=1}^{m} \log(1 + \hat{\tau}\hat{\eta}_j^{(k)}), \tag{19}$$

where $\hat{\tau}$ satisfies that $\sum_{k=1}^{N} \sum_{j=1}^{m} \hat{\eta}_j^{(k)}/(1 + \hat{\tau}\hat{\eta}_j^{(k)}) = 0$. Again, if $0 \notin [\min \hat{\eta}_j^{(k)}, \max \hat{\eta}_j^{(k)}]$, we set $R_{n,3} = +\infty$. The root $\hat{\tau}$ of the above equation can be found via the distributed Newton-Raphson iterations, that is, be implemented on each machine.

Since $\hat{\eta}_j^{(k)} = \hat{\eta}_i$, we have $R_{n,2} = R_{n,3}$. That is, they have the same finite-sample and asymptotic performance, and distributed test $R_{n,3}$ has the same power as ideal test $R_{n,2}$ with no memory constraint. The following theorem depicts the asymptotic null distributions of the ELR tests.

Theorem 3. Suppose conditions A2 and A4 hold. Then, under $H_0^{(2)}$, $R_{n,3} \to \chi_1^2$ in distribution.

To study the power of test, we consider testing $H_0^{(2)}$ against a sequence of contiguous alternatives:

$$H_{a,n}^{(2)} : \mu_\eta = a_n \sigma_\eta n^{-1/2}, \tag{20}$$

where $\mu_\eta = E(\hat{\eta}_1)$, σ_η is the standard deviation of $\hat{\eta}_1$ and greater than zero, and a_n is the same as in (11). In the following we present the alternative distributions of the ELR tests.

Theorem 4. Assume that conditions A2 and A4 hold. Then, under $H_{a,n}^{(2)}$, $R_{n,3} \to \chi_1^2(a^2)$ if $|a| < +\infty$, and $P(R_{n,3} \to +\infty) \to 1$ if $|a| = +\infty$.

From Theorems 3-4, given a significant level α, we conduct a variable selection procedure based on the following decision rule:

(i) If $R_{n,3} < \chi^2_{1,1-\alpha}$, then we cannot reject $H_0^{(2)}$. According to Occam's razor, we choose model (13) without the ℓth covariate.

(ii) If $R_{n,3} > \chi^2_{1,1-\alpha}$, it suggests the ℓth covariate is practically important for forecasting the response. We choose model (2) as the working model.

In the above decision rule, we choose model (2) when the null is rejected. This agrees with choosing the working model close to the true. Since model (2) is larger, it is closer to the true than model (13).

3.2   Knot selection with massive data

There are two popular ways of deciding the knots. One is to place equally spaced knot sequence, and the other is to use the quantile knot sequence from the empirical distribution of the underlying variable. The 1st knot choice can be easily computed since it is independent of the data. For the 2nd knot choice, it seems that the sample quantiles of the x-variables for massive or distributed data are not easy to get, but the median-searching algorithm in Harris (2012) can be adapted to the current situation. Specifically, we consider how to get the qth quantile of the sample $\{x_{i,1}, i = 1, \ldots, n\}$ for any $q \in (0,1)$. Let $x_{(1),1} \le x_{(2),1} \le \cdots \le x_{(n),1}$ be the order statistics. Then, the qth sample quantile is defined by

$$x_{(\lfloor h \rfloor),1} + (h - \lfloor h \rfloor)\{x_{(\lceil h \rceil),1} - x_{(\lfloor h \rfloor),1}\}, \tag{21}$$

where $\lfloor h \rfloor$ (or $\lceil h \rceil$) denotes the nearest integer to $h \equiv (n-1)q + 1$, which is less (or larger) than h. This is the default way of defining sample quantile in software R, and is equivalent to the Excel and Python optional "inclusive" methods.

According to (21), it suffices to find $x_{(l),1}$ for any $1 \le l \le n$. Let us split sample $\{x_{i,1}\}_{i=1}^n$ into N sets $\mathcal{F}_k = \{x_{j,1}^{(k)}, j = 1, \ldots, m\}$ for $1 \le k \le N$. Then our distributed algorithm proceeds as follows:

(i) Randomly select a set $\mathcal{F}_k$ and an element $a \in \mathcal{F}_k$;

(ii) Compute subset $\mathcal{C}_{k'} = \{x \in \mathcal{F}_{k'} : x \leq a\}$ for $k' = 1, \ldots, N$ and $\mathcal{N} = \sum_{k'=1}^{N} |\mathcal{C}_{k'}|$ with $|\mathcal{C}_{k'}|$ being the number of elements of $\mathcal{C}_{k'}$;

(iii) If $\mathcal{N} = l$, the algorithm stops and returns a as $x_{(l),1}$;

(iv) If $\mathcal{N} > l$, renew $\mathcal{F}_{k'} = \mathcal{C}_{k'}$ for $k' = 1, \ldots, N$ and go to step (i);

(v) If $\mathcal{N} < l$, renew $\mathcal{F}_{k'} = \mathcal{F}_{k'}/\mathcal{C}_{k'}$ for $k' = 1, \ldots, N$ and $l = l - \mathcal{N}$, and go to step (i).

As mentioned in Harris (2012), the computational complexity of the above algorithm is $O\{n/N + N \log(n/N)\}$. When $N \ll \sqrt{n}$, the computational complexity is simply $O(n/N)$, which decreases as N increases. For the split-and-conquer method, it usually assumes the technical condition $N \ll \sqrt{n}$, but our method works for any $1 \leq N \leq n$.

3.3 A distributed ELR algorithm

Computational details of the distributed ELR algorithm is listed in Algorithm 1. Note that the full sample estimators of coefficient vector $\mathbf{b}$ for models (2) and (13) are $\hat{\mathbf{b}}_{-\alpha} = \mathcal{A}_{\alpha}^{-1}\mathcal{B}_{\alpha}$, respectively for $\alpha = 0, \ell$. Let $e_{2,j}^{(k)} = \mathrm{y}_j^{(k)} - \mathbf{\Pi}(\mathbf{X}_j^{(k)})^{\top}\hat{\mathbf{b}}_{-0}$ and $e_{3,j}^{(k)} = \mathrm{y}_j^{(k)} - \mathbf{\Pi}_{-\ell}(\mathbf{X}_j^{(k)})^{\top}\hat{\mathbf{b}}_{-\ell}$. Then, similar to (12), we have the following fast calculation formulas for the LOOCV prediction errors:

$$\hat{\varepsilon}_{2,j}^{(k)} = (1 - p_{0,j}^{(k)})^{-1}e_{2,j}^{(k)} \quad \text{and} \quad \hat{\varepsilon}_{3,j}^{(k)} = (1 - p_{\ell,j}^{(k)})^{-1}e_{3,j}^{(k)},$$

where $p_{\alpha,j}^{(k)} = \mathbf{\Pi}_{-\alpha}(\mathbf{X}_j^{(k)})^{\top}\mathcal{A}_{\alpha}^{-1}\mathbf{\Pi}_{-\alpha}(\mathbf{X}_j^{(k)})$ for $\alpha = 0, \ell$. The above formulas have been incorporated into Algorithm 1. In this algorithm, we calculate the sufficient statistics $A_{-\alpha}^{(k)}$ and $B_{-\alpha}^{(k)}$ on individual machines, with which we obtain $\mathcal{A}_{\alpha}$ and $\mathcal{B}_{\alpha}$. Then the full sample coefficient estimator $\hat{\mathbf{b}}_{-\alpha}$ and the LOOCV prediction errors are calculated. At last, we evaluate test statistic $R_{n,3}$.

Algorithm 1: Distributed ELR algorithm

Input: $\{y_i, \mathbf{X}_i, 1 \leq i \leq n\}$, n, m, N, $\tau_0 = 0$, $\{\bar{\phi}_j\}_{j=1}^p$, ω and φ;

Output: $R_{n,3}$;

Initialization: Randomly partition $\{y_i, \mathbf{X}_i, 1 \leq i \leq n\}$ into N subsets $\{y_j^{(k)}, \mathbf{X}_j^{(k)}, j = 1, \ldots, m\}$ for $1 \leq k \leq N$ and distribute them on N machines;

Circulation: for $k = 1 : N$ **do**

> With the knot sequences $\{\bar{\phi}_j\}_{j=1}^p$, compute B-spline basis $\{\mathbf{\Pi}(\mathbf{X}_s^{(k)}), s = 1, \ldots, m\}$ and $\{\mathbf{\Pi}_{-\ell}(\mathbf{X}_s^{(k)}), s = 1, \ldots, m\}$;
>
> $A_{-0}^{(k)} = \sum_{s=1}^m \mathbf{\Pi}(\mathbf{X}_s^{(k)})\mathbf{\Pi}(\mathbf{X}_s^{(k)})^\top$ and $A_{-\ell}^{(k)} = \sum_{s=1}^m \mathbf{\Pi}_{-\ell}(\mathbf{X}_s^{(k)})\mathbf{\Pi}_{-\ell}(\mathbf{X}_s^{(k)})^\top$;
>
> $B_{-0}^{(k)} = \sum_{s=1}^m \mathbf{\Pi}(\mathbf{X}_s^{(k)})y_s^{(k)}$ and $B_{-\ell}^{(k)} = \sum_{s=1}^m \mathbf{\Pi}_{-\ell}(\mathbf{X}_s^{(k)})y_s^{(k)}$;

end

For $\alpha = 0, \ell$, compute $\mathcal{A}_\alpha = \sum_{k=1}^N A_{-\alpha}^{(k)}$, $\mathcal{B}_\alpha = \sum_{k=1}^N B_{-\alpha}^{(k)}$, $\hat{\mathbf{b}}_{-\alpha} = \mathcal{A}_\alpha^{-1}\mathcal{B}_\alpha$;

Circulation: for $k = 1 : N, j = 1 : m$ **do**

> $e_{2,j}^{(k)} = y_j^{(k)} - \mathbf{\Pi}(\mathbf{X}_j^{(k)})^\top\hat{\mathbf{b}}_{-0}$, $e_{3,j}^{(k)} = y_j^{(k)} - \mathbf{\Pi}_{-\ell}(\mathbf{X}_j^{(k)})^\top\hat{\mathbf{b}}_{-\ell}$;
>
> $p_{0,j}^{(k)} = \mathbf{\Pi}(\mathbf{X}_j^{(k)})^\top\mathcal{A}_0^{-1}\mathbf{\Pi}(\mathbf{X}_j^{(k)})$, $p_{\ell,j}^{(k)} = \mathbf{\Pi}_{-\ell}(\mathbf{X}_j^{(k)})^\top\mathcal{A}_\ell^{-1}\mathbf{\Pi}_{-\ell}(\mathbf{X}_j^{(k)})$;
>
> $\hat{\varepsilon}_{2,j}^{(k)} = (1 - p_{0,j}^{(k)})^{-1}e_{2,i}^{(k)}$, $\hat{\varepsilon}_{3,j}^{(k)} = (1 - p_{\ell,j}^{(k)})^{-1}e_{3,j}^{(k)}$;
>
> $\hat{\eta}_j^{(k)} = |\hat{\varepsilon}_{3,j}^{(k)}|^2 - |\hat{\varepsilon}_{2,j}^{(k)}|^2$;

end

if $\min_{j,k} \hat{\eta}_j^{(k)} \leq 0 \leq \max_{j,k} \hat{\eta}_j^{(k)}$ **then**

> **Iteration: for** $t = 1 : \omega$ **do**
>
>> **Circulation: for** $k = 1 : N$ **do**
>>
>>> $D_{1k} = \sum_{j=1}^m \frac{\hat{\eta}_j^{(k)}}{1 + \tau_{t-1}\hat{\eta}_j^{(k)}}$, $D_{2k} = \sum_{j=1}^m \frac{|\hat{\eta}_j^{(k)}|^2}{|1 + \tau_{t-1}\hat{\eta}_j^{(k)}|^2}$;
>>
>> **end**
>>
>> $D_1 = \sum_{k=1}^N D_{1k}$, $D_2 = -\sum_{k=1}^N D_{2k}$, $\tau_t = \tau_{t-1} - D_1/D_2$;
>>
>> **if** $|\tau_t - \tau_{t-1}| < \varphi$ **then**
>>
>>> $\hat{\tau} = \tau_t$;
>>>
>>> **Circulation: for** $k = 1 : N$ **do**
>>>
>>>> $R_{n,3}^{(k)} = 2\sum_{i=1}^m \log(1 + \hat{\tau}\hat{\eta}_i^{(k)})$
>>>
>>> **end**
>>>
>>> $R_{n,3} = \sum_{k=1}^N R_{n,3}^{(k)}$;
>>
>> **end**
>
> **end**

else

> $R_{n,3} = 10^{16}$;

end

Return $R_{n,3}$.

---

4 Simulations

To investigate the size and power of our ELR test for nonnested models, we conduct simulations for model selection in different situations. For each of the following examples, we run 600 simulations, and for each simulation we generated an iid sample from the DGP. For each simulation, the cubic B-splines were used to estimate the unknown functions in the working models, and the leave-one-out CV method was employed to calculate the APE. The number of knots is chosen by an adjusted APE criterion. Specifically, for model (2) $\{\kappa_j, j = 1, \ldots, p\}$ were chosen by minimizing the adjusted APE

$$APE_{adj} = \frac{1}{n - \kappa} \sum_{i=1}^{n} |\hat{\varepsilon}_{2,i}|^2,$$

where $\kappa = \sum_{j=1}^{p} \kappa_j$. For models (1) and (13), the APE_{adj} is defined similarly but with $(\hat{\varepsilon}_{2,i}, \kappa)$ replaced by $(\hat{\varepsilon}_{1,i}, \tilde{\kappa})$ and $(\hat{\varepsilon}_{3,i}, \kappa - \kappa_\ell)$, respectively.

In Example 1, we investigate if the proposed ELR test works for mis-specified nonnested models with heteroscedasticity. In Example 2, we compare our test with the GLR test in Fan and Jiang (2005, JASA) and the uniform Vuong (UNV) test in Liao and Shi (2020), we also study robustness of these tests. In Example 3, we consider our distributed ELR test for massive data.

Example 1. Consider model selection between varying-coefficient model $y_i = \beta_0(z_i) + \beta_1(z_i)x_{i,1} + \beta_2(z_i)x_{i,2} + u_i$ and additive model $y_i = \alpha + m_1(x_{i,1}) + m_2(x_{i,2}) + v_i$, with iid samples generated from

$$
\begin{aligned}
y_i &= 0.5(x_{i,1} + x_{i,2}) + \theta\{x_{i,1}\exp(1 + z_i) + x_{i,2}1(z_i > 0.5) + 1.5\cos(\pi z_i)\} \\
&\quad + \tau\{\exp(x_{i,1})\cos(x_{i,1}) + 0.5\sin(x_{i,2})\} + \sin(\pi x_{i,1})\varepsilon_i,
\end{aligned} \tag{22}
$$

where $(x_{i,1}, x_{i,2})$ are bivariate normally distributed with standard normal marginals and correlation coefficient 0.5, $z_i \sim U(0, 1)$, and $\varepsilon_i \sim N(0, 1)$.

The varying coefficient models and the additive model are nonnested. When $\theta = \tau = 0$, both models are correctly specified; when $\theta = 0$ and $\tau \neq 0$, the additive model is correctly specified; when $\theta \neq 0$ and $\tau = 0$, the varying coefficient model is correctly specified; when $\theta \neq 0$ and $\tau \neq 0$, both models are misspecified. We set different values of θ and τ to evaluate the size and power of our test. Since the GDP has changing variance, it allows us to evaluate the performance of our ELR test when the error is heteroscedastic.

Table 1: Null rejection rates (%) of ELR tests at significance level 5%(left cell) and 10%(right cell) for Example 1

n	(θ, τ)					
	(0, 0)	(0,0.07)	(0, 0.09)	(0, 0.12)	(0, 0.15)	(0, 0.18)
1000	(4.00,9.00)	(12.3,16.5)	(27.0,34.8)	(55.7,65.8)	(83.8,86.8)	(94.0,95.8)
1500	(5.33,9.33)	(38.3,44.7)	(63.5,70.8)	(87.0,91.2)	(96.5,97.8)	(99.2,100)
Model selection	Both	Additive	Additive	Additive	Additive	Additive
Sign of DAPE		+	+	+	+	+
n	(θ, τ)					
	(0, 0)	(0.05, 0)	(0.075, 0)	(0.1, 0)	(0.125, 0)	(0.15, 0)
1000	(4.00,9.00)	(45.8,58.3)	(68.5,80.7)	(88.7,94.5)	(96.7,98.3)	(99.0,99.3)
1500	(5.33,9.33)	(57.8,71.0)	(80.2,87.7)	(96.0,97.8)	(99.5,99.7)	(100,100)
Model selection	Both	Varying	Varying	Varying	Varying	Varying
Sign of DAPE		−	−	−	−	−
n	(θ, τ)					
	(0.05, 0.05)	(0.18, 0.1)	(0.18,0.05)	(0.05, 0.18)	(0.1, 0.18)	(0.18, 0.18)
1000	(13.3,22.3)	(96.5,98.5)	(99.7,100)	(64.5,74.3)	(16.3,24.0)	(29.8,39.5)
1500	(8.33,14,5)	(99.7,99.8)	(100,100)	(92.5,95.3)	(36.3,46.2)	(26.5,34.8)
Model selection	Varying	Varying	Varying	Additive	Additive	Varying
Sign of DAPE	−	−	−	+	+	−

DAPE - average of the differences of APEs between model (1) and model (2) in 600 simulations

For each paired values of (θ, τ), we calculated the null rejection rates of our ELR tests for testing problem (6) at 5% and 10% significance levels. The simulation results are summarized in Table 1. It is seen that our ELR test uniformly controls size over different significance levels, since the reject rates are all close to the nominal size at $(\theta, \tau) = (0, 0)$. When one of θ and τ goes far away from 0 and the other is fixed at 0, the alternative runs

further away from the null, and the rejection rate becomes higher and higher, which reveals that our test gets more and more powerful. When both θ and τ are nonzero, the two models are nonnested and misspecified, and the power gets higher as the distance between θ and τ increases. This demonstrates that our ELR test works great here. $\diamond$

Table 2: Null rejection rates (%) of the ELR, UNV, and GLR tests for Example 2

DGP	Test	n	τ					
			0	0.06	0.08	0.1	0.12	0.16
normal	ELR	1000	5.17	16.5	39.0	78.3	93.3	97.3
		1500	4.50	26.0	64.0	94.7	98.3	99.2
	UNV	1000	4.17	23.0	49.3	83.8	93.8	98.0
		1500	3.83	34.0	71.8	96.2	99.3	99.8
	GLR	1000	5.83	20.8	45.5	79.5	90.7	96.7
		1500	5.00	34.7	73.7	93.2	99.0	99.7
conditionally normal	ELR	1000	5.17	45.2	79.0	96.8	98.3	99.2
		1500	6.50	71.2	95.5	98.5	99.0	99.7
	UNV	1000	2.80	61.7	88.5	98.2	99.3	99.5
		1500	1.33	84.5	99.0	99.8	100	100
	GLR	1000	14.8	62.8	88.7	97.7	99.5	100
		1500	13.3	79.5	98.3	100	100	100
conditional t(6)	ELR	1000	5.17	23.5	65.2	90.3	96.7	98.3
		1500	5.67	46.8	89.0	98.2	99.3	100
	UNV	1000	2.00	32.3	75.3	94.2	97.8	98.5
		1500	0.83	55.0	93.0	99.2	100	100
	GLR	1000	13.8	40.0	73.5	95.0	98.2	99.8
		1500	12.3	54.0	91.2	99.3	99.5	100
mixed normal	ELR	1000	5.83	12.5	27.0	62.5	83.2	93.2
		1500	5.33	17.2	51.8	88.8	95.7	98.3
	UNV	1000	2.00	12.5	34.3	69.7	85.1	94.5
		1500	2.00	16.7	53.0	88.3	95.6	98.8
	GLR	1000	5.67	15.7	32.0	60.5	81.0	94.0
		1500	5.17	20.2	48.8	82.7	94.2	98.2

Example 2. Let us consider model comparison between varying-coefficient model $y_i = \beta_0(z_i) + \beta_1(z_i)x_{i,1} + u_i$ and additive model $y_i = \alpha + m_1(x_{i,1}) + m_2(x_{i,2}) + v_i$, when the true

DGP is

$$y_i = 0.5x_{i,1} + 0.25x_{i,1}\cos(x_{i,1}) + \tau \exp(x_{i,2})\cos(x_{i,2}) + \varepsilon_i,$$

where $(x_{i,1}, x_{i,2})$ are the same as in Example 1, and ε_i is $N(0,1)$ (normal), $\sin(x_{i,2})N(0,1)$ (conditionally normal), $\sin(x_{i,2})t(6)$ (conditional t(6)), and $0.95N(0,1)+0.05N(0,3^2)$ (mixed normal), respectively. This allows us to assess robustness of our ELR test for model comparison under a variety of noises. We set $z_i = x_{i,1}$ to make the additive model contains the varying coefficient model, so that the GLR test can be applied. These tests are also compared to the UNV test.

Table 2 reports the powers of the three tests. It is seen that the ELR test not only keeps the size but also is nearly most powerful and robust against the error distributions. For the normal and mixed normal errors, our ELR test has nearly the same power as the GLR test. As expected, the GLR and UNV tests cannot keep the size when the errors are heteroscedastic. ◇

Example 3. Consider comparing models $y_i = \alpha + m_1(x_{i,1}) + m_2(x_{i,2}) + v_i$ and $y_i = \alpha + m_2(x_{i,2}) + v_i$ with massive data, when the true DGP is

$$y_i = \tau \exp(x_{i,1})\cos(x_{i,1}) + 0.1x_{i,2}(1 + x_{i,2}) + \varepsilon_i, \tag{23}$$

where ε_i is $\sin(\pi x_{i,2})\mathcal{N}(0,1)$ and $(x_{i,1}, x_{i,2})$ are the same as in Example 1. Obviously, this is a model with heteroscedascity.

We set sample size $n = 21,000$ and $42,000$ for comparing full sample test $R_{n,2}$ with distributed test $R_{n,3}$. Table 3 reports the rejection rate of $H_0^{(2)}$ at 5% significance level. When $\tau = 0$, the null and the alternative coincide, and the power is the type I error probability. As shown in the table, all powers are close to the nominal level 5% at $\tau = 0$, indicating our tests keep the size. As τ increases, the alternative moves further away from the null, and the rejection rate of the null gets higher and higher. Furthermore, for different numbers of machines $N = 1, 50, 100, 150$, $R_{n,2}$ and $R_{n,3}$ have the same performance. This

Table 3: Null rejection rates (%) of ELR tests for Example 3

N	m	τ					
		0	0.01	0.015	0.02	0.025	0.03
1	21000	5.50	15.0	48.0	90.2	98.3	99.5
	42000	5.67	33.3	84.5	99.0	100	100
50	420	5.50	15.0	48.0	90.2	98.3	99.5
	840	5.67	33.3	84.5	99.0	100	100
100	210	5.50	15.0	48.0	90.2	98.3	99.5
	420	5.67	33.3	84.5	99.0	100	100
150	140	5.50	15.0	48.0	90.2	98.3	99.5
	280	5.67	33.3	84.5	99.0	100	100

implies that our distributed ELR tests can exactly recovery the results of the original ELR test with the whole data running on one machine.

5   A real example

We illustrate our method by analyzing the Boston housing dataset. This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston, MA. It is available at the StatLib archive (`http://lib.stat.cmu.edu/datasets/boston`). The dataset consists of median values of owner-occupied homes in 506 homes and several variables that might explain the variation of housing value (Harrison and Rubinfeld, 1978; Fan and Huang, 2005). Fan and Huang (2005) considered the following seven variables: CRIM (per capita crime rate by town), RM (average number of rooms per dwelling), TAX (full-value property-tax rate per $10,000), NOX (nitric oxides concentration parts per 10 million), PTRATIO (pupil-teacher ratio by town), AGE (proportion of owner-occupied units built prior to 1940), and LSTAT (lower status of the population). For simplicity, the variables CRIM, RM, log(TAX), NOX, PTRATIO and AGE are denoted by $x_1, x_2, \ldots, x_6$, respectively. Let y be the response (median value of owner-occupied homes) and $z = $ log(LSTAT).

The object is to study the association between y and $\mathbf{X} = (x_1, x_2, \ldots, x_6)$, given a sample $\{y_i, \mathbf{X}_i, z_i, i = 1, \ldots, n\}$ with size $n = 506$. Many authors analyzed the dataset using different models. Examples include the additive models in Opsomer and Ruppert (1998) and Fan and Jiang (2005), and the varying coefficient model in Fan and Huang (2005), among others. However, there is no formal model comparison among them. In the following we use our ELR test to do this work. In all cases, the significance level is taken as 5%.

(i) (Varying coefficient model vs additive model) Fan and Huang (2005) considered the varying coefficent model:

$$E(y_i | z_i, \mathbf{X}_i) = \beta_0(z_i) + \sum_{j=1}^{6} x_{i,j} \beta_j(z_i). \tag{24}$$

We are interested in further investigating whether the documented "nonlinearity" is the true nonlinearity between y_i and $\mathbf{X}_i$, or is due to the functional coefficients in a linear regression model. Thus, we consider the following nonparametric additive model for comparison:

$$E(y_i - \bar{y} | z_i, \mathbf{X}_i) = m_0(z_i) + \sum_{j=1}^{6} m_j(x_{i,j}), \tag{25}$$

with $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$, which contains the models studied in Opsomer and Ruppert (1998) and Fan and Jiang (2005). This reduces to model selection between models (24) and (25). Based on the sample, the value of ELR statistic is 19.33, greater than the critical vale $\chi^2_{1,0.95} = 3.84$, and the average difference of squared prediction errors between models (24) and (25) is given by $n^{-1} \sum_{i=1}^{n} \hat{\xi}_i = 19.36$. Hence, according to the decision rule below Theorem 2, we choose model (25).

(ii) (Comparison between additive models) Opsomer and Ruppert (1998) analyzed the dataset via a four dimensional additive model:

$$E(y_i - \bar{y} | z_i, \mathbf{X}_i) = m_0(z_i) + m_2(x_{i,2}) + m_3(x_{i,3}) + m_5(x_{i,5}). \tag{26}$$

28

Based on the GLR test for the above model, Fan and Jiang (2005) confirmed to fit the dataset with the following semiparametric model:

$$E(y_i - \bar{y}|z_i, \mathbf{X}_i) = a_0 z_i + m_2(x_{i,2}) + a_3 x_{i,3} + a_5 x_{i,5}. \tag{27}$$

First, we consider model selection between models (27) and (26) using our ELR test. The realized value of ELR statistic is 14.57, which is greater than the critical value $\chi^2_{1,0.95}$, and the average difference of squared prediction errors between models (27) and (26) is given by $n^{-1} \sum_{i=1}^n \hat{\xi}_i = 2.67$. This suggests us to choose (26), which agrees with the GLR test.

Next, we compare model (26) with model (25). The ELR statistic is $30.53 > \chi^2_{1,0.95}$. The average difference of squared prediction errors is given by $n^{-1} \sum_{i=1}^n \hat{\xi}_i = 4.39$. This leads to selection of model (25). That is, at least one of $m_1(\cdot)$, $m_4(\cdot)$ and $m_6(\cdot)$ are not zero.

Then, we test $H_{0\ell} : m_\ell(\cdot) = 0$ against $H_{1\ell} : m_\ell(\cdot) \neq 0$ for each $\ell = 1, 4, 6$ in model (25), using the ELR test. The results are reported in Table 4.   Obviously,

Table 4: ELR testing whether a nonparamatric function is zero

	$m_1(\cdot)$	$m_4(\cdot)$	$m_6(\cdot)$
ELR	8.20	10.92	0.82
Equivalent(=)	$\neq$	$\neq$	$=$

m_1 and m_4 are statistically significant, but $m_6(\cdot)$ not at 5% significance level, based on individual ELR tests or the multiple ELR test with the Bonferroni correction. This leads to the model

$$E(y_i - \bar{y}|z_i, \mathbf{X}_i) = m_0(z_i) + m_1(x_{i,1}) + m_2(x_{i,2}) + m_3(x_{i,3}) + m_4(x_{i,4}) + m_5(x_{i,5}). \tag{28}$$

Last, we compare model (28) with model (25). The ELR statistic is $0.82 < \chi^2_{1,0.95}$. Thus, models (25) and (28) are equivalent. Since model (28) is simpler, it is preferable according to the Occam's razor. This selection agrees to Table 4.

Appendix: Notations and Conditions

For ease of exposure, we introduce some notations which will be used throughout the remainder of the paper. Let $\mathbf{\Sigma}_A = E\{\mathbf{\Pi}(\mathbf{X})\mathbf{\Pi}(\mathbf{X})^\top\}$, $\mathbf{\Sigma}_{A,-l} = E\{\mathbf{\Pi}_{-\ell}(\mathbf{X})\mathbf{\Pi}_{-\ell}(\mathbf{X})^\top\}$, and $\mathbf{\Sigma}_C = E\{\mathbf{\Gamma}(\mathbf{X}, z)\mathbf{\Gamma}(\mathbf{X}, z)^\top\}$. Put $\widetilde{\mathbf{\Pi}}(\mathbf{X}_i) = \mathbf{\Sigma}_A^{-1/2}\mathbf{\Pi}(\mathbf{X}_i)$, $\widetilde{\mathbf{\Pi}}_{-\ell}(\mathbf{X}_i) = \mathbf{\Sigma}_{A,-l}^{-1/2}\mathbf{\Pi}_{-\ell}(\mathbf{X}_i)$, and $\widetilde{\mathbf{\Gamma}}(\mathbf{X}_i, z_i) = \mathbf{\Sigma}_C^{-1/2}\mathbf{\Gamma}(\mathbf{X}_i, z_i)$. For $\alpha = 0, -\ell$, let

$$\tilde{\mathbf{b}}_{-\alpha} = \arg\min_{\mathbf{b}} E\{y - \mathbf{\Pi}_{-\alpha}(\mathbf{X})^\top\mathbf{b}\}^2 \quad \text{and} \quad \tilde{\mathbf{c}} = \arg\min_{\mathbf{c}} E\{y - \mathbf{\Gamma}(\mathbf{X}, z)^\top\mathbf{c}\}^2.$$

Then, by the first order condition, the above minimizers admit closed formulas:

$$\tilde{\mathbf{b}}_{-\alpha} = \{E(\mathbf{\Pi}_{-\alpha}(\mathbf{X})\mathbf{\Pi}_{-\alpha}(\mathbf{X})^\top)\}^{-1}E\{\mathbf{\Pi}_{-\alpha}(\mathbf{X})y\}; \quad \tilde{\mathbf{c}} = \{E(\mathbf{\Gamma}(\mathbf{X}, z)\mathbf{\Gamma}(\mathbf{X}, z)^\top)\}^{-1}E\{\mathbf{\Gamma}(\mathbf{X}, z)y\}.$$

The population versions of prediction errors for models (1), (2) and (13) are $\tilde{\varepsilon}_{1,i} = y_i - \mathbf{\Gamma}(\mathbf{X}_i, z_i)^\top\tilde{\mathbf{c}}$, $\tilde{\varepsilon}_{2,i} = y_i - \mathbf{\Pi}(\mathbf{X}_i)^\top\tilde{\mathbf{b}}_{-0}$, and $\tilde{\varepsilon}_{3,i} = y_i - \mathbf{\Pi}(\mathbf{X}_i)_{-\ell}^\top\tilde{\mathbf{b}}_{-\ell}$, respectively. It is straightforward to verify that

$$\tilde{\varepsilon}_{1,i} = y_i - \widetilde{\mathbf{\Gamma}}(\mathbf{X}_i, z_i)^\top E\{\widetilde{\mathbf{\Gamma}}(\mathbf{X}_1, z_1)y_1\}, \tag{A.1}$$

$$\tilde{\varepsilon}_{2,i} = y_i - \widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top E\{\widetilde{\mathbf{\Pi}}(\mathbf{X}_1)y_1\}, \tag{A.2}$$

$$\tilde{\varepsilon}_{3,i} = y_i - \widetilde{\mathbf{\Pi}}_{-\ell}(\mathbf{X}_i)^\top E\{\widetilde{\mathbf{\Pi}}_{-\ell}(\mathbf{X}_1)y_1\}. \tag{A.3}$$

Then we define $\tilde{\xi}_i = \tilde{\varepsilon}_{1,i}^2 - \tilde{\varepsilon}_{2,i}^2$ and $\tilde{\eta}_i = \tilde{\varepsilon}_{3,i}^2 - \tilde{\varepsilon}_{2,i}^2$, which are population versions of $\hat{\xi}_i$ and $\hat{\eta}_i$, respectively.

To establish our theoretical results, we need some technical conditions. Let $\mathcal{H}_r$ be a space of functions whose dth order derivative is Hölder continuous of order v. That is, $\mathcal{H}_r = \{h(\cdot) : |h^{(d)}(a') - h^{(d)}(a)| \le C|a' - a|^v, a, a' \in [0, 1]\}$, where $h^{(d)}(\cdot)$ is dth derivative and $r = d+v$. If $v = 1$, then $h^{(d)}(\cdot)$ is Lipschitz continuous. Assume the following conditions hold:

A1 (Varying coefficient model) (i) The eigenvalues of matrix $E\{\mathbf{\Gamma}(\mathbf{X}, z)\mathbf{\Gamma}(\mathbf{X}, z)^\top\}$ are bounded away from 0 and ∞; (ii) Assume that $\beta_j \in \mathcal{H}_r$, and $\tilde{\kappa}_j = O(n^{1/(2r+1)})$ for some $r > 1.5$ and $0 \le j \le p$; (iii) Assume there exists some $\gamma > 2$ such that $E\|\mathbf{\Gamma}(\mathbf{X}, z)\|_2^{2\gamma} = O(\tilde{\kappa}^\gamma)$, and $E|y|^{2\gamma} < +\infty$, where $\tilde{\kappa} = \sum_{j=0}^p \tilde{\kappa}_j$.

A2 (Additive model) (i) The eigenvalues of matrix $E\{\mathbf{\Pi}(\mathbf{X})\mathbf{\Pi}(\mathbf{X})^\top\}$ are bounded away from 0 and ∞; (ii) Assume that $m_j(\cdot) \in \mathcal{H}_r$ and $\kappa_j = O(n^{1/(2r+1)})$ for $r > 1.5$ and $1 \le j \le p$; (iii) Assume there exists some $\gamma > 2$ such that $E\|\mathbf{\Pi}(\mathbf{X})\|_2^{2\gamma} = O(\kappa^\gamma)$, and $E|y|^{2\gamma} < +\infty$, where $\kappa = \sum_{j=1}^p \kappa_j$.

A3 (Varying coefficient and additive models) Assume that $E|\tilde{\varepsilon}_{1,1}|^{2\gamma} = O(1)$, $E|\tilde{\varepsilon}_{2,1}|^{2\gamma} = O(1)$, and $\mathrm{Var}(\tilde{\xi}_1) > c_1$ for some constants $c_1 > 0$ and $\gamma > 2$.

A4 (Additive model) Assume that $E|\tilde{\varepsilon}_{2,1}|^{2\gamma} = O(1)$, $E|\tilde{\varepsilon}_{3,1}|^{2\gamma} = O(1)$, and $\mathrm{Var}(\tilde{\eta}_1) > c_1$ for some constants $c_1 > 0$ and $\gamma > 2$.

The above conditions are wild. By Lemma 7 of Tang et al. (2013), condition A1(i) holds. Condition A2(i) is the same as condition A.2 of Belloni et al. (2015). Conditions A1(ii) and A2(ii) were assumed in Theorem 1 of Tang et al. (2013). For B-spline series, Newey (1997) assumed $\sup_{x_j}\|\mathbf{\Pi}_j(x_j)\|_2 = O(\sqrt{\kappa_j})$, which implies our condition $E\|\mathbf{\Pi}(\mathbf{X})\|_2^{2\gamma} = O(\kappa^\gamma)$ in A2(iii). Notice that $\mathbf{\Gamma}_j(z) = (B_{j,1}(z), \ldots, B_{j,\tilde{\kappa}_j}(z))^\top$ and $\mathbf{\Gamma}(\mathbf{X}, z) = (\mathbf{\Gamma}_0^\top(z), x_1\mathbf{\Gamma}_1^\top(z), \ldots, x_p\mathbf{\Gamma}_p^\top(z))^\top$. If $E(|x_j|^{2\gamma} \mid z_j = z)$ is a bounded function of z and $E\|\mathbf{\Gamma}_j(z)\|_2^{2\gamma} = O(\tilde{\kappa}^\gamma)$, then the condition $E\|\mathbf{\Gamma}(\mathbf{X}, z)\|_2^{2\gamma} = O(\tilde{\kappa}^\gamma)$ in A1(iii) holds.

Since the squared prediction errors in the ELR test are compared, conditions A3 and A4 assume that the γth moments of their population versions must be bounded away from $+\infty$, i.e. $E(|\tilde{\varepsilon}_{j,1}|^{2\gamma}) = O(1)$ for $j = 1, 2, 3$. This can be relaxed if one compares the median of prediction errors, but it will complicate the technical proofs of theorems. Furthermore, it is assumed in condition A3 that $\mathrm{Var}(\tilde{\xi}_1) > c_1$. This condition, combined with Lemma 3(i), ensures that $\sigma_\xi > 0$. Otherwise, there is no need to develop a test for comparson of the two competing models. Similarly, in condition A4 it is reasonable to assume $\mathrm{Var}(\tilde{\eta}_1) > c_1$.

Supplementary Material

To save space, all technical proofs of theorems are included in the online supplementary material.

Supplementary material for "Nonnested model selection based on empirical likelihood"

Now we give technical proofs of our theorems. To streamline our arguments, we first introduce some technical lemmas whose proofs are reported after the proofs of theorems.

Lemma 1. Assume conditions A1 - A3 hold. Then, for $j = 1, 2$,

(i) $\max_{1 \le i \le n} |\tilde{\varepsilon}_{j,i} - \hat{\varepsilon}_{j,i}| = O_P(n^{\frac{1}{2\gamma} + \frac{3}{4r+2} - \frac{1}{2}})$; (ii) $n^{-1} \sum_{i=1}^{n} (\hat{\varepsilon}_{j,i} - \tilde{\varepsilon}_{j,i})^2 = O_P(n^{\frac{2}{2r+1} - 1})$.

Lemma 2. Assume conditions A1 - A3 hold. Then

(i) $n^{-1} \sum_{i=1}^{n} (\hat{\xi}_i - \tilde{\xi}_i) = o_P(n^{-1/2})$; (ii) $n^{-1} \sum_{i=1}^{n} (\hat{\xi}_i^2 - \tilde{\xi}_i^2) = o_P(1)$.

Lemma 3. Assume conditions A1 - A3 hold. Then

(i) $\mathrm{Var}(\tilde{\xi}_1) = \sigma_\xi^2 + o(1)$;

(ii) $\max_{1 \le i \le n} |\tilde{\xi}_i| = O_P(n^{1/\gamma})$, $\max_{1 \le i \le n} |\hat{\xi}_i| = O_P(n^{1/\gamma})$, $E|\tilde{\xi}_1|^\gamma = O(1)$, and $n^{-1} \sum_{i=1}^{n} \tilde{\xi}_i^2 = E\tilde{\xi}_1^2 + o_P(1)$;

(iii) under $H_{a,n}^{(1)}$, $a_n^{-1} \sqrt{n} E\tilde{\xi}_1 / \sigma_\xi \to 1$ when $|a| = +\infty$, and $\sqrt{n} E\tilde{\xi}_1 / \sigma_\xi \to a$ when $|a| < \infty$;

(iv) under $H_{a,n}^{(1)}$, if $|a| < +\infty$, then $n^{-1} \sum_{i=1}^{n} \tilde{\xi}_i = O_P(n^{-1/2})$ and $n^{-1} \sum_{i=1}^{n} \hat{\xi}_i = O_P(n^{-1/2})$.

Lemma 4. Assume coditions A1 - A3 hold. Under $H_{a,n}^{(1)}$, if $|a| < \infty$, then $\hat{\lambda} = O_P(n^{-1/2})$.

Proofs of Theorems 1-2. Since Theorem 1 can be proven along the same line as Theorem 2 (with $a_n = 0$), we omit the proof of Theorem 1.

Case (i): $|a| < \infty$. Notice that $\hat{\lambda}$ solves the equation $\sum_{i=1}^{n} \hat{\xi}_i / (1 + \hat{\lambda}\hat{\xi}_i) = 0$, which can be rewritten as

$$0 = \sum_{i=1}^{n} \hat{\xi}_i \{1 - \hat{\lambda}\hat{\xi}_i + (\hat{\lambda}\hat{\xi}_i)^2 / (1 + \hat{\lambda}\hat{\xi}_i)\}.$$

Then

$$\hat{\lambda} = \left(\sum_{i=1}^{n}\hat{\xi}_i^2\right)^{-1}\sum_{i=1}^{n}\hat{\xi}_i\{1+(\hat{\lambda}\hat{\xi}_i)^2/(1+\hat{\lambda}\hat{\xi}_i)\}$$

$$= \left(\sum_{i=1}^{n}\hat{\xi}_i^2\right)^{-1}\sum_{i=1}^{n}\hat{\xi}_i + \left(\sum_{i=1}^{n}\hat{\xi}_i^2\right)^{-1}\sum_{i=1}^{n}\hat{\lambda}^2\hat{\xi}_i^3/(1+\hat{\lambda}\hat{\xi}_i). \tag{A.4}$$

Applying Taylor's expansion to $\sum_{i=1}^{n}\log(1+\hat{\lambda}\hat{\xi}_i)$ leads to

$$\sum_{i=1}^{n}\log(1+\hat{\lambda}\hat{\xi}_i) = \sum_{i=1}^{n}\left\{\hat{\lambda}\hat{\xi}_i - (\hat{\lambda}\hat{\xi}_i)^2/2 + \frac{(\hat{\lambda}\hat{\xi}_i)^3}{3(1+c_i\hat{\lambda}\hat{\xi}_i)^3}\right\}$$

$$= \hat{\lambda}\sum_{i=1}^{n}\hat{\xi}_i - \hat{\lambda}\left(\sum_{i=1}^{n}\hat{\xi}_i^2\right)\hat{\lambda}/2 + \hat{\lambda}^3\sum_{i=1}^{n}\frac{\hat{\xi}_i^3}{3(1+c_i\hat{\lambda}\hat{\xi}_i)^3}, \tag{A.5}$$

where $c_i \in [0,1]$. By Lemma 2 and Lemma 3(ii), we obtain that

$$n^{-1}\sum_{i=1}^{n}\hat{\xi}_i^2 = n^{-1}\sum_{i=1}^{n}\tilde{\xi}_i^2 + n^{-1}\sum_{i=1}^{n}(\hat{\xi}_i^2 - \tilde{\xi}_i^2)$$

$$= E|\tilde{\xi}_1|^2 + o_P(1), \tag{A.6}$$

which, combined with $\mathrm{Var}(\tilde{\xi}_1) = E|\tilde{\xi}_1|^2 - (E\tilde{\xi}_1)^2$ and $(E\tilde{\xi}_1)^2 = O(n^{-1})$ in Lemma 3(iii), yields that

$$n^{-1}\sum_{i=1}^{n}\hat{\xi}_i^2 = \mathrm{Var}(\tilde{\xi}_1) + o_P(1). \tag{A.7}$$

This, combined with Lemmas 3 and 4, implies that

$$\left|\hat{\lambda}^3\sum_{i=1}^{n}\frac{\hat{\xi}_i^3}{3(1+c_i\hat{\lambda}\hat{\xi}_i)^3}\right| \le O_P(1)|\hat{\lambda}|^3\max_{1\le i\le n}|\hat{\xi}_i|\sum_{i=1}^{n}\hat{\xi}_i^2 = O_P(n^{-3/2}n^{1/\gamma}n) = o_P(1); \tag{A.8}$$

$$\left|\left(\sum_{i=1}^{n}\hat{\xi}_i^2\right)^{-1}\sum_{i=1}^{n}\hat{\lambda}^2\hat{\xi}_i^3/(1+\hat{\lambda}\hat{\xi}_i)\right| \le O_P(1)\hat{\lambda}^2\max_{1\le i\le n}|\hat{\xi}_i| = O_P(n^{1/\gamma-1}).$$

Let $\bar{\xi}_n = n^{-1}\sum_{i=1}^{n}\hat{\xi}_i$. Then, by Lemmas 3-4, we have $|\bar{\xi}_n| = O_P(n^{-1/2})$ and $|\hat{\lambda}| = O_P(n^{-1/2})$. Then, it follows from (A.4) and (A.7) that

$$\hat{\lambda} = \bar{\xi}_n/\mathrm{Var}(\tilde{\xi}_1) + o_P(n^{-1/2}).$$

By (A.5), (A.7) and (A.8), we have

$$2\sum_{i=1}^{n}\log(1+\hat{\lambda}\hat{\xi}_i) = 2n\hat{\lambda}\bar{\xi}_n - n\mathrm{Var}(\tilde{\xi}_1)\hat{\lambda}^2 + o_P(1).$$

2

Hence,

$$R_{n,1} = 2\sum_{i=1}^{n} \log(1 + \hat{\lambda}\hat{\xi}_i) = n\bar{\xi}_n^2/\text{Var}(\tilde{\xi}_1) + o_P(1).$$

Denoted by $\bar{\xi}_n^* = n^{-1}\sum_{i=1}^{n}\tilde{\xi}_i$. Applying Lemma 2, we obtain that

$$R_{n,1} = 2\sum_{i=1}^{n} \log(1 + \hat{\lambda}\hat{\xi}_i) = n|\bar{\xi}_n^*|^2/\text{Var}(\tilde{\xi}_1) + o_P(1). \tag{A.9}$$

Since $\{\tilde{\xi}_i\}_{i=1}^{n}$ are iid and $E|\tilde{\xi}_n|^\gamma = O(1)$ for $\gamma > 2$ in Lemma 3(ii), by the Lindeberg-Feller central limit theorem, we establish that

$$\sqrt{n}(\bar{\xi}_n^* - E\tilde{\xi}_1)/\text{Var}(\tilde{\xi}_1) = n^{-1/2}\sum_{i=1}^{n}(\tilde{\xi}_i - E\tilde{\xi}_1)/\text{Var}(\tilde{\xi}_1) \to \mathcal{N}(0,1).$$

Under $H_{a,n}^{(1)}$, we know from Lemma 3(iii) that $\sqrt{n}E\tilde{\xi}_1/\sqrt{\text{Var}(\tilde{\xi}_1)} \to a$. Therefore,

$$\sqrt{n}\bar{\xi}_n^*/\sqrt{\text{Var}(\tilde{\xi}_1)} \to \mathcal{N}(a,1).$$

Then, by (A.9),

$$R_{n,1} = 2\sum_{i=1}^{n} \log(1 + \hat{\lambda}\hat{\xi}_i) \to \chi_1^2(a^2). \tag{A.10}$$

Case (ii): $a = \infty$. Let $\hat{\lambda}_* = n^{-1/2}\text{sgn}(E\tilde{\xi}_i)$. By Lemma 3(ii), $\max_{1\le i\le n}\hat{\xi}_i = o_P(n^{1/\gamma})$ for $\gamma > 2$. Then

$$\max_{1\le i\le n} \hat{\lambda}_*\hat{\xi}_i = o_P(1). \tag{A.11}$$

Since $\hat{\lambda} = \arg\max_\lambda \sum_{i=1}^{n} \log(1 + \lambda\hat{\xi}_i)$, we have

$$R_{n,1} = 2\sum_{i=1}^{n} \log(1 + \hat{\lambda}\hat{\xi}_i) \ge 2\sum_{i=1}^{n} \log(1 + \hat{\lambda}_*\hat{\xi}_i).$$

Then, using (A.11) and Taylor's expansion, we establish that

$$\begin{aligned}
R_{n,1} &\ge 2\sum_{i=1}^{n} \hat{\lambda}_*\hat{\xi}_i - \sum_{i=1}^{n} \hat{\lambda}_*^2\hat{\xi}_i^2/(1 + c_i\hat{\lambda}_*\hat{\xi}_i)^2 \\
&\ge 2\sum_{i=1}^{n} \hat{\lambda}_*\hat{\xi}_i - 2\sum_{i=1}^{n} \hat{\lambda}_*^2\hat{\xi}_i^2\{1 + o_P(1)\} \\
&= 2n^{-1/2}\sum_{i=1}^{n} \hat{\xi}_i\text{sgn}(E\tilde{\xi}_i) - 2n^{-1}\sum_{i=1}^{n} \hat{\xi}_i^2\{1 + o_P(1)\}, \tag{A.12}
\end{aligned}$$

3

where $c_i \in [0, 1]$. By Lemma 3(iii), we have $E\tilde{\xi}_1^2 = O(1)$. This, combined with $a_n \to \infty$ and (A.6), produces that $n^{-1}\sum_{i=1}^n \hat{\xi}_i^2 = E\tilde{\xi}_1^2 + o_P(1) = o_P(a_n)$. Hence,

$$R_{n,1} \geq 2n^{-1/2}\sum_{i=1}^n \hat{\xi}_i \mathrm{sgn}(E\tilde{\xi}_i) + o_P(a_n).$$

Using Lemma 2, we get

$$n^{-1}\sum_{i=1}^n \hat{\xi}_i \mathrm{sgn}(E\tilde{\xi}_i) = n^{-1}\sum_{i=1}^n \mathrm{sgn}(E\tilde{\xi}_i)\tilde{\xi}_i + o_P(a_n n^{-1/2}) = |E\tilde{\xi}_i| + o_P(a_n n^{-1/2}).$$

Then

$$R_{n,1} \geq 2\sqrt{n}|E\tilde{\xi}_i| + o_P(a_n). \tag{A.13}$$

By Lemma 3(i)-(iii) and condition A3, we know that, for large n, $|E\tilde{\xi}_i| > 0.5\sqrt{c_1}|a_n|n^{-1/2}$. This, together with (A.13) and $a_n \to \infty$, leads to

$$P(R_{n,1} \to \infty) \to 1. \tag{A.14}$$

Proofs of Theorems 3-4. Since Theorem 3 can be shown in the same way as Theorem 4, we only prove Theorem 4. The asymptotic results for $R_{n,2}$ follow along the same line as that for Theorem 2 by replacing $\hat{\xi}_i$ and $\tilde{\xi}_i$ with $\hat{\eta}_i$ and $\tilde{\eta}_i$, respectively. Since $R_{n,3} = R_{n,2}$, we complete the proof of Theorem 4. ◇

Proof Lemma 1. (i) We first show that

$$\max_{1 \leq i \leq n} \|\mathbf{\Sigma}_n^{(-i)} - \mathbf{\Sigma}_A\|_2 = O_P(\kappa n^{-1/2}). \tag{A.15}$$

Put $\mathbf{\Sigma}_n = (n-1)^{-1}\sum_{i=1}^n \mathbf{\Pi}(\mathbf{X}_i)\mathbf{\Pi}(\mathbf{X}_i)^\top$. By condition A2(iii) and the inequality $(E|b|^2)^{1/2} \leq (E|b|^\gamma)^{1/\gamma}$ with $b = \|\mathbf{\Pi}(\mathbf{X})\|_2^2$ for $\gamma > 2$, we have $E\|\mathbf{\Pi}(\mathbf{X})\|_2^4 \leq \{E\|\mathbf{\Pi}(\mathbf{X})\|_2^{2\gamma}\}^{2/\gamma} = O(\kappa^2)$. Then

$$
\begin{aligned}
E\|\mathbf{\Sigma}_n - \mathbf{\Sigma}_A\|_2^2 &\leq \mathrm{Trace}\{E(\mathbf{\Sigma}_n - \mathbf{\Sigma}_A)(\mathbf{\Sigma}_n - \mathbf{\Sigma}_A)\} \\
&= (n-1)^{-2}\sum_{k=1}^\kappa \sum_{l=1}^\kappa \sum_{i=1}^n [E\{\Pi_k^2(\mathbf{X}_i)\Pi_l^2(\mathbf{X}_i)\} - \Sigma_{A,kl}^2] \\
&= n(n-1)^{-2}\{E\|\mathbf{\Pi}(\mathbf{X})\|_2^4 - \mathrm{Trace}(\mathbf{\Sigma}_A^2)\} \\
&= O(\kappa^2/n), \tag{A.16}
\end{aligned}
$$

4

where $\Sigma_{A,kl} = E\{\Pi_k^2(\mathbf{X}_i)\Pi_l^2(\mathbf{X}_i)\}$. Hence,

$$\|\mathbf{\Sigma}_n - \mathbf{\Sigma}_A\|_2 = O_P(\kappa/\sqrt{n}) \quad \text{and} \quad \|\mathbf{\Sigma}_n\|_2 \le \|\mathbf{\Sigma}_A\|_2 + O_P(\kappa/\sqrt{n}). \tag{A.17}$$

Furthermore,

$$\|\mathbf{\Sigma}_n^{(-i)} - \mathbf{\Sigma}_A\|_2 \le \|\mathbf{\Sigma}_n^{(-i)} - \mathbf{\Sigma}_n\|_2 + \|\mathbf{\Sigma}_n - \mathbf{\Sigma}_A\|_2 = (n-1)^{-1}\|\mathbf{\Pi}(\mathbf{X}_i)\|_2^2 + O_P(\kappa/\sqrt{n}).$$

Note that, by condition A2, $E\|\mathbf{\Pi}(\mathbf{X})\|_2^{2\gamma} = O(\kappa^\gamma)$. It follows from Markov's inequality that

$$\max_{1\le i\le n} \|\mathbf{\Pi}(\mathbf{X}_i)\|_2^{2\gamma} \le n n^{-1}\sum_{i=1}^n \|\mathbf{\Pi}(\mathbf{X}_i)\|_2^{2\gamma} = O_P(n\kappa^\gamma). \tag{A.18}$$

That is, $\max_{1\le i\le n}\|\mathbf{\Pi}(\mathbf{X}_i)\|_2^2 = O_P(n^{1/\gamma}\kappa)$. Then

$$\max_{1\le i\le n}\|\mathbf{\Sigma}_n^{(-i)} - \mathbf{\Sigma}_A\|_2 = O_P\{(n^{1/\gamma-1} + n^{-1/2})\kappa\} = O_P(\kappa n^{-1/2}).$$

That is, (A.15) holds. Note that $E\mathrm{y}_i\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top E\{\widetilde{\mathbf{\Pi}}(\mathbf{X}_1)\mathrm{y}_1\} = \|\mathbf{\Sigma}_A^{-1/2}E\{\mathbf{\Pi}(\mathbf{X}_1)\mathrm{y}_1\}\|_2^2$. It follows from (A.2) that

$$E|\tilde\varepsilon_{2,i}|^2 = E|\mathrm{y}_i - \widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top E\{\widetilde{\mathbf{\Pi}}(\mathbf{X}_1)\mathrm{y}_1\}|^2 = E|\mathrm{y}_1|^2 - \|\mathbf{\Sigma}_A^{-1/2}E\{\mathbf{\Pi}(\mathbf{X}_1)\mathrm{y}_1\}\|_2^2.$$

This, together with condition A2, implies that $\|\mathbf{\Sigma}_A^{-1/2}E\{\mathbf{\Pi}(\mathbf{X}_1)\mathrm{y}_1\}\|_2^2 \le E|\mathrm{y}_1|^2 = O(1)$. Then, with $\lambda_{\max}(\mathbf{\Sigma}_A) = O(1)$ and $\lambda_{\min}(\mathbf{\Sigma}_A) > 0$ in condition A2, it is easy to see that

$$\|E\{\mathbf{\Pi}(\mathbf{X}_1)\mathrm{y}_1\}\|_2 = O(1) \quad \text{and} \quad \|\mathbf{\Sigma}_A^{-1}E\{\mathbf{\Pi}(\mathbf{X}_1)\mathrm{y}_1\}\|_2 = O(1). \tag{A.19}$$

Denoted by $\boldsymbol{\mu}_k = \mathbf{\Pi}(\mathbf{X}_k)\mathrm{y}_k - E\{\mathbf{\Pi}(\mathbf{X}_k)\mathrm{y}_k\}$. Applying Cauchy-Schwarz's inequality, (A.15), (A.18) and (A.19), we obtain that

$$
\begin{aligned}
|\tilde\varepsilon_{2,i} - \hat\varepsilon_{2,i}| &\le \|\mathbf{\Pi}(\mathbf{X}_i)\|_2 \Big\| \frac{1}{n-1}\sum_{k=1(\ne i)}^n (\mathbf{\Sigma}_n^{(-i)})^{-1}\boldsymbol{\mu}_k + \{(\mathbf{\Sigma}_n^{(-i)})^{-1} - \mathbf{\Sigma}_A^{-1}\}E\{\mathbf{\Pi}(\mathbf{X}_k)\mathrm{y}_k\}\Big\|_2 \\
&\le O_P(n^{\frac{1}{2\gamma}}\sqrt{\kappa})\Big\|\frac{1}{n-1}\sum_{k=1(\ne i)}^n \boldsymbol{\mu}_k\Big\|_2 + O_P(n^{\frac{1}{2\gamma}}\sqrt{\kappa})\|(\mathbf{\Sigma}_n^{(-i)})^{-1}(\mathbf{\Sigma}_n^{(-i)} - \mathbf{\Sigma}_A)\mathbf{\Sigma}_A^{-1}\|_2 \\
&= O_P(n^{\frac{1}{2\gamma}}\sqrt{\kappa})\Big\|\frac{1}{n-1}\sum_{k=1(\ne i)}^n \boldsymbol{\mu}_k\Big\|_2 + O_P(\kappa^{3/2}n^{\frac{1}{2\gamma}-\frac{1}{2}}),
\end{aligned}
$$

5

uniformly for $1 \le i \le n$. Note that $E\|\mathbf{\Pi}(\mathbf{X}_1)\|_2^{2\gamma} = O(\kappa^\gamma)$ and $E|\mathrm{y}_1|^{2\gamma} = O(1)$ (condition A2), it follows that

$$E\|\mathbf{\Pi}(\mathbf{X}_1)\mathrm{y}_1\|_2^2 \le \sqrt{E\|\mathbf{\Pi}(\mathbf{X}_1)\|_2^4 E|\mathrm{y}_1|^4} = O(\kappa). \tag{A.20}$$

Then

$$E\big\|\sum_{k=1}^n \boldsymbol{\mu}_k\big\|_2^2 = \sum_{k=1}^n E\|\boldsymbol{\mu}_k\|_2^2 = O(n\kappa). \tag{A.21}$$

Recalling that $E|\mathrm{y}_1|^{2\gamma} = O(1)$, we have $\max_{1\le i\le n}|\mathrm{y}_i|^{2\gamma} = O_P(n)$. This, together with (A.18), yields that

$$\max_{1\le i\le n}\|\mathbf{\Pi}(\mathbf{X}_i)\mathrm{y}_i\|_2 \le \max_{1\le i\le n}\|\mathbf{\Pi}(\mathbf{X}_i)\|_2 \max_{1\le i\le n}|\mathrm{y}_i| = O_P(n^{1/\gamma}\sqrt{\kappa}). \tag{A.22}$$

Then, applying (A.20)-(A.22) and the inequality

$$\max_{1\le i\le n}\big\|\sum_{k=1(\ne i)}^n \boldsymbol{\mu}_k\big\|_2 \le \big\|\sum_{k=1}^n \boldsymbol{\mu}_k\big\|_2 + \max_{1\le i\le n}\|\mathbf{\Pi}(\mathbf{X}_i)\mathrm{y}_i\|_2 + E\|\mathbf{\Pi}(\mathbf{X}_1)\mathrm{y}_1\|_2,$$

we obtain that

$$\max_{1\le i\le n}\big\|\frac{1}{n-1}\sum_{k=1(\ne i)}^n \boldsymbol{\mu}_k\big\|_2 = O_P(\sqrt{\kappa/n} + n^{1/\gamma-1}\sqrt{\kappa}). \tag{A.23}$$

Thus, $\max_{1\le i\le n}|\tilde{\varepsilon}_{2,i} - \hat{\varepsilon}_{2,i}| = O_P(\kappa^{3/2}n^{\frac{1}{2\gamma}-\frac{1}{2}}) = O_P(n^{\frac{1}{2\gamma}+\frac{3}{4r+2}-\frac{1}{2}})$. Similarly, we can also show that $\max_{1\le i\le n}|\tilde{\varepsilon}_{1,i} - \hat{\varepsilon}_{1,i}| = O_P(n^{\frac{1}{2\gamma}+\frac{3}{4r+2}-\frac{1}{2}})$.

(ii) Let $\hat{\varepsilon}_{2,i}^* = \frac{1}{n-1}\sum_{k=1}^n \mathbf{\Pi}(\mathbf{X}_i)^\top \boldsymbol{\Sigma}_n^{-1}\mathbf{\Pi}(\mathbf{X}_k)\mathrm{y}_k$. Then

$$n^{-1}\sum_{i=1}^n |\hat{\varepsilon}_{2,i} - \tilde{\varepsilon}_{2,i}|^2 \le 2n^{-1}\sum_{i=1}^n |\hat{\varepsilon}_{2,i}^* - \tilde{\varepsilon}_{2,i}|^2 + 2n^{-1}\sum_{i=1}^n |\hat{\varepsilon}_{2,i} - \hat{\varepsilon}_{2,i}^*|^2.$$

By the inequality $(a+b)^2 \le 2(a^2+b^2)$, we have

$$\begin{aligned}
n^{-1}\sum_{i=1}^n |\hat{\varepsilon}_{2,i} - \hat{\varepsilon}_{2,i}^*|^2 &\le 2n^{-1}\sum_{i=1}^n \big|\frac{1}{n-1}\sum_{k=1(\ne i)}^n \mathbf{\Pi}(\mathbf{X}_i)^\top \big[\{\boldsymbol{\Sigma}_n^{(-i)}\}^{-1} - \boldsymbol{\Sigma}_n^{-1}\big]\mathbf{\Pi}(\mathbf{X}_k)\mathrm{y}_k\big|^2 \\
&\quad + \frac{2}{n(n-1)^2}\sum_{i=1}^n \big|\mathbf{\Pi}(\mathbf{X}_i)^\top \boldsymbol{\Sigma}_n^{-1}\mathbf{\Pi}(\mathbf{X}_i)\mathrm{y}_i\big|^2 \\
&\equiv r_{n,1} + r_{n,2}.
\end{aligned}$$

6

Applying (A.19) and (A.23), we obtain that

$$\max_{1 \le i \le n} \Big\| \frac{1}{n-1} \sum_{k=1(\ne i)}^{n} \boldsymbol{\mu}_k + E \boldsymbol{\Pi}(\mathbf{X}_1) \mathbf{y}_1 \Big\|_2 = O_P(1).$$

This, combined with the Cauchy-Schwarz inequality, (A.15) and (A.17), yields that

$$
\begin{aligned}
r_{n,1} \;\le\;& 2n^{-1} \sum_{i=1}^{n} \| \boldsymbol{\Pi}(\mathbf{X}_i) \|_2^2 \cdot \| \{ \boldsymbol{\Sigma}_n^{(-i)} \}^{-1} - \boldsymbol{\Sigma}_n^{-1} \|_2^2 \cdot \Big\| \frac{1}{n-1} \sum_{k=1(\ne i)}^{n} \boldsymbol{\mu}_k + E \boldsymbol{\Pi}(\mathbf{X}_1)\mathbf{y}_1 \Big\|_2^2 \\
\le\;& O_P(1) \frac{1}{(n-1)^2 n} \sum_{i=1}^{n} \| \boldsymbol{\Pi}(\mathbf{X}_i) \|_2^2 \cdot \| \boldsymbol{\Pi}(\mathbf{X}_i) \|_2^4 \cdot \| \boldsymbol{\Sigma}_n^{-1} \|_2^2 \cdot \| \{ \boldsymbol{\Sigma}_n^{(-i)} \}^{-1} \|_2^2 \\
=\;& O_P(n^{-2}) n^{-1} \sum_{i=1}^{n} \| \boldsymbol{\Pi}(\mathbf{X}_i) \|_2^6
\end{aligned}
$$

and

$$
r_{n,2} \;\le\; \frac{2}{n(n-1)^2} \sum_{i=1}^{n} \| \boldsymbol{\Pi}(\mathbf{X}_i) \|_2^2 \cdot \| \boldsymbol{\Sigma}_n^{-1} \|_2^2 \cdot \| \boldsymbol{\Pi}(\mathbf{X}_i)\mathbf{y}_i \|_2^2 = O_P(n^{-2}) n^{-1} \sum_{i=1}^{n} \| \boldsymbol{\Pi}(\mathbf{X}_i) \|_2^4 \mathbf{y}_i^2.
$$

By condition A2, we have $\max_{1 \le i \le n} |\mathbf{y}_i| = O_P(n^{1/(2\gamma)})$ and

$$E \| \boldsymbol{\Pi}(\mathbf{X}_1) \|_2^4 \le \{ E \| \boldsymbol{\Pi}(\mathbf{X}_1) \|_2^{2\gamma} \}^{2/\gamma} = O(\kappa^2).$$

Then, applying (A.18), we establish that

$$n^{-1} \sum_{i=1}^{n} \| \boldsymbol{\Pi}(\mathbf{X}_i) \|_2^6 \le \max_{1 \le i \le n} \| \boldsymbol{\Pi}(\mathbf{X}_i) \|_2^2 n^{-1} \sum_{i=1}^{n} \| \boldsymbol{\Pi}(\mathbf{X}_i) \|_2^4 = O_P(\kappa^3 n^{1/\gamma})$$

and

$$n^{-1} \sum_{i=1}^{n} \| \boldsymbol{\Pi}(\mathbf{X}_i) \|_2^4 \mathbf{y}_i^2 \le \max_{1 \le i \le n} \mathbf{y}_i^2 n^{-1} \sum_{i=1}^{n} \| \boldsymbol{\Pi}(\mathbf{X}_i) \|_2^4 = O_P(\kappa^2 n^{1/\gamma}).$$

Thus, $n^{-1} \sum_{i=1}^{n} |\hat{\varepsilon}_{2,i} - \hat{\varepsilon}_{2,i}^*|^2 = O_P(\kappa^3 n^{1/\gamma - 2})$. It follows that

$$n^{-1} \sum_{i=1}^{n} |\hat{\varepsilon}_{2,i} - \tilde{\varepsilon}_{2,i}|^2 \le 2n^{-1} \sum_{i=1}^{n} |\hat{\varepsilon}_{2,i}^* - \tilde{\varepsilon}_{2,i}|^2 + O_P(\kappa^3 n^{1/\gamma - 2}). \tag{A.24}$$

Put $\mathbf{v} = \boldsymbol{\Sigma}_A^{-1} E \{ \boldsymbol{\Pi}(\mathbf{X}_k)\mathbf{y}_k \}$. Then

$$
\begin{aligned}
\hat{\varepsilon}_{2,i}^* - \tilde{\varepsilon}_{2,i} \;=\;& \frac{1}{n-1} \sum_{k=1}^{n} \Big[\boldsymbol{\Pi}(\mathbf{X}_i)^\top \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\mu}_k + \boldsymbol{\Pi}(\mathbf{X}_i)^\top \{ \boldsymbol{\Sigma}_n^{-1} - \boldsymbol{\Sigma}_A^{-1} \} E \{ \boldsymbol{\Pi}(\mathbf{X}_k)\mathbf{y}_k \} \Big] \\
=\;& \frac{1}{n-1} \sum_{k=1}^{n} \boldsymbol{\Pi}(\mathbf{X}_i)^\top \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\mu}_k + \frac{1}{n-1} \sum_{k=1}^{n} \boldsymbol{\Pi}(\mathbf{X}_i)^\top \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_n) \mathbf{v} \\
\equiv\;& I_{i,1} + I_{i,2}. \tag{A.25}
\end{aligned}
$$

By the Cauchy-Schwarz inequality, (A.17), (A.19), (A.21), $\lambda_{\max}(\boldsymbol{\Sigma}_A) = O(1)$, and $\lambda_{\min}(\boldsymbol{\Sigma}_A) > 0$ in condition A2, it holds that

$$n^{-1} \sum_{i=1}^{n} I_{i,2}^2 \leq \frac{(n-1)}{n} \|\boldsymbol{\Sigma}_n\|_2 \cdot \|\boldsymbol{\Sigma}_n^{-1}\|_2^2 \cdot \|(\boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_n)\|_2^2 \cdot \|\mathbf{v}\|_2^2 = O_P(\kappa^2/n); \qquad \text{(A.26)}$$

$$n^{-1} \sum_{i=1}^{n} I_{i,1}^2 \leq \frac{(n-1)}{n} \|\boldsymbol{\Sigma}_n\|_2 \cdot \|\boldsymbol{\Sigma}_n^{-1}\|_2^2 \cdot \frac{1}{(n-1)^2} \|\sum_{k=1}^{n} \boldsymbol{\mu}_k\|_2^2 = O_P(\kappa/n). \qquad \text{(A.27)}$$

Naturally, combining (A.25)-(A.27) leads to

$$n^{-1} \sum_{i=1}^{n} |\hat{\varepsilon}_{2,i}^* - \tilde{\varepsilon}_{2,i}|^2 \leq 2n^{-1} \sum_{i=1}^{n} I_{i,1}^2 + 2n^{-1} \sum_{i=1}^{n} I_{i,2}^2 = O_P(\kappa^2/n),$$

which, together with (A.24), yields that $n^{-1} \sum_{i=1}^{n} |\hat{\varepsilon}_{2,i} - \tilde{\varepsilon}_{2,i}|^2 = O_P(n^{2/(2r+1)-1})$. Along the same line, we can also show that $n^{-1} \sum_{i=1}^{n} |\hat{\varepsilon}_{1,i} - \tilde{\varepsilon}_{1,i}|^2 = O_P(n^{2/(2r+1)-1})$. $\diamond$

Proof of Lemma 2. (i) By the definitions of $\hat{\xi}_i$ and $\tilde{\xi}_i$, we have

$$n^{-1} \sum_{i=1}^{n} (\hat{\xi}_i - \tilde{\xi}_i) = n^{-1} \sum_{i=1}^{n} (\hat{\varepsilon}_{1,i}^2 - \tilde{\varepsilon}_{1,i}^2) - n^{-1} \sum_{i=1}^{n} (\hat{\varepsilon}_{2,i}^2 - \tilde{\varepsilon}_{2,i}^2).$$

We will show each term on the righthand side of the above equation is $o_P(n^{-1/2})$. In the following we only show this for the 2nd term, since it can be done similarly for the 1st term.

Notice that

$$|\hat{\varepsilon}_{2,i}|^2 - |\tilde{\varepsilon}_{2,i}|^2 = (\hat{\varepsilon}_{2,i} - \tilde{\varepsilon}_{2,i})^2 + 2\tilde{\varepsilon}_{2,i}(\hat{\varepsilon}_{2,i} - \tilde{\varepsilon}_{2,i}).$$

By Lemma 1, $n^{-1} \sum_{i=1}^{n} |\hat{\varepsilon}_{2,i} - \tilde{\varepsilon}_{2,i}|^2 = o_P(n^{-1/2})$. Then it suffices to show that

$$n^{-1} \sum_{i=1}^{n} \tilde{\varepsilon}_{2,i}(\tilde{\varepsilon}_{2,i} - \hat{\varepsilon}_{2,i}) = o_P(n^{-1/2}). \qquad \text{(A.28)}$$

The reader who does not wish to study the lengthy proof may skip to the proof of (ii).

Let $\boldsymbol{\delta}_i = \widetilde{\boldsymbol{\Pi}}(\mathbf{X}_i)[\mathbf{y}_i - \widetilde{\boldsymbol{\Pi}}(\mathbf{X}_i)^\top E\{\widetilde{\boldsymbol{\Pi}}(\mathbf{X}_1)\mathbf{y}_1\}] = \tilde{\varepsilon}_{2,i}\widetilde{\boldsymbol{\Pi}}(\mathbf{X}_i)$. Then $E\boldsymbol{\delta}_i^\top \boldsymbol{\delta}_j = 0$ for $i \neq j$, and

$$E\|\boldsymbol{\delta}_i\|_2^2 \leq 2E\|\widetilde{\boldsymbol{\Pi}}(\mathbf{X}_i)\mathbf{y}_i\|_2^2 + 2\|\boldsymbol{\Sigma}_A^{-1}E\{\boldsymbol{\Pi}(\mathbf{X}_1)\mathbf{y}_1\}\|_2^2 \cdot E\{\|\boldsymbol{\Pi}(\mathbf{X}_i)\|_2^2 \cdot \|\widetilde{\boldsymbol{\Pi}}(\mathbf{X}_i)\|_2^2\}. \qquad \text{(A.29)}$$

8

Note that $E\|\mathbf{\Pi}(\mathbf{X}_i)\|_2^4 \leq \{E\|\mathbf{\Pi}(\mathbf{X}_i)\|_2^{2\gamma}\}^{2/\gamma} = O(\kappa^2)$ and $\lambda_{\min}(\mathbf{\Sigma}_A) > 0$. It follows from (A.19), (A.20) and (A.29) that $E\|\boldsymbol{\delta}_i\|_2^2 = O(\kappa^2)$. Hence,

$$\Big\|n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_{2,i}\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)\Big\|_2 = O_P(\kappa n^{-1/2}). \tag{A.30}$$

Using the Cauchy-Schwarz inequality, (A.19), $\lambda_{\min}(\mathbf{\Sigma}_A) > 0$ and $E\|\mathbf{\Pi}(\mathbf{X}_i)\|_2^{2\gamma} = O(\kappa^\gamma)$ for $\gamma > 2$, we establish that

$$\begin{aligned}
E|\tilde{\varepsilon}_{2,i}| \cdot \|\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)\|_2^2 &\leq E\|\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)\|_2^2 |\mathrm{y}_i| + \|\mathbf{\Sigma}_A^{-1}E\{\mathbf{\Pi}(\mathbf{X}_1)\mathrm{y}_1\}\|_2 \cdot E\{\|\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)\|_2^2 \cdot \|\mathbf{\Pi}(\mathbf{X}_i)\|_2\} \\
&\leq O(1)\sqrt{E\|\mathbf{\Pi}(\mathbf{X}_i)\|_2^4 E\mathrm{y}_i^2} + O(1)E\|\mathbf{\Pi}(\mathbf{X}_i)\|_2^3 \\
&= O(\kappa^{3/2}).
\end{aligned}$$

Thus,

$$n^{-1}\sum_{i=1}^n |\tilde{\varepsilon}_{2,i}| \cdot \|\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)\|_2^2 = O_P(\kappa^{3/2}). \tag{A.31}$$

Denoted by $\mathbf{F}_i = \mathbf{\Sigma}_A^{1/2}\{(\mathbf{\Sigma}_n^{(-i)})^{-1} - \mathbf{\Sigma}_A^{-1}\}\mathbf{\Sigma}_A^{1/2}$, $\mathbf{F} = \mathbf{\Sigma}_A^{1/2}\{\mathbf{\Sigma}_n^{-1} - \mathbf{\Sigma}_A^{-1}\}\mathbf{\Sigma}_A^{1/2}$, $\mathbf{G}_i = (n-1)^{-1}\sum_{k=1(\neq i)}^n \widetilde{\mathbf{\Pi}}(\mathbf{X}_k)\mathrm{y}_k$, $\mathbf{G} = (n-1)^{-1}\sum_{k=1}^n \widetilde{\mathbf{\Pi}}(\mathbf{X}_k)\mathrm{y}_k$, $\mathbf{L}_i = \frac{1}{n-1}\sum_{k=1(\neq i)}^n [\widetilde{\mathbf{\Pi}}(\mathbf{X}_k)\mathrm{y}_k - E\{\widetilde{\mathbf{\Pi}}(\mathbf{X}_k)\mathrm{y}_k\}]$, and $\mathbf{L} = \frac{1}{n-1}\sum_{k=1}^n [\widetilde{\mathbf{\Pi}}(\mathbf{X}_k)\mathrm{y}_k - E\{\widetilde{\mathbf{\Pi}}(\mathbf{X}_k)\mathrm{y}_k\}]$. Let

$I_1 = n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_{2,i}\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top \mathbf{F}\mathbf{G}$,

$I_2 = n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_{2,i}\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top \{-(\mathbf{F}_i - \mathbf{F})(\mathbf{G}_i - \mathbf{G})\}$,

$I_3 = n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_{2,i}\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top \mathbf{F}_i(\mathbf{G}_i - \mathbf{G})$,

$I_4 = n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_{2,i}\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top (\mathbf{F}_i - \mathbf{F})\mathbf{G}_i$,

$I_5 = n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_{2,i}\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top (\mathbf{L}_i - \mathbf{L})$, and $I_6 = n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_{2,i}\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top \mathbf{L}$.

Then it can be rewritten that

$$n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_{2,i}(\hat{\varepsilon}_{2,i} - \tilde{\varepsilon}_{2,i}) = n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_{2,i}\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top \mathbf{F}_i\mathbf{G}_i + n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_{2,i}\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top \mathbf{L}_i = \sum_{j=1}^6 I_j.$$

Hence, by (A.28), it suffices to show that $I_j = o_P(n^{-1/2})$. Applying the Cauchy-Schwarz inequality and (A.30)-(A.31), we establish that

$$|I_1| \leq \Big\|n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_{2,i}\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)\Big\|_2 \|\mathbf{F}\|_2 \|\mathbf{G}\|_2 = O_P(\kappa/\sqrt{n})\|\mathbf{F}\|_2\|\mathbf{G}\|_2;$$

9

$$|I_3| \leq \frac{1}{n(n-1)} \sum_{i=1}^{n} \left\| \tilde{\varepsilon}_{2,i} \widetilde{\mathbf{\Pi}}(\mathbf{X}_i) \right\|_2 \cdot \|\mathbf{F}_i\|_2 \cdot \|\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)\mathbf{y}_i\|_2$$

$$\leq \max_{1 \leq i \leq n} \|\mathbf{F}_i\|_2 |\mathbf{y}_i| \frac{1}{n(n-1)} \sum_{i=1}^{n} |\tilde{\varepsilon}_{2,i}| \cdot \left\| \widetilde{\mathbf{\Pi}}(\mathbf{X}_i) \right\|_2^2$$

$$= O_P(\kappa^{3/2}/n) \max_{1 \leq i \leq n} |\mathbf{y}_i| \max_{1 \leq i \leq n} \|\mathbf{F}_i\|_2.$$

Combining (A.15), (A.17), $E|\mathbf{y}_i|^\gamma = O(1)$, $\lambda_{\min}(\mathbf{\Sigma}_A) > 0$ and $\lambda_{\max}(\mathbf{\Sigma}_A) = O(1)$, we arrive

at $\max_{1 \leq i \leq n} |\mathbf{y}_i| = O_P(n^{\frac{1}{2\gamma}})$,

$$\max_{1 \leq i \leq n} \|\mathbf{F}_i\|_2 \leq \max_{1 \leq i \leq n} \|\mathbf{\Sigma}_A^{1/2}\|_2^2 \cdot \|\mathbf{\Sigma}_n^{(-i)}\|_2 \cdot \|\mathbf{\Sigma}_A^{-1}\|_2 \cdot \|\mathbf{\Sigma}_n^{(-i)} - \mathbf{\Sigma}_A\|_2 = O_P(\kappa/\sqrt{n}),$$

and $\|\mathbf{F}\|_2 \leq \|\mathbf{\Sigma}_A^{1/2}\|_2^2 \cdot \|\mathbf{\Sigma}_n\|_2 \cdot \|\mathbf{\Sigma}_A^{-1}\|_2 \cdot \|\mathbf{\Sigma}_n - \mathbf{\Sigma}_A\|_2 = O_P(\kappa/\sqrt{n})$. Using (A.19) and (A.20),

we get

$$E\|\mathbf{G}\|_2^2 = \frac{1}{(n-1)^2} \sum_{i=1}^{n} \sum_{j=1}^{n} E\widetilde{\mathbf{\Pi}}(\mathbf{X}_i)^\top \widetilde{\mathbf{\Pi}}(\mathbf{X}_j)\mathbf{y}_i\mathbf{y}_j$$

$$= \frac{n}{(n-1)^2} E\|\widetilde{\mathbf{\Pi}}(\mathbf{X}_j)\mathbf{y}_i\|_2^2 + \frac{n}{(n-1)} \|E\widetilde{\mathbf{\Pi}}(\mathbf{X}_j)\mathbf{y}_i\|_2^2$$

$$= O(1).$$

Thus, $\|\mathbf{G}\|_2 = O_P(1)$. Since $\gamma > 2$ and $r > 1.5$, we have $I_1 = O_P(\kappa^2/n) = o_P(n^{-1/2})$

and $I_3 = O_P(n^{\frac{1}{2\gamma}} \kappa^{5/2} n^{-3/2}) = o_P(n^{-1/2})$. Similarly, we can show that $I_j = o_P(n^{-1/2})$ for

$j = 2, 4, 5, 6$.

(ii) Notice that

$$\hat{\xi}_i^2 - \tilde{\xi}_i^2 = (\hat{\xi}_i - \tilde{\xi}_i)^2 + 2\tilde{\xi}_i(\hat{\xi}_i - \tilde{\xi}_i).$$

It follows from the Cauchy-Schwarz inequality that

$$\left| n^{-1} \sum_{i=1}^{n} (\hat{\xi}_i^2 - \tilde{\xi}_i^2) \right| \leq n^{-1} \sum_{i=1}^{n} (\hat{\xi}_i - \tilde{\xi}_i)^2 + 2|\tilde{\xi}_i| \cdot |\hat{\xi}_i - \tilde{\xi}_i|$$

$$\leq 2n^{-1} \left(\sum_{i=1}^{n} |\tilde{\xi}_i|^2 \sum_{i=1}^{n} |\hat{\xi}_i - \tilde{\xi}_i|^2 \right)^{1/2} + n^{-1} \sum_{i=1}^{n} (\hat{\xi}_i - \tilde{\xi}_i)^2.$$

By Jensen's inequality and condition A3, we obtain that

$$E|\tilde{\xi}_1|^\gamma = E|\tilde{\varepsilon}_{1,1}^2 - \tilde{\varepsilon}_{2,1}^2|^\gamma \leq 2^{\gamma-1} E|\tilde{\varepsilon}_{1,1}|^{2\gamma} + 2^{\gamma-1} E|\tilde{\varepsilon}_{2,1}|^{2\gamma} = O(1). \tag{A.32}$$

Thus, by condition A3 and Markov's inequality, $n^{-1} \sum_{i=1}^{n} \tilde{\xi}_i^2 = O_P(1)$. To complete the proof, it's sufficient to show that $n^{-1} \sum_{i=1}^{n} |\hat{\xi}_i - \tilde{\xi}_i|^2 = o_P(1)$.

By the definitions of $\hat{\xi}_i$, $\tilde{\xi}_i$, and Jessen's inequality, we establish that

$$
\begin{aligned}
n^{-1} \sum_{i=1}^{n} (\hat{\xi}_i - \tilde{\xi}_i)^2 &= n^{-1} \sum_{i=1}^{n} \Big\{ (\tilde{\varepsilon}_{1,i} - \hat{\varepsilon}_{1,i})^2 - (\tilde{\varepsilon}_{2,i} - \hat{\varepsilon}_{2,i})^2 \\
&\qquad + 2\tilde{\varepsilon}_{1,i}(\hat{\varepsilon}_{1,i} - \tilde{\varepsilon}_{1,i}) - 2\tilde{\varepsilon}_{2,i}(\hat{\varepsilon}_{2,i} - \tilde{\varepsilon}_{2,i}) \Big\}^2 \\
&\leq \frac{4}{n} \sum_{i=1}^{n} \sum_{j=1}^{2} (\tilde{\varepsilon}_{j,i} - \hat{\varepsilon}_{j,i})^4 + \frac{16}{n} \sum_{i=1}^{n} \sum_{j=1}^{2} \tilde{\varepsilon}_{j,i}^2 (\tilde{\varepsilon}_{j,i} - \hat{\varepsilon}_{j,i})^2.
\end{aligned}
$$

Applying Lemma 1, we establish that

$$
\begin{aligned}
n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} (\tilde{\varepsilon}_{j,i} - \hat{\varepsilon}_{j,i})^4 &\leq \max_{1 \leq i \leq n, 1 \leq j \leq 2} (\tilde{\varepsilon}_{j,i} - \hat{\varepsilon}_{j,i})^2 n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} (\tilde{\varepsilon}_{j,i} - \hat{\varepsilon}_{j,i})^2 \\
&= O_P(n^{\frac{1}{\gamma} + \frac{5}{2r+1} - 2});
\end{aligned}
$$

$$
\begin{aligned}
n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} \tilde{\varepsilon}_{j,i}^2 (\tilde{\varepsilon}_{j,i} - \hat{\varepsilon}_{j,i})^2 &\leq \max_{1 \leq i \leq n, 1 \leq j \leq 2} \tilde{\varepsilon}_{j,i}^2 n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} (\tilde{\varepsilon}_{j,i} - \hat{\varepsilon}_{j,i})^2 \\
&= \max_{1 \leq i \leq n, 1 \leq j \leq 2} \tilde{\varepsilon}_{j,i}^2 O_P(n^{\frac{2}{2r+1} - 1}).
\end{aligned}
$$

By condition A3, $E|\tilde{\varepsilon}_{2,1}|^{2\gamma} = O(1)$ for $\gamma > 2$. Then, by the Markov inequality,

$$
P(n^{-1} \max_{1 \leq i \leq n} |\tilde{\varepsilon}_{2,i}|^{2\gamma} > M_n) \leq M_n^{-1} n^{-1} \sum_{i=1}^{n} E|\tilde{\varepsilon}_{2,i}|^{2\gamma} \to 0,
$$

as $M_n \to \infty$. Thus, $\max_{1 \leq i \leq n} |\tilde{\varepsilon}_{2,i}|^2 = O_P(n^{1/\gamma})$. Similarly, $\max_{1 \leq i \leq n} |\tilde{\varepsilon}_{1,i}|^2 = O_P(n^{1/\gamma})$. Hence,

$$
\max_{1 \leq i \leq n, 1 \leq j \leq 2} |\tilde{\varepsilon}_{j,i}|^2 = O_P(n^{1/\gamma}). \tag{A.33}
$$

Combining the above results with $\gamma > 2$ and $r > 1.5$ leads to

$$
n^{-1} \sum_{i=1}^{n} (\hat{\xi}_i - \tilde{\xi}_i)^2 = O_P(n^{\frac{1}{\gamma} + \frac{5}{2r+1} - 2}) + O_P(n^{\frac{1}{\gamma} + \frac{2}{2r+1} - 1}) = o_P(1).
$$

$\diamond$

Proof of Lemma 3. (i) By Lemma 2, we have

$$n^{-1}\sum_{i=1}^{n}(\tilde{\xi}_i - \hat{\xi}_i) = o_P(n^{-1/2}) \quad \text{and} \quad n^{-1}\sum_{i=1}^{n}(\hat{\xi}_i^2 - \tilde{\xi}_i^2) = o_P(1). \tag{A.34}$$

Let $\boldsymbol{\omega}_{1,n} = n^{-1}\sum_{i=1}^{n}(\hat{\xi}_i^2 - \tilde{\xi}_i^2)/(E\hat{\xi}_i^2 + E\tilde{\xi}_i^2)$. By condition A3, we have $E\tilde{\xi}_i^2 \geq \text{Var}(\tilde{\xi}_i) > c_1$.

This, combined with (A.34), ensures that

$$\boldsymbol{\omega}_{1,n} = o_P(1) \quad \text{and} \quad \sup_n E|\boldsymbol{\omega}_{1,n}| \leq 1.$$

Applying Theorem A (Serfling, 1980, page 14), we obtain that $|E\boldsymbol{\omega}_{1,n}| \leq E|\boldsymbol{\omega}_{1,n}| \to 0$.

Then, it is easy to see that

$$E\hat{\xi}_i^2/E\tilde{\xi}_i^2 \to 1. \tag{A.35}$$

Let $\mathcal{X}_n = n^{-1}\sum_{i=1}^{n}(\hat{\xi}_i - E\tilde{\xi}_i)$ and $\mathcal{Y}_n = n^{-1}\sum_{i=1}^{n}(\tilde{\xi}_i - E\tilde{\xi}_i)$. Using the identity $\mathcal{X}_n^2 - \mathcal{Y}_n^2 = (\mathcal{X}_n - \mathcal{Y}_n)^2 + 2\mathcal{Y}_n(\mathcal{X}_n - \mathcal{Y}_n)$, we get

$$\mathcal{X}_n^2 - \mathcal{Y}_n^2 = \left\{n^{-1}\sum_{i=1}^{n}(\hat{\xi}_i - \tilde{\xi}_i)\right\}^2 + 2\mathcal{Y}_n n^{-1}\sum_{i=1}^{n}(\hat{\xi}_i - \tilde{\xi}_i). \tag{A.36}$$

By Markov's inequality and (A.32), it holds that

$$P\left(|\mathcal{Y}_n| > c_n n^{-1/2}\right) \leq n^{-1}c_n^{-2}\sum_{i=1}^{n}E(\tilde{\xi}_i - E\tilde{\xi}_i)^2 \to 0$$

for any $c_n \to \infty$. Hence,

$$\mathcal{Y}_n = O_P(n^{-1/2}). \tag{A.37}$$

This, combined with (A.34) and (A.36), yields that

$$\mathcal{X}_n^2 - \mathcal{Y}_n^2 = o_P(n^{-1}).$$

Define $\boldsymbol{\omega}_{2,n} = (\mathcal{X}_n^2 - \mathcal{Y}_n^2)/(E\mathcal{X}_n^2 + E\mathcal{Y}_n^2)$. Since $E\mathcal{Y}_n^2 = n^{-1}\text{Var}(\tilde{\xi}_1) \geq n^{-1}c_1$, we have

$\boldsymbol{\omega}_{2,n} = o_P(1)$ and $\sup_n E|\boldsymbol{\omega}_{2,n}| \leq 1$. Hence, $E\boldsymbol{\omega}_{2,n} = o(1)$. Similar to (A.35), we get

$$E\mathcal{X}_n^2/E\mathcal{Y}_n^2 \to 1. \tag{A.38}$$

Denoted by $\boldsymbol{\omega}_{3,n} = \{n\mathrm{Var}(\tilde{\xi}_1)\}^{-1/2}\sum_{i=1}^n(\tilde{\xi}_i - \hat{\xi}_i)$. Then, by (A.34) and $\mathrm{Var}(\tilde{\xi}_1) > c_1$, $\boldsymbol{\omega}_{3,n} = o_P(1)$. Furthermore,

$$
\begin{aligned}
E|\boldsymbol{\omega}_{3,n}|^2 &= \{\mathrm{Var}(\tilde{\xi}_1)/n\}^{-1}E(\mathcal{X}_n - \mathcal{Y}_n)^2 \\
&\leq 2(E\mathcal{Y}_n^2)^{-1}(E\mathcal{X}_n^2 + E\mathcal{Y}_n^2) \\
&= 2 + 2E\mathcal{X}_n^2/E\mathcal{Y}_n^2.
\end{aligned}
$$

It follows from (A.38) that $\sup_n E|\boldsymbol{\omega}_{3,n}|^2$ is bounded. Thus,

$$
E\boldsymbol{\omega}_{3,n} \to 0, \text{ or equivalently } \sqrt{n}(E\tilde{\xi}_1 - E\hat{\xi}_1)/\sqrt{\mathrm{Var}(\tilde{\xi}_1)} \to 0. \tag{A.39}
$$

Since $\mathrm{Var}(\tilde{\xi}_1) > c_1 > 0$, $E(\tilde{\xi}_i - \hat{\xi}_i) \to 0$. This, combined with (A.35), yields that $\mathrm{Var}(\hat{\xi}_i) - \mathrm{Var}(\tilde{\xi}_i) = o(1)$, or equivalently

$$
\mathrm{Var}(\tilde{\xi}_i) = \sigma_\xi^2 + o(1). \tag{A.40}
$$

(ii) By (A.32), we obtain that $\max_{1 \leq i \leq n}|\tilde{\xi}_i|^\gamma \leq \sum_{i=1}^n|\tilde{\xi}_i|^\gamma = O_P(n)$, which implies that $\max_{1 \leq i \leq n}|\tilde{\xi}_i| = O_P(n^{1/\gamma})$. Using Lemma 1(i) and the definitions of $\hat{\xi}_i$ and $\tilde{\xi}_i$, we obtain that

$$
\begin{aligned}
|\hat{\xi}_i - \tilde{\xi}_i| &= \left|(\tilde{\varepsilon}_{1,i} - \hat{\varepsilon}_{1,i})^2 - (\tilde{\varepsilon}_{2,i} - \hat{\varepsilon}_{2,i})^2 + 2\tilde{\varepsilon}_{1,i}(\hat{\varepsilon}_{1,i} - \tilde{\varepsilon}_{1,i}) - 2\tilde{\varepsilon}_{2,i}(\hat{\varepsilon}_{2,i} - \tilde{\varepsilon}_{2,i})\right| \\
&\leq 2\max_{1 \leq i \leq n, 1 \leq j \leq 2}(\tilde{\varepsilon}_{j,i} - \hat{\varepsilon}_{j,i})^2 + 4\max_{1 \leq i \leq n, 1 \leq j \leq 2}|\tilde{\varepsilon}_{j,i}(\hat{\varepsilon}_{j,i} - \tilde{\varepsilon}_{j,i})| \\
&= O_P(n^{\frac{1}{\gamma} + \frac{3}{2r+1} - 1}) + O_P(n^{\frac{1}{2\gamma} + \frac{3}{4r+2} - \frac{1}{2}})\max_{1 \leq i \leq n, 1 \leq j \leq 2}|\tilde{\varepsilon}_{j,i}|,
\end{aligned}
$$

which, combined with (A.33) and $r > 1.5$, yields that $\max_{1 \leq i \leq n}|\hat{\xi}_i - \tilde{\xi}_i| = O_P(n^{1/\gamma})$. Thus, $\max_{1 \leq i \leq n}|\hat{\xi}_i| \leq \max_{1 \leq i \leq n}|\tilde{\xi}_i| + \max_{1 \leq i \leq n}|\hat{\xi}_i - \tilde{\xi}_i| = O_P(n^{1/\gamma})$. Note that $\tilde{\xi}_i$ are iid. It follows from (A.32) that $n^{-1}\sum_{i=1}^n\tilde{\xi}_i^2 = E\tilde{\xi}_i^2 + o_P(1)$.

(iii). Case 1: $|a| = +\infty$. By (A.39) and (A.40), we have

$$
a_n^{-1}\sqrt{n}E(\hat{\xi}_i/\sigma_\xi - \tilde{\xi}_i/\sigma_\xi) \to 0.
$$

Under $H_{a,n}^{(1)}$, we know that $a_n^{-1}\sqrt{n}E\hat{\xi}_i/\sigma_\xi = 1$ and $a_n^{-1}\sqrt{n}E\tilde{\xi}_i/\sigma_\xi \to 1$.

13

Case 2: $|a| < +\infty$. By (A.39) and (A.40), we have $\sqrt{n}E(\hat{\xi}_i/\sigma_\xi - \tilde{\xi}_i/\sigma_\xi) \to 0$. Thus, under $H_{a,n}^{(1)}$, we get

$$\sqrt{n}E\hat{\xi}_i/\sigma_\xi = a_n \to a \quad \text{and} \quad \sqrt{n}E\tilde{\xi}_i/\sigma_\xi \to a. \tag{A.41}$$

(iv). By (A.41), we have $E(\tilde{\xi}_1)/\sigma_\xi = O(n^{-1/2})$ when $|a| < \infty$. Then

$$n^{-1}\sum_{i=1}^{n}\tilde{\xi}_i/\sigma_\xi = E(\tilde{\xi}_1/\sigma_\xi) + \frac{1}{n\sigma_\xi}\sum_{i=1}^{n}(\tilde{\xi}_i - E\tilde{\xi}_i)$$

$$= \frac{1}{n\sigma_\xi}\sum_{i=1}^{n}(\tilde{\xi}_i - E\tilde{\xi}_i) + O(n^{-1/2}). \tag{A.42}$$

By condition A3, $\text{Var}(\tilde{\xi}_1) > c_1$. It follows from (A.40) that $\sigma_\xi \geq c_1 + o(1)$. This, combined with (A.37), yields that $n^{-1}\sum_{i=1}^{n}\tilde{\xi}_i = O_P(n^{-1/2})$. In addition, applying Lemma 2 and the triangle inequality, we obtain that

$$\left|n^{-1}\sum_{i=1}^{n}\hat{\xi}_i\right| \leq \left|n^{-1}\sum_{i=1}^{n}(\hat{\xi}_i - \tilde{\xi}_i)\right| + \left|n^{-1}\sum_{i=1}^{n}\tilde{\xi}_i\right| = O_P(n^{-1/2}).$$

$\diamond$

Proof of Lemma 4. Since $n^{1/\gamma - 1/2} = o(1)$, there exists a sequence ϕ_n such that $\phi_n = o(n^{-1/\gamma})$ and $n^{-1/2} = o(\phi_n)$. Define $\Lambda_n = \{\lambda : |\lambda| \leq \phi_n\}$. Then, by the result $\max_{1 \leq i \leq n}|\hat{\xi}_i| = O_P(n^{1/\gamma})$ in Lemma 3(ii), we get $\max_{1 \leq i \leq n, \lambda \in \Lambda_n}|\lambda\hat{\xi}_i| = o_P(1)$. Let

$$\bar{\lambda} = \arg\min_{\lambda \in \Lambda_n}\sum_{i=1}^{n}\log(1 + \lambda\hat{\xi}_i).$$

Then $\max_{1 \leq i \leq n}|\bar{\lambda}\hat{\xi}_i| = o_P(1)$. Using Taylor's expansion, with probability going to 1, we obtain that

$$0 \leq \sum_{i=1}^{n}\log(1 + \bar{\lambda}\hat{\xi}_i) = \bar{\lambda}\sum_{i=1}^{n}\hat{\xi}_i - \frac{\bar{\lambda}^2}{2}\sum_{i=1}^{n}\frac{\hat{\xi}_i^2}{(1 + c_i^*\bar{\lambda}\hat{\xi}_i)^2}$$

$$\leq |\bar{\lambda}|\left|\sum_{i=1}^{n}\hat{\xi}_i\right| - \bar{\lambda}^2 c\sum_{i=1}^{n}\hat{\xi}_i^2 \tag{A.43}$$

for some constants $0 \leq c_i^* \leq 1$ and $0 < c \leq 1$. By condition A3, $\text{Var}(\tilde{\xi}_1) > c_1$, and by Lemma 3(ii), $n^{-1}\sum_{i=1}^{n}|\tilde{\xi}_i|^2 = E|\tilde{\xi}_1|^2 + o_P(1)$. Then, applying Lemma 2(ii), we establishes that

$$n^{-1}\sum_{i=1}^{n}\hat{\xi}_i^2 \geq c_1 + o_P(1). \tag{A.44}$$

14

By Lemma 3(iv), we have $n^{-1} \sum_{i=1}^{n} \hat{\xi}_i = O_P(n^{-1/2})$, which, combined with (A.43)-(A.44), yields that

$$\bar{\lambda} = O_P(n^{-1/2}) = o_P(\phi_n).$$

Thus, with probability tending to 1, $\bar{\lambda}$ is in the interior of Λ_n. Since $\sum_{i=1}^{n} \log(1 + \lambda \hat{\xi}_i)$ is concave, $P(\hat{\lambda} = \bar{\lambda}) \to 1$. Hence, $\hat{\lambda} = O_P(n^{-1/2})$. $\diamond$

References

Serfling, R. J. (1980). Approximation theorems of mathematical statistics. *New York: Wiley*.