A Nondegenerate Vuong Test and Post Selection Confidence Intervals for Semi/Nonparametric Models *

Zhipeng Liao † Xiaoxia Shi ‡

October 8, 2019

Abstract

This paper proposes a new model selection test for the statistical comparison of semi/nonparametric models based on a general quasi-likelihood ratio criterion. An important feature of the new test is its uniformly exact asymptotic size in the overlapping nonnested case, as well as in the easier nested and strictly nonnested cases. The uniform size control is achieved without using pre-testing, sample-splitting, or simulated critical values. We also show that the test has nontrivial power against all $\sqrt{n}$-local alternatives and against some local alternatives that converge to the null faster than $\sqrt{n}$. Finally, we provide a framework for conducting uniformly valid post model selection inference for model parameters. The finite sample performance of the nondegenerate test and that of the post model selection inference procedure are illustrated in a mean-regression example by Monte Carlo.

JEL Classification: C14, C31, C32

Keywords: Asymptotic Size, Model Selection/Comparison Test, Post Model Selection Inference, Semi/Nonparametric Models

*We acknowledge our helpful discussions with Ivan Canay, Xiaohong Chen, Denis Chetverikov, Jinyong Hahn, Bruce E. Hansen, Michael Jansson, Whitney Newey, Rosa Matzkin, Peter C.B. Phillips, Quang H. Vuong, and participants in econometrics workshops at Cemmap/UCL, Chinese University of Hong Kong, Duke, Erasmus University Rotterdam, Harvard/MIT, HKUST, Northwestern University, SUFE, Singapore Management University, Tilburg University, Tinbergen Institute, UCLA, UC-Riverside, UCSD, USC, UW-Madison and Yale. Any errors are our own.

†Department of Economics, UC Los Angeles, 8379 Bunche Hall, Mail Stop: 147703, Los Angeles, CA 90095. Email: zhipeng.liao@econ.ucla.edu.

‡Department of Economics, University of Wisconsin at Madison. Email: xshi@ssc.wisc.edu.

1 Introduction

Model selection is an important issue in many empirical work. For example, in economic studies, there are often competing theories for one phenomenon. Even when there is only one theory, it can rarely pin down an empirical model to take to the data. Model selection tests are tools to determine the best model out of multiple competing models with a pre-specified statistical confidence level. One such test was proposed in Vuong (1989) to select from two parametric likelihood models according to their Kullback-Leibler information criterion (KLIC). The test determines the statistical significance of KLIC difference and, when the difference is significant, draws the directional conclusion that one model is closer to the truth than the other. This test has been widely used in empirical work due to its straightforward interpretation and implementation,[1] and it has been extended to many settings besides the likelihood one.

The studentized quasi-likelihood ratio (QLR) test statistic used in Vuong (1989) may have different asymptotic distributions under the null hypothesis, depending on whether the asymptotic variance of the QLR is degenerate. The degeneracy is unknown when the models compared are overlapping nonnested. In this case, a test based on such a test statistic and a standard critical value may not be uniformly valid and adding a pretest of the degeneracy does not provide a satisfactory solution, as shown in Shi (2015b). What is especially troubling is that the QLR-based test has a bias term that favors complex models. As a result, a user could manipulate the model selection result by unnecessarily increasing or decreasing the complexity of certain models. Shi (2015b) develops a solution in the context of parametric models, but Shi's test does not apply to semi/nonparametric models where the problem is in fact exacerbated.

The first contribution of this paper is to extend the conceptual idea of Shi (2015b) to semi/nonparametric models. Like Shi's test, our test corrects for bias caused by difference in model complexity and achieves uniform asymptotic validity regardless of model relationship. Unlike Shi's test, our revised QLR statistic is uniformly asymptotically normal, leading to a very simple testing procedure. The nonparametric component in one or both of the models, while making the asymptotic theory much more complicated, remarkably simplifies the testing procedure relative to Shi (2015b). We use linear sieve approximation for the nonparametric components (ref, e.g., Chen (2007)). As such, the asymptotic theory also provides a good approximation for parametric models with a large number of parameters.

The second contribution of this paper is a valid inference for the model parameters after the model selection test. Post model selection inference on one hand is unavoidable in most

[1]See, e.g., Cameron and Heckman (1998),Coate and Conlin (2004),Paulson et al. (2006), Gowrisankaran and Rysman (2012), Moines and Pouget (2013), Barseghyan et al. (2013), Karaivanov and Townsend (2014), Kendall et al. (2015), Gandhi and Serrano-Padial (2015), to name only a few.

applications, and on the other hand is difficult to do correctly. For example, if post-model selection confidence intervals are constructed as if no model selection had been conducted, Leeb and Pötscher (2005) show that the resulting confidence intervals may have coverage probabilities very different from the nominal level. In this paper, we provide two types of uniformly asymptotically valid confidence intervals for parameters post model selection.

The rest of the introduction is devoted to the discussion of related literature.

The literature on the QLR model selection test. Although the QLR test proposed in Vuong (1989) has been widely used in the empirical studies and extended to many non-likelihood settings,[2] its property on the size control draws researchers' attention only recently. As mentioned above, the model selection part of this paper extends the conceptual idea of Shi (2015b) to semi/non-parametric models and propose a test with uniform size control for semi/non-parametric models. A few other papers in the literature of the Vuong test also achieve uniform asymptotic size control. These include Li (2009), Schennach and Wilhelm (2017), Hsu and Shi (2017) and Shi (2015a). These papers do not deal with semi/non-parametric models and each achieves uniform size control by a different technique. Li (2009) achieves uniformity thanks to the simulation noise brought about by numerical integration. Schennach and Wilhelm (2017) employ a sophisticated split-sample technique. Hsu and Shi (2017) introduce artificial noise to their test statistic. Shi (2015a) uses a pretest with a diverging threshold.

The consistent model specification testing literature. Although the main advantage of our revised QLR test is in the overlapping nonnested cases, it can be applied to and has uniform asymptotic similarity in the nested cases as well. In such cases, our test is a model specification test of the nested model against the alternative of the nesting model. As such, it is related to Hong and White (1995), Fan and Li (1996), Lavergne and Vuong (2000), and Aït-Sahalia et al. (2001) among others (see e.g. Aït-Sahalia et al. (2001) for a comprehensive literature review). Our test reduces to the heteroskedasticity-robust version of Hong and White (1995) based on series regression when a parametric conditional mean model is compared to a nonparametric one, and reduces to a series regression-based version of Aït-Sahalia et al.'s (2001) test when two nested nonparametric regressions are compared based on a weighted mean-squared error criterion. Our test applies to the testing problems in Fan and Li (1996) and Lavergne and Vuong (2000) but differs from the tests therein.

Post model selection inference. Our post model selection (PMS) inference has two parts. The first part regards conditional inference on model-specific parameters. This part is inspired by Tibshirani et al. (2016), who provide valid p-values and confidence intervals for post Lasso inference in a linear regression context with Gaussian noise. Their result is extended in Tibshirani et al. (2015) and Tian and Taylor (2015) to other linear regressions settings. We generalize Tibshirani

[2]Extensions include Lavergne and Vuong (1996), Rivers and Vuong (2002), Kitamura (2000), among others.

et al. (2016) to post model test inference for general semi-nonparametric models, and provide asymptotically exact confidence intervals without imposing special structures on the models or requiring knowledge of a variance-covariance matrix. The second part of our PMS inference analysis regards inference on common parameters of the two models. This part shares the objective of the methods surveyed in Belloni et al. (2014). However, this type of post selection inference is highly context specific, and the surveyed methods do not apply to post selection inference in general models.

The nonnested hypotheses literature. Since Vuong's (1989) test is most commonly used to select between nonnested models, it is often linked to the literature of nonnested hypotheses featuring Cox (1961, 1962), Atkinson (1970), Pesaran (1974), Pesaran and Deaton (1978), Mizon and Richard (1986), Gourieroux and Monfort (1995), Ramalho and Smith (2002), and Bontemps et al. (2008) among others. This literature does not share the objective of Vuong's test. Rather than focusing on the relative fit of the models, earlier part of this literature focuses on testing the correct specification of one model with power directed toward the other model. Later part of this literature focuses on the ability of one model to encompass empirical features of the other model. To our knowledge, the uniform validity of these tests when the models under consideration are overlapping nonnested has not been studied, and may be an interesting topic for future research.[3]

The rest of the paper is organized as follows. Section 2 sets up our testing framework and gives three examples. Section 3 describes our test in detail. Section 4 establishes the asymptotic size and the local power of our test. Section 5 illustrates the construction of our test in the mean-regression context. Section 6 provides the uniformly valid post model selection inference procedures. Section 7 shows Monte Carlo results of a mean-regression example. Section 8 applies the proposed nondegenerate test and conditional confidence interval to a schooling choice example, and Section 9 concludes. Proofs and other supplemental materials are included in the Supplemental Appendix.

Notation. Let C, C_1 and C_2 be generic positive constants whose values do not change with the sample size. For any column vector a, let a' denote its transpose and $\|a\|$ its ℓ_2-norm. For any square matrix A, $A(i,j)$ denotes the element in the ith row and jth column of A, $\|A\|$ denotes the operator norm, and A^+ denotes its Moore-Penrose inverse. Let $\rho_{\min}(A)$ and $\rho_{\max}(A)$ be the smallest and largest eigenvalues of A in terms of absolute value, respectively. Let $tr(A)$ denote the trace of matrix A. For any square matrices A_1 and A_2, $diag(A_1, A_2)$ denotes the block diagonal matrix with A_1 being the leading submatrix. Let $N(\mu, \Sigma)$ stand for a normal random vector with mean μ and variance-covariance matrix Σ. For any (possibly random) positive sequences $\{a_n\}_{n=1}^{\infty}$

[3]The lack of uniform size control of the Cox test when the DGP space is not restricted is illustrated in Loh (1985). However, uniform size control under reasonable restrictions on the DGP space for the Cox test and other nonnested hypotheses tests is still an interesting problem yet to be explored.

and $\{b_n\}_{n=1}^{\infty}$, $a_n = O_P(b_n)$ means that $\lim_{c\to\infty} \limsup_n \Pr(a_n/b_n > c) = 0$; and $a_n = o_P(b_n)$ means that for all $\varepsilon > 0$, $\lim_{n\to\infty} \Pr(a_n/b_n > \varepsilon) = 0$. For any $p \in (0,1)$, let z_p denote the p quantile of the standard normal distribution.

2 General Setup

2.1 Setup

Let $Z \in \mathcal{Z} \subseteq R^{d_z}$ be an observable random vector with distribution F_0. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two models about F_0; that is, $\mathcal{M}_1$ and $\mathcal{M}_2$ are two sets of probability distributions on R^{d_z} defined by modeling assumptions. We are interested in testing the null hypothesis of equal fit:

$$H_0 : f(\mathcal{M}_1, F_0) = f(\mathcal{M}_2, F_0), \tag{2.1}$$

where $f(\cdot,\cdot)$ is a generic measure of fit. The alternative hypothesis can be either

$$H_1^{\text{2-sided}} : f(\mathcal{M}_1, F_0) \neq f(\mathcal{M}_2, F_0) \text{ or } H_1^{\text{1-sided}} : f(\mathcal{M}_1, F_0) > f(\mathcal{M}_2, F_0). \tag{2.2}$$

The two-sided test indicates that the two models have (statistically) significantly different fit for the observed data when it rejects H_0, and the one-sided test indicates that model $\mathcal{M}_1$ fits the observed data significantly better when it rejects H_0. It is the goal of this paper to develop a simple test of equal fitting with uniform asymptotic validity and good power properties.

The fit measure $f(\cdot,\cdot)$ is context-specific and should be chosen to best suit the empirical model comparison need. We focus on a given fit measure of the following form:

$$f(\mathcal{M}_j, F_0) = \max_{\alpha_j \in \mathcal{A}_j} E_{F_0}[m_j(Z;\alpha_j)] = E_{F_0}\left[m_j(Z;\alpha^*_{F_0,j})\right], \text{ for } j = 1,2, \tag{2.3}$$

where $E_{F_0}[\cdot]$ denotes the expectation taken under F_0, $m_j(\cdot;\cdot)$ is a user-chosen link function that is the central component of the fit measure, α_j is the parameter in model $\mathcal{M}_j$, $\mathcal{A}_j$ is the possibly infinite-dimensional parameter space, and $\alpha^*_{F_0,j}$ is the pseudo-true parameter value of model j defined as $\alpha^*_{F_0,j} = \arg\max_{\alpha_j \in \mathcal{A}_j} E_{F_0}[m_j(Z;\alpha_j)]$.[4]

To fix ideas, consider the most common examples of $\mathcal{M}_j$ and $f(\mathcal{M}_j, F_0)$, $j = 1,2$:

[4]Following the literature (see, e.g., Stone (1985) and Ai and Chen (2007)), we assume that the pseudo true parameter $\alpha^*_{F_0,j}$ exists, is unique, and lies in the interior of $\mathcal{A}_j$ for $j = 1,2$ throughout the paper. The sufficient conditions to ensure the existence of the pseudo true parameter $\alpha^*_{F_0}$ in general semi/nonparametric models are: (i) the population function $Q_{F_0}(\alpha) = E_{F_0}[m(Z,\alpha)]$ is continuous at any $\alpha \in \mathcal{A}$ under certain metric d (e.g., the L_2-metric or the uniform metric); and (ii) the parameter space $\mathcal{A}$ is compact with respect to d. Low level sufficient conditions for the existence and uniqueness of $\alpha^*_{F_0,j}$ in specific models can be found in Stone (1985) and Ai and Chen (2007). See Section 5 for more discussion in the regression models.

Example 1 (Likelihood Ratio) *Consider $Z = (W', X')'$. Many structural models used in empirical economics can be written as a conditional likelihood model of Z given X, i.e. (ignoring the model index j)*

$$\mathcal{M} = \left\{ F : dF_{Z|X}(z|x)/d\mu_z = \phi(z|x;\alpha),\ \forall z,\ \text{for some } \alpha \in \mathcal{A} \right\}, \tag{2.4}$$

where $F_{Z|X}$ is the conditional distribution of Z given X implied by F, $dF_{Z|X}(z|x)/d\mu_z$ is the Radon-Nykodym density of $F_{Z|X}$ with respect to a basic measure (μ_z) on the space of Z, ϕ is a known function, α is a possibly infinite-dimensional unknown parameter, and $\mathcal{A}$ is its parameter space. For such a model, a natural fit measure is the population conditional log-likelihood, which is the $f(\mathcal{M}, F_0)$ defined in equation (2.3) with

$$m(Z;\alpha) = \log \phi(Z|X;\alpha). \tag{2.5}$$

Note that with $f(\mathcal{M}, F_0)$ defined this way, $\{f(\mathcal{M}, F_0) - f(\{F_0\}, F_0)\}$ is the Kullback-Leibler pseudo-distance from model $\mathcal{M}$ to the true distribution F_0. Vuong's (1989) original test is designed for such a likelihood context with α for both models being finite-dimensional, although Shi (2015b) shows that it may have size distortion. Shi (2015b) proposes a uniformly valid procedure for the parametric likelihood case.

Example 2 (Squared Error) *Consider $Z = (Y, X')'$, where Y is a dependent variable, X is a vector of regressors. A mean-regression model may be written as*

$$\mathcal{M} = \left\{ F : E_F[Y|X = x] = g(x;\alpha),\ \forall x,\ \text{for some } \alpha \in \mathcal{A} \right\}, \tag{2.6}$$

where $g(\cdot;\cdot)$ is a known regression function, α is a possibly infinite-dimensional unknown parameter and $\mathcal{A}$ is its parameter space.[5] *For such a model, a commonly used fit measure is the population regression mean-squared error, which is $f(\mathcal{M}, F_0)$ defined in equation (2.3) with*

$$m(Z;\alpha) = -\left|Y - g(X;\alpha)\right|^2/2. \tag{2.7}$$

Example 3 (Check Function) *Consider $Z = (Y, X')'$, where Y is a dependent variable, X is a vector of regressors. A quantile-regression model may be written as*

$$\mathcal{M} = \left\{ P : Q_{\tau,F}(Y|X = x) = g(x;\alpha),\ \forall x,\ \text{for some } \alpha \in \mathcal{A} \right\}, \tag{2.8}$$

[5]Sometimes, regression models are used without explicitly or implicitly assuming the best fitting regression function to be $E(Y|X = x)$. Nonetheless, the regression mean-squared error criterion often still is used to compare the models. In those cases, the test developed in this paper still applies.

where $Q_{\tau,F}(Y|X)$ is the conditional τ-th quantile of Y given X under F with $\tau \in (0,1)$, $g(\cdot;\cdot)$ is a known regression function, α is a possibly infinite-dimensional unknown parameter, and $\mathcal{A}$ is its parameter space. Similar to the example above, a reasonable fit measure is the expected check function of Y from the best conditional τ-th quantile function in the model, which is $f(\mathcal{M}, F_0)$ defined in equation (2.3) with

$$m(Z;\alpha) = (I\{Y \leq g(X;\alpha)\} - \tau)[Y - g(X;\alpha)] \tag{2.9}$$

where $I\{\cdot\}$ denotes the indicator function.

2.2 Model Relationships

The following terms for model relationships are mentioned in the introduction, and will be used in later sections when we discuss the uniform validity of our test in detail.

Definition 1 (Strictly Nonnested) *Models $\mathcal{M}_1$ and $\mathcal{M}_2$ are strictly nonnested if there does not exist a pair $(\alpha_1, \alpha_2) \in \mathcal{A}_1 \times \mathcal{A}_2$ such that $m_1(z;\alpha_1) = m_2(z;\alpha_2)\ \forall\ z \in \mathcal{Z}$.*

Definition 2 (Overlapping) *Models $\mathcal{M}_1$ and $\mathcal{M}_2$ are overlapping if they are not strictly nonnested.*

Definition 3 (Nested) *Model $\mathcal{M}_1$ nests model $\mathcal{M}_2$ if, for each $\alpha_2 \in \mathcal{A}_2$, there exists an $\alpha_1 \in \mathcal{A}_1$ such that $m_1(z;\alpha_1) = m_2(z;\alpha_2)$ for any $z \in \mathcal{Z}$.*

Clearly, the overlapping case include the nested case. If the models are overlapping but not nested, we say that the models are *overlapping nonnested*. If the models are mutually nested (i.e. $\mathcal{M}_1$ nests $\mathcal{M}_2$, and $\mathcal{M}_2$ nests $\mathcal{M}_1$), then the models are *observationally equivalent*.[6] We exclude the case where the models are observationally equivalent from our discussion, since in this trivial case, H_0 always holds regardless of the true data distribution and no statistical method can distinguish the two. The model relationship determines whether the random variable $m_1(Z;\alpha^*_{F_0,1}) - m_2(Z;\alpha^*_{F_0,2})$ is always, never, or sometimes degenerate (i.e., almost surely zero) under H_0.[7] [8] Since whether $m_1(Z;\alpha^*_{F_0,1}) - m_2(Z;\alpha^*_{F_0,2})$ is degenerate or not affects the asymptotic distribution of standard quasi-likelihood ratio statistic, uniformity issue arises when its status is unknown.

[6] This definition of model equivalence is consistent with that in Pesaran and Ulloa (2008).

[7] This variable is clearly not almost surely zero under H_1, because its mean is different from zero.

[8] Some readers may confuse the degeneracy of $m_1(Z;\alpha^*_1) - m_2(Z;\alpha^*_2)$ under H_0 with the observational equivalence of the models $\mathcal{M}_1$ and $\mathcal{M}_2$. The former does not imply the latter, as one can easily see in the following simplistic example. Let $\mathcal{M}_1$ be a mean-regression model $E[Y|X] = \alpha_1(X)$ with the space $\mathcal{A}_1$ of α_1 including the zero function, and let $\mathcal{M}_2$ be another mean-regression model $E[Y|X] = 0$. Then our H_0 is the same as the hypothesis that $E[Y|X] = 0$ *a.s.*. Under H_0, the difference in squared residuals is degenerate to zero. But the models $\mathcal{M}_1$ and $\mathcal{M}_2$ are clearly not observationally equivalent.

As we will see, the test statistic that we construct is asymptotically standard normal under H_0 regardless of whether $m_1(Z;\alpha^*_{F_0,1}) - m_2(Z;\alpha^*_{F_0,2})$ is degenerate. This leads to a test that is uniformly asymptotically valid across all cases and all types of model relationship. Such uniformity is of practice importance for a number of reasons. First, in many nonnested model selection scenarios, the competing models are not completely incompatible to each other, in which case they are overlapping. Second, establishing strict nonnestedness is difficult for structural models used in empirical analysis. Using our test obviates the need for doing this. Third, even when the models are strictly nonnested, tests ignoring the uniformity issue may still have severe size distortion (over-rejection) in finite samples when both models can closely describe the data distribution, while our test does not suffer from this kind of distortion.

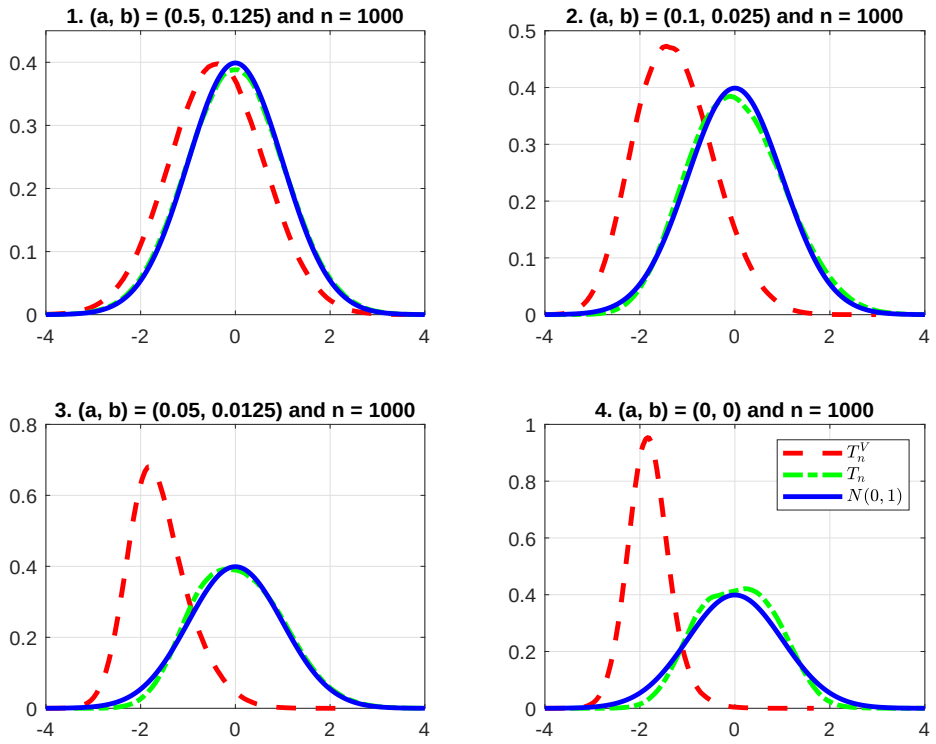
2.3 Illustration of the Uniformity Issue

To further illustrate the uniformity issue, we presents a simple simulation study in Figure 1 compares two parametric linear regression models based on their mean-squared error. We show both the distribution of the standardized QLR statistic (as used in Vuong (1989), T_n^V) and our test statistic (T_n) in the figure. Here, model 1 has two regressors and model 2 has 17 regressors.

The red dashed line represents the finite sample density of T_n^V defined in (3.5) below. In the pointwise asymptotic framework, under H_0, T_n^V has asymptotic standard normal distribution when the latent parameters $(a, b) \neq 0$ and asymptotic weighted chi-square distribution when $(a, b) = 0$. Suppose that one conducts model selection test using the critical value from the standard normal distribution. Although such a test is justified by the asymptotic distribution of T_n^V when (a, b) are not zero, we see that it over-rejects under the null even in this case, as illustrated in the first three scenarios considered in Figure 1. When the latent parameters (a, b) are close to zero, this test is severely over-sized and strongly in favor of the large model, i.e., model 2. As the figure also shows, the standard normal distribution is a poor approximation to the finite sample density of T_n^V when (a, b) are not far enough away from zero, this also suggests that it is tricky to use pre-testing of the latent model structure construct a valid model selection test.

The green dash-dotted line represents the finite sample density of our revised QLR statistic T_n defined in (3.16) below. It is clear that the distribution of T_n is robust against small values of (a, b), and its finite sample density is very close to the standard normal. Thus, the test using T_n and critical value from the standard normal has better size control than the test based on T_n^V and it is also not biased by the relative complexities of the two models.

Figure 1: Finite Sample Densities of T_n^V and T_n under the Null Hypothesis



Notes: (i). The simulated data is generated from the equation $Y_i = 0.5X_{1,i} + aX_{2,i} + b\sum_{k=1}^{16} X_{2+k,i} + u_i$, where (a, b) is set to different values in the four subgraphs and the values guarantee equal fitting of the candidate models, and $(X_{1,i}, ..., X_{18,i}, u_i)'$ is a standard normal random vector; (ii) model 1: $Y_i = X_{1,i}\theta_{1,1} + aX_{2,i}\theta_{1,2} + u_{1,i}$ is compared with model 2: $Y_i = X_{2,i}\theta_{2,2} + b\sum_{k=1}^{16} X_{2+k,i}\theta_{2,2+k} + u_{2,i}$ in their expected squared errors; (iii) the finite sample densities of the existing QLR statistic T_n^V and our statistic T_n are approximated using 1,000,000 simulated samples.

3 Description of Our Model Selection Test

Suppose that there is an i.i.d. sample $\{Z_i\}_{i=1}^n$ of Z. In this section we describe our test for (2.1) based on this sample. The construction of the test is grounded on the asymptotic expansion established in the next section. We focus on the steps of the construction in this section for easy reference for potential users of the test.

We use linear sieve approximation for the unknown functions, and use sieve M-estimator for estimation.[9] The specific procedure is explained now. For $j = 1, 2$, let $\mathcal{A}_{j,k_j}$ denote a finite

[9]Many properties of the sieve M-estimator, including consistency, rate of convergence and asymptotic normality are established in the literature. See, e.g., Chen (2007) for a recent survey on this topic.

dimensional approximation of the parameter space $\mathcal{A}_j$, which satisfies

$$\mathcal{A}_{j,k_j} = \{\alpha_{j,k_j}(\cdot) : \alpha_{j,k_j}(\cdot) = \alpha_j(\beta_{j,k_j}) \equiv P_{j,k_j}(\cdot)'\beta_{j,k_j} : \beta_{j,k_j} \in B_{j,k_j} \subseteq R^{k_j}\}, \tag{3.1}$$

where $P_{j,k_j}(\cdot) = \left[p_{j,1}(\cdot), \ldots, p_{j,k_j}(\cdot)\right]'$ is a k_j-dimensional vector of user-chosen approximating functions such as polynomials and splines, k_j is a positive integer which may diverge with the sample size n. In the rest of the paper, we write $\alpha_{k_j}(\cdot) = \alpha_{j,k_j}(\cdot)$, $P_{k_j}(\cdot) = P_{j,k_j}(\cdot)$ and $\beta_{k_j} = \beta_{j,k_j}$ for $j = 1, 2$ for ease of notation.

To construct the test, we first estimate the fit of each model with the sample analogue estimator. For $j = 1, 2$, define

$$\widehat{f}(\mathcal{M}_j, F_0) = n^{-1}\sum_{i=1}^{n} m_j(Z_i; \widehat{\alpha}_{k_j}) \tag{3.2}$$

where $\widehat{\alpha}_{k_j} = \alpha_j(\widehat{\beta}_{k_j})$ is an M-estimator defined with

$$\widehat{\beta}_{k_j} = \arg\max_{\beta_{k_j} \in B_{j,k_j}} n^{-1}\sum_{i=1}^{n} m_j\left[Z_i; \alpha_j(\beta_{k_j})\right]. \tag{3.3}$$

For notation simplicity, we define the pseudo-density ratio:

$$\ell(Z; \alpha) = m_1(Z; \alpha_1) - m_2(Z; \alpha_2) \tag{3.4}$$

where $\alpha = (\alpha_1, \alpha_2) \in \mathcal{A}_1 \times \mathcal{A}_2$. We also define $\alpha^*_{F_0} = (\alpha^*_{F_0,1}, \alpha^*_{F_0,2})$, $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$, $\mathbf{k} = (k_1, k_2)$, $\beta_{\mathbf{k}} = (\beta'_{k_1}, \beta'_{k_2})'$, $\mathcal{A}_{\mathbf{k}} = \mathcal{A}_{1,k_1} \times \mathcal{A}_{2,k_2}$, $\alpha_{\mathbf{k}} = \alpha(\beta_{\mathbf{k}}) = (\alpha_1(\beta_{k_1}), \alpha_2(\beta_{k_2}))$, and $\widehat{\alpha}_{\mathbf{k}} = (\widehat{\alpha}_{k_1}, \widehat{\alpha}_{k_2})$.

Since the null hypothesis H_0 is equivalent to $E_{F_0}[\ell(Z; \alpha^*_{F_0})] = 0$, one may be tempted to suggest treating $E_{F_0}[\ell(Z; \alpha^*_{F_0})]$ as a parameter and constructing a Student t-like test for this hypothesis. In other words, the suggestion would be to construct the test statistic

$$T_n^V \equiv \bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}})(n^{-1/2}\widehat{\omega}_n(\widehat{\alpha}_{\mathbf{k}}))^{-1}, \tag{3.5}$$

where $\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}})$ is the sample analogue estimator of $E_{F_0}[\ell(Z; \alpha^*_{F_0})]$ and $n^{-1/2}\widehat{\omega}_n(\widehat{\alpha}_{\mathbf{k}})$ is the sample analogue of its standard deviation:

$$\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}}) = n^{-1}\sum_{i=1}^{n}\ell(Z_i; \widehat{\alpha}_{\mathbf{k}}) \text{ and } \widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k}}) = n^{-1}\sum_{i=1}^{n}[\ell(Z_i; \widehat{\alpha}_{\mathbf{k}}) - \bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}})]^2. \tag{3.6}$$

Then one would construct tests of the form: $\varphi_n^{V,\text{2-sided}}(p) = 1\{|T_n^V| > z_{1-p/2}\}$ or $\varphi_n^{V,\text{1-sided}}(p) = 1\{T_n^V > z_{1-p}\}$. In fact, such tests are analogous extensions of Vuong's (1989) (one-step) test to the semi/non-parametric context. Thus, we refer to them as the "naive extension" tests hereafter.

The rationale behind the naive extension test is that $n^{1/2}\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}}) = n^{1/2}\bar{\ell}_n(\alpha^*_{F_0}) + o_p(1) \to_d N(0, \omega^2_{F_0,*})$ and $\widehat{\omega}^2_n = \omega^2_{F_0,*} + o_p(1)$, where $\omega^2_{F_0,*} = Var_{F_0}(\ell(Z; \alpha^*_{F_0}))$. However, this asymptotic approximation can be very poor when $\omega^2_{F_0,*}$ is close to or equal to zero. When the models are overlapping nonnested, both small positive values and the zero value are possible for $\omega^2_{F_0,*}$ under H_0, depending on the *unknown* data distribution F_0. Thus, the naive extension test often fails to have the correct level in a finite sample.[10]

The intuition of the failure of the naive extension test can be seen from the following heuristic second order expansion of the QLR statistic.[11] Let

$$\beta^*_{k_j} = \arg \max_{\beta_{k_j} \in B_{j,k_j}} E_{F_0}\left[m_j(Z; \alpha_j(\beta_{k_j}))\right], \tag{3.7}$$

where we suppress the dependence of $\beta^*_{k_j}$ on F_0 for notational convenience. We assume that the sieve coefficients $\beta^*_{k_j}$ are in the interior of their spaces B_{j,k_j} for any k_j. Let $\alpha^*_{k_j}(\cdot) = P_{k_j}(\cdot)'\beta^*_{k_j}$. Then $\alpha^*_{k_j}$ is the sieve approximator of the pseudo true parameter $\alpha^*_{F_0,j}$ on the finite dimensional space $\mathcal{A}_{j,k_j}$. Let $\ell_{\alpha,\mathbf{k}}(Z;\alpha)$ be the "score" function of $\ell(Z;\alpha)$ evaluated at $\alpha \in \mathcal{A}_{\mathbf{k}}$. When $\ell(Z; \alpha(\beta_{\mathbf{k}}))$ is differentiable in $\beta_{\mathbf{k}}$, we can let

$$\ell_{\alpha,\mathbf{k}}(Z;\alpha_{\mathbf{k}}) = \partial \ell(Z;\alpha_{\mathbf{k}})/\partial\beta_{\mathbf{k}} \text{ and } \bar{\ell}_{\alpha,\mathbf{k},n}(\alpha^*_{\mathbf{k}}) = n^{-1}\sum_{i=1}^n \ell_{\alpha,\mathbf{k}}(Z_i;\alpha^*_{\mathbf{k}}) \tag{3.8}$$

where $\alpha^*_{\mathbf{k}} = (\alpha^*_{k_1}, \alpha^*_{k_2})$. Then a second order Taylor expansion of $\bar{\ell}_n(\alpha^*_{\mathbf{k}})$ around $\widehat{\alpha}_{\mathbf{k}}$ gives:

$$\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}}) - E_{F_0}[\ell(Z;\alpha^*_{F_0})] \approx \bar{\ell}_n(\alpha^*_{\mathbf{k}}) - E_{F_0}[\ell(Z;\alpha^*_{F_0})] - 2^{-1}\bar{\ell}_{\alpha,\mathbf{k},n}(\alpha^*_{\mathbf{k}})' H^{-1}_{F_0,\mathbf{k}} \bar{\ell}_{\alpha,\mathbf{k},n}(\alpha^*_{\mathbf{k}}), \tag{3.9}$$

where

$$H_{F_0,\mathbf{k}} = diag\left(\frac{\partial^2 E_{F_0}[m_1(Z;\alpha_{k_1})]}{\partial\beta_{k_1}\partial\beta'_{k_1}}, -\frac{\partial^2 E_{F_0}[m_2(Z;\alpha_{k_2})]}{\partial\beta_{k_2}\partial\beta'_{k_2}}\right) = diag\left(H_{F_0,k_1}, -H_{F_0,k_2}\right). \tag{3.10}$$

Appropriate conditions and the central limit theorem imply that $n^{1/2}\left\{\bar{\ell}_n(\alpha^*_{\mathbf{k}}) - E_{F_0}[\ell(Z;\alpha^*_{F_0})]\right\} \to_d N(0, \omega^2_{F_0,*})$ and $n^{1/2}\bar{\ell}_{\alpha,\mathbf{k},n}(\alpha^*_{F_0}) \to_d N(0, D_{F_0,\mathbf{k}})$, where

$$D_{F_0,\mathbf{k}} = E_{F_0}[\ell_{\alpha,\mathbf{k}}(Z;\alpha^*_{\mathbf{k}})\ell_{\alpha,\mathbf{k}}(Z;\alpha^*_{\mathbf{k}})']. \tag{3.11}$$

[10] A pretest for whether $\ell(\cdot;\alpha^*_{F_0}) = 0$ could be performed. But the two-step procedure may (a) not be uniformly asymptotically valid if the pretest does not use a conservative critical value, and (b) not be powerful because the pretest makes rejection difficult.

[11] The use of higher order expansion to develop more robust asymptotic theory is not new. It has been used in many contexts including, for example, Jun and Pinkse (2012).

The latter implies that $n\bar{\ell}_{\alpha,n}(\alpha^*_{\mathbf{k}})' H^{-1}_{F_0,\mathbf{k}} \bar{\ell}_{\alpha,n}(\alpha^*_{\mathbf{k}})$ is approximately $\sum_{j=1}^{|\mathbf{k}|} \lambda_j \chi^2_j(1)$, where $|\mathbf{k}| = k_1 + k_2$, $\{\chi^2_j(1)\}_{j=1}^{|\mathbf{k}|}$ are independent chi-squares with one degree of freedom and $\{\lambda_j\}_{j=1}^{|\mathbf{k}|}$ are the eigenvalues of $D_{F_0,\mathbf{k}} H^{-1}_{F_0,\mathbf{k}}$. Thus,

$$n\left\{\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}}) - E_{F_0}[\ell(Z;\alpha^*_{F_0})]\right\} \approx n^{1/2} N(0, \omega^2_{F_0,*}) - 2^{-1} \sum_{j=1}^{|\mathbf{k}|} \lambda_j \chi^2_j(1). \tag{3.12}$$

Note that since $E[\chi^2_j(1)] = 1$, we have $E[\sum_{j=1}^{|\mathbf{k}|} \lambda_j \chi^2_j(1)] = \sum_{j=1}^{|\mathbf{k}|} \lambda_j$, which is typically nonzero and can be of comparable scale as $n^{1/2}\omega_{F_0,*}$, the standard deviation of $n\bar{\ell}_n(\alpha^*_{F_0})$. This means that, even when the null hypothesis H_0 holds ($E_{F_0}[\ell(Z;\alpha^*_{F_0})] = 0$), the numerator of the statistic T^V_n may not be centered around zero, causing the naive extension test to be biased. A similar expansion of the denominator unveils that $n\widehat{\omega}_n(\widehat{\alpha}_{\mathbf{k}})^2$ is a biased estimator of $\omega^2_{F_0,*}$ as well, and the dominating term of the bias is coincidentally $2^{-1}\sum_{j=1}^{|\mathbf{k}|} \lambda^2_j$. Thus, the naive extension test not only has a numerator bias that leads it to favor one model over the other when both have equal fit, but also has a denominator bias that tends to make it conservative. The two biases could cancel each other in certain context, but in general do not, and can exacerbate each other when the power against one-sided alternatives is considered.

Our nondegenerate test corrects the two biases by estimating and removing them. Specifically, we construct estimators $\hat{\lambda}_j : j = 1, \ldots, |\mathbf{k}|$ and propose the bias removed statistics:

$$\tilde{\ell}_n = \bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}}) + (2n)^{-1} \sum_{j=1}^{|\mathbf{k}|} \hat{\lambda}_j \text{ and } \tilde{\omega}^2_n = \widehat{\omega}^2_n(\widehat{\alpha}_{\mathbf{k}}) - (2n)^{-1} \sum_{j=1}^{|\mathbf{k}|} \hat{\lambda}^2_j. \tag{3.13}$$

Then the approximation in (3.12) implies that under H_0,

$$n\tilde{\ell}_n \approx n^{1/2} N(0, \omega^2_{F_0,*}) - 2^{-1} \sum_{j=1}^{|\mathbf{k}|} \lambda_j (\chi^2_j - 1). \tag{3.14}$$

Recall that $|\mathbf{k}| \to \infty$ as $n \to \infty$ in semi/non-parametric models, and apply the central limit theorem on the sum of independent mean-zero variables $\chi^2_j - 1 : j = 1, \ldots, |\mathbf{k}|$ to find that the second term is approximately normal as well. We also show that the two terms are asymptotically independent, suggesting that $n\tilde{\ell}_n$ is asymptotically mean-zero normal under H_0. Moreover, $n\tilde{\omega}^2_n$ also consistently estimate the variance of this mean-zero normal limit. As a result, we have

$$T^0_n = \frac{n\tilde{\ell}_n}{n^{1/2}\tilde{\omega}_n} \to_d N(0,1), \text{ as } n \to \infty. \tag{3.15}$$

There is a minor issue with using T_n^0 as our test statistic because $\tilde{\omega}_n^2$ is defined as the difference of two non-negative terms. In finite sample, this difference can be zero or negative even though the probability of that happening approaches zero as $n \to \infty$. To avoid this finite sample irregularity, we recommend a slightly regularized version:

$$T_n = \frac{n\tilde{\ell}_n}{n^{1/2}\widehat{\sigma}_n}, \text{ where } \widehat{\sigma}_n^2 = \max\left\{\tilde{\omega}_n^2, (2n)^{-1}\sum_{i=1}^{|\mathbf{k}|}\hat{\lambda}_j^2\right\}. \tag{3.16}$$

The regularization has no effect on the asymptotic distribution as we show that $(2n)^{-1}\sum_{i=1}^{|\mathbf{k}|}\hat{\lambda}_j^2$ is less than or equal to $\tilde{\omega}_n^2$ asymptotically. Thus, we still have $T_n \to_d N(0,1)$ as $n \to \infty$.

Estimating $\lambda_j : j = 1, \ldots, |\mathbf{k}|$ is straightforward as they are eigenvalues of $D_{F_0,\mathbf{k}}H_{F_0,\mathbf{k}}^{-1}$. It is in fact unnecessary to estimate these eigenvalues individually since it is clear from the discussion above that all we need are the two sums: $\sum_{j=1}^{\mathbf{k}}\lambda_j$ and $\sum_{j=1}^{\mathbf{k}}\lambda_j^2$, which are equal to $tr(D_{F_0,\mathbf{k}}H_{F_0,\mathbf{k}}^{-1})$ and $tr((D_{F_0,\mathbf{k}}H_{F_0,\mathbf{k}}^{-1})^2)$, respectively, by matrix algebra identities. These can be constructed in a plug-in manner once we have estimates $\widehat{D}_n$ and $\widehat{H}_n$ for $D_{F_0,\mathbf{k}}$ and $H_{F_0,\mathbf{k}}$. When $\ell(Z;\cdot)$ is differentiable, we let

$$\widehat{D}_n = n^{-1}\sum_{i=1}^{n}\ell_{\alpha,\mathbf{k}}(Z_i;\widehat{\alpha}_{\mathbf{k}})\ell_{\alpha,\mathbf{k}}(Z_i;\widehat{\alpha}_{\mathbf{k}})' \text{ and } \widehat{H}_n = n^{-1}\sum_{i=1}^{n}\frac{\partial^2\ell(Z_i;\widehat{\alpha}_{\mathbf{k}})}{\partial\beta_{\mathbf{k}}\partial\beta_{\mathbf{k}}'}. \tag{3.17}$$

The score functions $\ell_{\alpha,\mathbf{k}}(Z_i;\widehat{\alpha}_{\mathbf{k}})$ and estimators of the Hessian matrix are available case by case in the literature when differentiability does not hold. For example, suitable choices for the nonparametric quantile regression example are given in Belloni et al. (2011).

The two-sided test and the one-sided test of H_0 in (2.1) of nominal size p ($\in (0,1)$) are, therefore,

$$\varphi_n^{\text{2-sided}}(p) = 1\{|T_n| > z_{1-p/2}\} \text{ and } \varphi_n^{\text{1-sided}}(p) = 1\{T_n > z_{1-p}\} \tag{3.18}$$

respectively. The test does not select a better fitting model when it does not reject the null hypothesis. Such indeterminacy reflects the data fact that the fit of the two models are not statistically significantly different. In practice, if a model must be selected, one needs to analyze other, perhaps nonstatistical, features of the models. Often times the researcher has a preferred model based on features such as dimensionality and interpretability, and can set that one as the benchmark model. The benchmark model is selected when the null of equal fit is not rejected.

We show the uniform asymptotic validity of the above tests in the next section. Specifically, we show that:

$$\lim_{n\to\infty}\inf_{F_0\in\mathcal{F}_0}E_{F_0}[\varphi_n(p)] = \lim_{n\to\infty}\sup_{F_0\in\mathcal{F}_0}E_{F_0}[\varphi_n(p)] = p, \tag{3.19}$$

where $\varphi_n = \varphi_n^{\text{2-sided}}$ or $\varphi_n = \varphi_n^{\text{1-sided}}$, and $\mathcal{F}_0$ is the set of data generating processes (DGPs) that

the null hypothesis and the assumptions (given below) allow, which shows that the tests that we propose are asymptotically size-exact and similar.

4 Uniform Asymptotic Validity

In this section, we establish the uniform asymptotic validity and the local power of our test under high-level assumptions. These assumptions are verified in a nonparametric mean-regression example and in a quantile-regression example in Supplemental Appendices C and D respectively.

We begin by stating the regularity conditions on the DGP space $\mathcal{F}$ and null DGP space $\mathcal{F}_0$. In the assumptions below, $\{\xi_{\mathbf{k}}\}_{\mathbf{k}}$ is an array of non-decreasing positive numbers which may diverge with $|\mathbf{k}| = k_1 + k_2$, and may not depend on F_0.

Assumption 4.1 *The set $\mathcal{F}$ is the set of F_0's such that*

(a) $\{Z_i\}_{i\geq 1}$ *are i.i.d. draws from* F_0*;*

(b) *for every* $\mathbf{k}$, $E_{F_0}[\ell(Z;\alpha(\beta_{\mathbf{k}}))]$ *is twice-differentiable in* $\beta_{\mathbf{k}}$*;*

(c) *the sieve approximator* $\alpha^*_{\mathbf{k}}$ *satisfies* $E_{F_0}\left[\ell_{\alpha,\mathbf{k}}(Z;\alpha^*_{\mathbf{k}})\right] = \mathbf{0}_{|\mathbf{k}|}$ *for every* $\mathbf{k}$*;*

(d) $E_{F_0}\left[\ell(Z;\alpha^*_{F_0})^2\right] < C$*, and for every* $\mathbf{k}$, $E_{F_0}\left[\|\ell_{\alpha,\mathbf{k}}(Z;\alpha^*_{\mathbf{k}})\|^4\right] \leq C\xi_{\mathbf{k}}\,|\mathbf{k}|$*;*

(e) $E_{F_0}\left[\left|(\ell(Z;\alpha^*_{F_0}) - E_{F_0}(\ell(Z;\alpha^*_{F_0})))/\omega_{F_0,*}\right|^4\right] < C$ *whenever* $\omega^2_{F_0,*} \equiv Var_{F_0}[\ell(Z;\alpha^*_{F_0})] > 0$*;*

(f) *for* $j = 1, 2$, $-C \leq \rho_{\min}(H_{F_0,k_j}) \leq \rho_{\max}(H_{F_0,k_j}) \leq -C^{-1}$ *and* $\rho_{\max}(D_{F_0,\mathbf{k}}) \leq C$ *for all* $\mathbf{k}$*.*

Assumption 4.2 $\mathcal{F}_0 = \left\{F_0 \in \mathcal{F} : E_{F_0}\left[\ell(Z;\alpha^*_{F_0})\right] = 0\right\}$.

Assumption 4.1(b) ensures that the matrix $H_{F_0,\mathbf{k}}$ in (3.10) is well defined. Assumption 4.1(c) generally follows from the first order optimality condition of $\alpha^*_{\mathbf{k}}$. Let $\lambda_{F_0,1}, \ldots, \lambda_{F_0,|\mathbf{k}|}$ denote the $|\mathbf{k}|$ eigenvalues of $D^{1/2}_{F_0,\mathbf{k}} H^{-1}_{F_0,\mathbf{k}} D^{1/2}_{F_0,\mathbf{k}}$, and let

$$\sigma^2_{F_0,n} \equiv \omega^2_{F_0,*} + (2n^2)^{-1}(n-1)\omega^2_{F_0,U,\mathbf{k}} \tag{4.1}$$

where $\omega^2_{F_0,U,\mathbf{k}} \equiv \sum_{j=1}^{|\mathbf{k}|} \lambda^2_{F_0,j} \equiv tr((D_{F_0,\mathbf{k}} H^{-1}_{F_0,\mathbf{k}})^2)$. Assumptions 4.1(d) and (f) together ensure that $\omega^2_{F_0,*}$, $D_{F_0,\mathbf{k}}$, $\omega^2_{F_0,U,\mathbf{k}}$, and $\sigma^2_{F_0,n}$ are well defined. The array $\xi_{\mathbf{k}}$ depends on the approximating function used. For example, it is the order of k_j^2 on the jth direction if power series is used for model j, and it is the order of k_j if Fourier or spline series is used. Assumption 4.1(e) implies the Linderberg condition on the pseudo-density ratio.

The definition of the supremum (infimum) operator implies that, to show the uniformity results (3.19), it is sufficient to consider all sequences of DGPs $\{F_n\}_{n\geq 1}$ in $\mathcal{F}_0$. Moreover, to study the local power properties, we need to consider sequences of DGPs $\{F_n\}_{n\geq 1}$ in $\mathcal{F}\backslash\mathcal{F}_0$. In general, we consider sequences $\{F_n\}_{n\geq 1}$ in $\mathcal{F}$. For any $F_n \in \mathcal{F}$, we let $\alpha^*_{j,n}$ abbreviate $\alpha^*_{F_n,j}$, and let α^*_n abbreviate $(\alpha^*_{1,n}, \alpha^*_{2,n})$. Let $\bar{\ell}_{\alpha,n}(\alpha) = n^{-1}\sum_{i=1}^n \ell_{\alpha,\mathbf{k}}(Z_i;\alpha)$ for any $\alpha \in \mathcal{A}$.

Assumption 4.3 *Under any sequence of DGP's* $\{F_n\}_{n\geq 1}$ *such that* $F_n \in \mathcal{F}$ *for all* n*, we have*
(a) $\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}}) = \bar{\ell}_n(\alpha_n^*) - 2^{-1}\overline{\ell}_{\alpha,n}(\alpha_{\mathbf{k}}^*)' H_{F_n,\mathbf{k}}^{-1} \overline{\ell}_{\alpha,n}(\alpha_{\mathbf{k}}^*) + o_p(n^{-1/2}\sigma_{F_n,n})$;
(b) $(n\sigma_{F_n,n}^2)^{-1} = o(1)$ *and* $|\mathbf{k}|\xi_{\mathbf{k}}(n^2\sigma_{F_n,n}^2)^{-1} = o(1)$.

Assumption 4.3(a) is a second order expansion of $\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}})$ around α_n^*. We verify this assumption in the nonparametric mean regression example (Supplemental Appendix C) and the nonparametric quantile regression example (Supplemental Appendix D). With the formula of this expansion, we can add more details to the heuristic discussion in Section 3. The variance of the leading term, $n^{-1}\omega_{F_n,*}^2$, in the expansion comes from estimating the expectation, and the variance of the second term, approximately $2^{-1}n^{-2}\omega_{F_n,U,\mathbf{k}}^2$, comes from estimating α_n^*. The quantity $\omega_{F_n,*}^2$ can be either zero or positive in the overlapping nonnested case. Indeed, it can converge to zero at any rate in that case. On the other hand, the quantity $\omega_{F_n,U,\mathbf{k}}^2$ typically is nonzero.[12] The relative magnitude of the two terms is proportional to $\frac{n\omega_{F_n,*}^2}{\omega_{F_n,U,\mathbf{k}}^2}$, which can be zero or positive. It is such ambiguity of the relative asymptotic order of the two expansion terms that makes a uniformly valid test difficult to construct.[13]

Assumption 4.3(b) is an important condition for the uniform asymptotic validity of our test. The first part of it ensures that the approximation residual in Assumption 4.3 (a) diminishes at a fast enough rate as the sample size grows. The second part of the assumption allows us to apply a U-statistic central limit theorem to the quadratic term $2^{-1}\overline{\ell}_{\alpha,n}(\alpha_{\mathbf{k}}^*)' H_{F_n,\mathbf{k}}^{-1} \overline{\ell}_{\alpha,n}(\alpha_{\mathbf{k}}^*)$. To understand this assumption, note that $\sigma_{F_n,n}^2 = \omega_{F_n,*}^2 + (2n^2)^{-1}(n-1)\omega_{F_n,U,\mathbf{k}}^2$. If $\omega_{F_n,*}^2$ is bounded below by a positive constant (as is typical for strictly nonnested models), Assumption 4.3(b) is satisfied as long as $|\mathbf{k}|\xi_{\mathbf{k}}n^{-2} = o(1)$ as $n \to \infty$, which simply requires that the number of sieve terms not to grow too fast. Otherwise, Assumption 4.3(b) imposes restriction on the U-statistic variance $\omega_{F_n,U,\mathbf{k}}^2 \equiv tr\left((H_{F_n,\mathbf{k}}^{-1} D_{F_n,\mathbf{k}})^2\right)$. Specifically, it requires, as $n \to \infty$, that

$$\omega_{F_n,U,\mathbf{k}}^2 \to \infty \text{ and } |\mathbf{k}|\xi_{\mathbf{k}}(n\omega_{F_n,U,\mathbf{k}}^2)^{-1} = o(1). \tag{4.2}$$

This is satisfied if $|\mathbf{k}|$ grows with n and there are not too many near zero eigenvalues for the matrix $H_{F_n,\mathbf{k}}^{-1} D_{F_n,\mathbf{k}}$. Both can be assessed in practice because $\mathbf{k}$ is user-chosen and $H_{F_n,\mathbf{k}}^{-1} D_{F_n,\mathbf{k}}$ can be consistently estimated. Moreover, the requirement that $|\mathbf{k}|$ grows with n is natural and

[12]For example, consider $\mathcal{M}_1$: $Y = X_1'\beta_1 + X_2'\beta_2 + u$ and $\mathcal{M}_2$: $Y = X_1'\beta_1 + u$. Suppose that $X = (X_1', X_2')'$ is uncorrelated with u and $E_{F_0}[XX'] = I_{|\mathbf{k}|}$ for simplicity. The null hypothesis H_0 is equivalent to $\beta_2 = 0$ and there is $\ell(Z;\alpha_n^*) = 0$ under H_0 as a result. Yet, $2^{-1}\overline{\ell}_{\alpha,n}(\alpha_n^*)' H_{F_n,\mathbf{k}}^{-1} \overline{\ell}_{\alpha,n}(\alpha_n^*) = 2^{-1}n^{-2}\sum_{i=1}^n \sum_{j=1}^n u_i u_j X_{2,i}' X_{2,j}$ which is clearly not degenerate. See Hong and White (1995) for more sophisticated examples.

[13]Ambiguity of this type also arises in the analysis of weak instruments and weak identification, where the common techniques include pretesting with conservative critical value, Anderson-Rubin type robust procedures, and conditional likelihood inference. The first two in general do not yield asymptotically similar tests, indicating power loss under some data generating processes, while the last one is not a general technique that can be applied here.

necessary in the literature of series estimation of semi/nonparametric models.[14]

Under the above assumptions, the following intermediate result holds.

Theorem 4.1 *Suppose that Assumptions* 4.1 *and* 4.3 *hold. Then under any sequence* $\{F_n\}_{n\geq 1}$ *such that* $F_n \in \mathcal{F}$ *for all* n*, we have*

$$\frac{n(\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}}) - E_{F_n}[\ell(Z;\alpha_n^*)]) + (1/2)tr(\widehat{D}_n(\alpha_{\mathbf{k}}^*)H_{F_n,\mathbf{k}}^{-1})}{n^{1/2}\sigma_{F_n,n}} \to_d N(0,1), \tag{4.3}$$

where $\widehat{D}_n(\alpha_{\mathbf{k}}^*) = n^{-1}\sum_{i=1}^n \ell_{\alpha,\mathbf{k}}(Z_i;\alpha_{\mathbf{k}}^*)\ell_{\alpha,\mathbf{k}}(Z_i;\alpha_{\mathbf{k}}^*)'$.

Remark 1 *Note that Theorem 4.1 applies whether or not* $F_n \in \mathcal{F}_0$*. In the case that* $F_n \in \mathcal{F}_0$ *for all* n*, it again covers two special sub-cases:* (i) *The statistic* $n^{1/2}\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}})$ *is non-degenerate (*$F_n = F$ *for some* F *and for all* n*, and* $\omega_{F,*}^2 > 0$*);* (ii) *the statistic* $n^{1/2}\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}})$ *is degenerate (*$F_n = F$ *for some* F *and for all* n*, and* $\omega_{F,*}^2 = 0$*). More importantly, it allows* $\omega_{F_n,*}^2$ *to converge to zero at all rates, and thus covers all types of DGP sequences in the overlapping nonnested case.*

When $\omega_{F_n,*}^2$ *converges to zero at an equal or faster rate than* n^{-1} *or is exactly zero, the asymptotic normality in (4.3) is achieved by the central limit theorem of U-statistic which requires that* $|\mathbf{k}|$ *grows with* n*. The normal approximation of the U-statistic is widely used in the literature of model specification test. See e.g., Hall (1984), Hong and White (1995), Horowitz and Härdle (1994), Fan and Li (1996), Aït-Sahalia et al. (2001) and Donald et al. (2003). Theorem 4.1 shares similar features with the results in these papers, in that they also require the number of approximating functions to diverge with* n *or the bandwidth of kernel functions to go to zero with* n*.*

In order to use the intermediate result in Theorem 4.1, we need to construct consistent estimators of $\widehat{D}_n(\alpha_{\mathbf{k}}^*)$, $H_{F_n,\mathbf{k}}$, and $\sigma_{F_n,n}^2$. The estimators that we consider are respectively the $\widehat{D}_n$, the $\widehat{H}_n$, and the $\widehat{\sigma}_n^2$ defined in the previous section. Assumption 4.4 below ensures their consistency. In this assumption, $\delta_n = \min\left\{n^{1/2}\sigma_{F_n,n}|\mathbf{k}|^{-1}, 1\right\}$, and $\ell_F(\alpha) = E_F[\ell(Z;\alpha)]$ for all $F \in \mathcal{F}$ and $\alpha \in \mathcal{A}$.

Assumption 4.4 *Under any sequence of DGP's* $\{F_n\}_{n\geq 1}$ *with* $F_n \in \mathcal{F}$ *for all* n*, we have:*

(a) $\|\widehat{H}_n - H_{F_n,\mathbf{k}}\| = o_p(\delta_n)$, $\|\widehat{D}_n - \widehat{D}_n(\alpha_{\mathbf{k}}^*)\| = o_p(\delta_n)$ *and* $\|\widehat{D}_n(\alpha_{\mathbf{k}}^*) - D_{F_n,\mathbf{k}}\| = o_p(\delta_n)$*;*

(b) $n^{-1}\sum_{i=1}^n |\ell(Z_i,\widehat{\alpha}_{\mathbf{k}}) - \ell(Z_i,\alpha_n^*)|^2 = \bar{\ell}_{\alpha,n}(\alpha_{\mathbf{k}}^*)'(H_{F_n,\mathbf{k}}^{-1}D_{F_n,\mathbf{k}}H_{F_n,\mathbf{k}}^{-1})\bar{\ell}_{\alpha,n}(\alpha_{\mathbf{k}}^*) + o_p(\sigma_{F_n,n}^2)$*;*

(c) $n^{-1}\sum_{i=1}^n(\ell(Z_i,\alpha_n^*) - \ell_{F_n}(\alpha_n^*))\,[\ell(Z_i,\widehat{\alpha}_{\mathbf{k}}) - \ell(Z_i,\alpha_n^*)] = o_p(\sigma_{F_n,n}^2)$*;*

(d) $|\mathbf{k}|\,n^{-1} = o(1)$*.*

[14]The asymptotic theory established in this paper also provides a good approximation for the comparison of parametric models with fixed but large $|\mathbf{k}|$. Simulation results in Supplemental Appendix F show that our test works well even when $|\mathbf{k}|$ is only 7.

Conditions in Assumption 4.4 are verified in the nonparametric mean-regression example in Supplemental Appendix C. Under this assumption, we can easily show that the large sample bias of $n\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n})$ can be estimated up to the appropriate rate:

Lemma 4.1 *Suppose that Assumptions* 4.1(c) and (e)-(g), and 4.4(a) *hold. Then under any sequence* $\{F_n\}_{n\geq 1}$ *such that* $F_n \in \mathcal{F}$ *for all* n*, we have*

$$tr(\widehat{D}_n\widehat{H}_n^{-1}) - tr(\widehat{D}_n(\alpha_{\mathbf{k}}^*)H_{F_n,\mathbf{k}}^{-1}) = o_p(n^{1/2}\sigma_{F_n,n}).$$

Next, we derive the convergence of $\widehat{\sigma}_n^2$. First, we show the convergence of $\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k}})$ in the following lemma.

Lemma 4.2 *Suppose that Assumptions* 4.1, 4.3 *and* 4.4 *hold. Then under any sequence* $\{F_n\}_{n\geq 1}$ *such that* $F_n \in \mathcal{F}$ *for all* n*, we have*

$$\left|\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k}}) - \left(\omega_{F_n,*}^2 + n^{-1}\omega_{F_n,U,\mathbf{k}}^2\right)\right| = o_p(\sigma_{F_n,n}^2).$$

Remark 2 *Note that* $\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k}})$ *may be viewed as a sample-analogue estimator of* $\omega_{F_n,*}^2$*. Lemma* 4.2 *shows that, in general,* $\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k}})$ *over-estimates* $\omega_{F_n,*}^2$*. In fact, it even over-estimates the overall asymptotic variance of the size-corrected quasi-likelihood ratio statistic:* $\sigma_{F_n,n}^2$*, by* $(2n^2)^{-1}(n+1)\omega_{F_n,U,\mathbf{k}}^2$*. The upward bias is due to the estimation error in* $\widehat{\alpha}_{\mathbf{k}}$*.*

Lemma 4.2 suggests that $\sigma_{F_n,n}^2$ can be consistently estimated by estimating and then removing the large-sample bias $(2n^2)^{-1}(n+1)\omega_{F_n,U,\mathbf{k}}^2$ from $\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k}})$. This motivates the estimator $\widehat{\sigma}_n^2$ defined in the previous section. In the definition of $\widehat{\sigma}_n^2$, $tr((\widehat{D}_n\widehat{H}_n^{-1})^2)$ is used to estimate $\omega_{F_n,U,\mathbf{k}}^2$. The lemma below shows that this estimator of $\omega_{F_n,U,\mathbf{k}}^2$ is consistent in an appropriate sense, and so is the resulting bias-removed estimator of $\sigma_{F_n,n}^2$.

Lemma 4.3 *Suppose that Assumptions* 4.1, 4.3 *and* 4.4 *hold. Then under any sequence* $\{F_n\}_{n\geq 1}$ *such that* $F_n \in \mathcal{F}$ *for all* n*, we have*

(a) $tr((\widehat{D}_n\widehat{H}_n^{-1})^2) - \omega_{F_n,U,\mathbf{k}}^2 = o_p(n\sigma_{F_n,n}^2)$*, and*

(b) $\tilde{\omega}_n^2 - \sigma_{F_n,n}^2 = o_p(\sigma_{F_n,n}^2)$*, where* $\tilde{\omega}_n^2 = \widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k}}) - (2n)^{-1}tr((\widehat{D}_n\widehat{H}_n^{-1})^2)$ *as defined in (*3.13*).*

Lemma 4.3 is used to show the consistency of $\widehat{\sigma}_n^2$: $\widehat{\sigma}_n^2 - \sigma_{F_n,n}^2 = o_p(\sigma_{F_n,n}^2)$. This along with Theorem 4.1 and Lemmas 4.1–4.2 immediately leads to the uniform asymptotic size control and the asymptotic similarity results in (3.19). These results also immediately lead to a local power formula because the assumptions used for them do not require $F_n \in \mathcal{F}_0$. These are summarized in the theorem below.

Theorem 4.2 *Suppose that Assumptions* 4.1-4.4 *hold. Then:*
(a) (3.19) *holds for* $\varphi_n = \varphi_n^{\text{2-sided}}$ *and* $\varphi_n = \varphi_n^{\text{1-sided}}$.
(b) *Under any sequence* $F_n \in \mathcal{F}$ *such that* $F_n \to F_0$ *for some* $F_0 \in \mathcal{F}_0$ *in the Kolmogorov-Smirnov distance, and that* $n^{1/2} E_{F_n}[\ell(Z;\alpha_n^*)] / \sigma_{F_n,n} \to c$ *for some constant* $c \in R$, *we have*

$$\lim_{n\to\infty} E_{F_n}\left[\varphi_n^{\text{2-sided}}(p)\right] = 2 - \Phi(z_{1-p/2} - c) - \Phi(z_{1-p/2} + c), \text{ and}$$
$$\lim_{n\to\infty} E_{F_n}\left[\varphi_n^{\text{1-sided}}(p)\right] = 1 - \Phi(z_{1-p} - c),$$

where $\Phi(\cdot)$ *is the CDF of the standard normal distribution.*

Remark 3 *Note that* $\sigma_{F_n,n} = O(1)$, *and it can be* $o(1)$ *when* $\omega^2_{F_n,*} \to 0$. *Thus, part (b) of the theorem implies that the test has nontrivial power against all local alternatives with* $E_{F_n}[\ell(Z;\alpha_n^*)]$ *converging to 0 at the rate* $n^{1/2}$, *and against alternatives with* $E_{F_n}[\ell(Z;\alpha_n^*)]$ *converging to 0 at a rate faster than* $n^{1/2}$ *if* $\omega^2_{F_n,*} \to 0$. *Such power property is not shared by a pre-test based model selection test like that in Shi (2015a), or a model selection test that uses added noise to augment the variance either through sample splitting or other means.*

Remark 4 *As we have discussed, Shi (2015b) proposes a nondegenerate test for the parametric case. Her test statistic, if directly applied to the sieve approximation of the semi/nonparametric models, takes the following form*

$$T_n^{para}(c) = \frac{n\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}}) + 2^{-1} tr(\widehat{D}_n \widehat{H}_n^{-1})}{n^{1/2}\left(\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k}}) + cn^{-1} tr((\widehat{D}_n \widehat{H}_n^{-1})^2)\right)^{1/2}}, \tag{4.4}$$

where $c \geq 0$ *is a tuning parameter. Compared with* $T_n^{para}(c)$, *our test statistic* T_n *has the same numerator but a different denominator. By Lemma 4.3(b),* $\frac{\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k}})}{\sigma^2_{F_n,n}} - \frac{(2n)^{-1} tr((\widehat{D}_n \widehat{H}_n^{-1})^2)}{\sigma^2_{F_n,n}} \to_p 1$, *which implies that* $\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k}}) > (2n)^{-1} tr((\widehat{D}_n \widehat{H}_n^{-1})^2)$ *with probability approaching one. This and the definition of* $\widehat{\sigma}_n^2$ *together imply that* $\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k}}) \geq \widehat{\sigma}_n^2$ *with probability approaching one, which in turn implies that* $|T_n^{para}(c)| \leq |T_n|$ *with probability approaching one for any* $c \geq 0$. *On the other hand, the critical value of the test proposed in Shi (2015b) by construction is not smaller than the critical value of our test. Therefore the asymptotic theory established in this section automatically justifies the test proposed in Shi (2015b) in terms of asymptotic size control when applied to the semi/nonparametric models. However, when* $|\mathbf{k}|$ *is large, there are a large number of nuisance parameters (which are not consistently estimable) for Shi's (2015b) approach to consider, which makes it difficult to use. In contrast, our test is much easier to use, also has asymptotic size control, and has better power in the semi/nonparametric setting, where the better power is implied by its bigger test statistic and smaller critical value. Moreover, the asymptotic standard normal*

distribution of our test statistic T_n also makes the post model selection inference easy in practice as we discuss in later sections.

5 Example: Semi/Nonparametric Mean-Regression

In this section we illustrate the construction of our test using the nonparametric mean-regression example. We verify the high-level assumptions in this example in Supplemental Appendix C. Another illustrating example—quantile-regression—is given in Supplemental Appendix D, where we also verify the high level assumptions.

For $j = 1, 2$, model j is to maximize $E_{F_0}[-2^{-1} |Y - \alpha_j(X_j)|^2]$ over $\alpha_j \in \mathcal{A}_j$, where $\alpha_j(x)$ is a possibly infinite dimensional parameter, $\mathcal{A}_j$ is its parameter space, and F_0 denotes the joint distribution of $Z \equiv (Y, X_1, X_2)$. The regressors X_1 and X_2 of the two models may be nested, overlapping, or strictly non-nested sets of variables. Even when the regressors are strictly nonnested sets of variables (i.e., there are no common regressors across the two regressions), the two regression models are still overlapping according to the definitions in Section 2.2 because it is possible that $\alpha_1(X_1) = \alpha_2(X_2) = Constant$.

The model studied in this section covers a richer class of models than it looks. Depending on what one sets $\mathcal{A}_j$ to be, it can represent a fully nonparametric mean-regression model, a partial linear model, a separable model, or a parametric linear model. See below for an example. We do not require that there exists an $\alpha_j \in \mathcal{A}_j$ such that $\alpha_j(X_j) = E_{F_0}[Y|X_j]$ a.s.

The sieve approximating functions for this case have to do with the structure of $\mathcal{A}_j$. For example, suppose that we have a partial linear model $\alpha_j(X_j) = \beta_1 X_{j,1} + g(X_{j,2})$. Then, we should let $P_{k_j}(X_j) = \left[p_{j,1}(X_j), \ldots, p_{j,k_j}(X_j)\right]'$ such that $p_{j,1}(X_j) = X_{j,1}$ and the rest of the sequence of $p_{j,\ell}(X_j)$'s be an appropriate sieve approximation of $g(X_{j,2})$, such as splines or polynomials on $X_{j,2}$.

The sieve M-estimator is simply the sieve least squares estimator:

$$\widehat{\alpha}_{k_j}(\cdot) = P_{k_j}(\cdot)'\widehat{\beta}_{k_j} \text{ with } \widehat{\beta}_{k_j} = (\mathbf{P}'_{k_j,n}\mathbf{P}_{k_j,n})^{-1}\mathbf{P}'_{k_j,n}\mathbf{Y}_n, \tag{5.1}$$

where $\mathbf{P}_{k_j,n} = \left[P_{k_j}(X_{j,1}), \ldots, P_{k_j}(X_{j,n})\right]'$ for $j = 1, 2$, and $\mathbf{Y}_n = (Y_1, \ldots, Y_n)'$. The link function is

$$\ell(Z; \alpha) = 2^{-1}|Y - \alpha_2(X_2)|^2 - 2^{-1}|Y - \alpha_1(X_1)|^2. \tag{5.2}$$

Using the above two displays, the pseudo-likelihood ratio and the standard error statistics can be constructed easily following (3.6).

The pseudo true parameter $\alpha_j^*(\cdot)$ is defined as $\alpha_j^* = \arg\max_{\alpha_j \in \mathcal{A}_j} E_{F_0}[-2^{-1} |Y - \alpha_j(X_j)|^2]$, which depends on the functional form restrictions imposed on the parameter space $\mathcal{A}_j$. If there is no functional form restriction, then $\alpha_j^*(X_j) = E_{F_0}[Y|X_j]$. If an additive form is imposed,

i.e., $\alpha_j(X_j) = g(X_{j,1}) + \ldots + g(X_{j,q})$ for some finite q, the pseudo true parameter exists and is unique under general conditions (see Condition 1 and Lemma 1 in Stone (1985)). When a partially linear form is imposed, i.e., $\alpha_j(X_j) = X'_{j,1}\beta_1 + g(X_{j,2})$, then the pseudo true parameter $\alpha_j^*(X_j) = X'_{j,1}\beta_1^* + g^*(X_{j,2})$ where

$$\beta_1^* = \left(E_{F_0}\left[X_{j,1}^* X_{j,1}^{*\prime}\right]\right)^{-1} E_{F_0}\left[X_{j,1}^* Y^*\right] \text{ and } g^*(X_{j,2}) = E_{F_0}\left[Y - X'_{j,1}\beta_1^* | X_{j,2}\right] \tag{5.3}$$

where $X_{j,1}^* = X_{j,1} - E_{F_0}\left[X_{j,1}|X_{j,2}\right]$ and $Y^* = Y - E_{F_0}\left[Y|X_{j,2}\right]$.

Let $u_{k_j} = Y - \alpha_{k_j}^*(X_j)$, where $\alpha_{k_j}^*(\cdot) = P_{k_j}(\cdot)'\beta_{k_j,F_0}^*$ and

$$\beta_{k_j,F_0}^* = \underset{\beta_{k_j}\in R^{k_j}}{\arg\min} E_{F_0}\left[\left|Y - P_{k_j}(X_j)'\beta_{k_j}\right|^2\right]. \tag{5.4}$$

By the first order optimality condition for $u_{k_j} = Y - P_{k_j}(X_j)'\beta_{k_j,F_0}^*$, we have $E_{F_0}[u_{k_j}P_{k_j}(X_j)] = \mathbf{0}_{k_j\times 1}$. With the sieve approximation in (3.1), $\ell(Z;\alpha(\beta_{\mathbf{k}}))$ is differentiable in $\beta_{\mathbf{k}}$. Thus, the score function can be obtained by the chain rule:

$$\ell_{\alpha,\mathbf{k}}(Z;\alpha) = \left((Y - \alpha_1(X_1))P_{k_1}(X_1)', -(Y - \alpha_2(X_2))P_{k_2}(X_2)'\right)'. \tag{5.5}$$

Then, the expectation of the outer product of the score function evaluated at $\alpha_{\mathbf{k}}^*$ is

$$D_{F_0,\mathbf{k}} = \begin{pmatrix} E_{F_0}[u_{k_1}^2 P_{k_1}(X_1)P_{k_1}(X_1)'] & -E_{F_0}[u_{k_1}u_{k_2}P_{k_1}(X_1)P_{k_2}(X_2)'] \\ -E_{F_0}[u_{k_1}u_{k_2}P_{k_2}(X_2)P_{k_1}(X_1)'] & E_{F_0}[u_{k_2}^2 P_{k_2}(X_2)P_{k_2}(X_2)'] \end{pmatrix}, \tag{5.6}$$

and the population Hessian matrix is:

$$H_{F_0,\mathbf{k}} = diag\left(-E_{F_0}[P_{k_1}(X_1)P_{k_1}(X_1)'],\ E_{F_0}[P_{k_2}(X_2)P_{k_2}(X_2)']\right). \tag{5.7}$$

It is natural to use the plug-in estimators of $D_{F_0,\mathbf{k}}$ and $H_{F_0,\mathbf{k}}$:

$$\widehat{D}_{n,\mathbf{k}} = \begin{pmatrix} n^{-1}\sum_{i=1}^n \widehat{u}_{1,i}^2 P_{k_1}\left(X_{1,i}\right) P_{k_1}(X_{1,i})' & -n^{-1}\sum_{i=1}^n \widehat{u}_{1,i}\widehat{u}_{2,i} P_{k_1}\left(X_{1,i}\right) P_{k_2}(X_{2,i})' \\ -n^{-1}\sum_{i=1}^n \widehat{u}_{1,i}\widehat{u}_{2,i} P_{k_2}\left(X_{2,i}\right) P_{k_1}(X_{1,i})' & n^{-1}\sum_{i=1}^n \widehat{u}_{2,i}^2 P_{k_2}\left(X_{2,i}\right) P_{k_2}\left(X_{2,i}\right)' \end{pmatrix}, \tag{5.8}$$

where the residual $\widehat{u}_{j,i} = Y_i - \widehat{\alpha}_{k_j}(X_{j,i})$; and

$$\widehat{H}_{n,\mathbf{k}} = diag\left(-n^{-1}\sum_{i=1}^n P_{k_1}\left(X_{1,i}\right) P_{k_1}(X_{1,i})',\ n^{-1}\sum_{i=1}^n P_{k_2}\left(X_{2,i}\right) P_{k_2}\left(X_{2,i}\right)'\right). \tag{5.9}$$

Finally, the test statistic may be constructed easily using the above quantities following (3.16).

6 Uniformly Valid Post Selection Test Inference

Up to this point, we have focused on how to properly conduct model selection that takes into account sample noise. Sometimes, model selection is the sole purpose of a research project (e.g., Coate and Conlin (2004) and Gandhi and Serrano-Padial (2015)). But, sometimes, one is also interested in the model parameters that are estimated using the same data set on which the model selection test is conducted. Leeb and Pötscher (2005) show the size-distortion of naive post-model-selection (PMS) inference that does not account for the randomness of model selection. Uniformly valid post model selection test inference procedures for possibly misspecified semi/nonparametric models have not been developed in the literature.

The QLR model selection test framework treats the parameters in the two models as separate parameters in the sense that there is no across-model restrictions. In practice, while some parameters of a model may only have meaningful interpretation in its own model environment, it is also possible that a parameter from one model and a parameter from the other model represent the same economic parameter of interest. Thus, we treat these two different scenarios separately when considering post model selection test inference.

In the first scenario, the parameter of interest is only well-defined in model $\mathcal{M}_j$ ($j = 1$ or 2), and the researcher is interested in it only when $\mathcal{M}_j$ is selected by the model selection test. In this scenario, we would like to make the inference conditional on the event that $\mathcal{M}_1$ is selected. Leeb and Pötscher (2006) pointed out that in general it is impossible to approximate the conditional distribution of the parameter estimator given that the model is selected. Instead of studying the conditional distribution, we take a different route, and construct confidence interval for the parameter using a conditionally asymptotically pivotal statistic. We devote subsection 6.2 to this approach.

In the second scenario, the parameter of interest, θ, is well-defined in both models: it equals $\psi_1(\alpha_1)$ in model $\mathcal{M}_1$ and equals $\psi_2(\alpha_2)$ in model $\mathcal{M}_2$ for two known functionals $\psi_1 : \mathcal{A}_1 \to R$ and $\psi_2 : \mathcal{A}_2 \to R$. Its (pseudo)-true value is determined by the better fitting model:

$$\theta^* = \psi_1(\alpha_1^*)1(f(\mathcal{M}_1, F_0) \geq f(\mathcal{M}_2, F_0)) + \psi_2(\alpha_2^*)1(f(\mathcal{M}_1, F_0) < f(\mathcal{M}_2, F_0)). \tag{6.1}$$

For example, if the competing models are two regression models, θ^* could be the expected point prediction from the better fitting model. We devote subsection 6.3 below to this problem.

To prepare for subsections 6.2 and 6.3, we let $\psi_1(\alpha_1^*)$ and $\psi_2(\alpha_2^*)$ be estimated by the plug-in estimators $\psi_1(\widehat{\alpha}_{k_1})$ and $\psi_2(\widehat{\alpha}_{k_2})$ respectively. Both subsections 6.2 and 6.3 rely on the joint normal limiting distribution of $(\psi_1(\widehat{\alpha}_{k_1}), \psi_2(\widehat{\alpha}_{k_2}), \bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}}))'$ (after proper re-centering and rescaling), which we derive in the next subsection.

6.1 Preliminaries

We first introduce some notation. Let $\ell_{\alpha,k_1}(Z;\alpha_1)$ denote the sub-vector of the first k_1 coordinates of $\ell_{\alpha,\mathbf{k}}(Z;\alpha)$, and let $\ell_{\alpha,k_2}(Z;\alpha_2)$ denote minus the sub-vector of the last k_2 coordinates of $\ell_{\alpha,\mathbf{k}}(Z;\alpha)$. Let $D_{F_0,k_j} = E_{F_0}[\ell_{\alpha,k_j}(Z;\alpha^*_{F_0,j})\ell_{\alpha,k_j}(Z;\alpha^*_{F_0,j})']$ for $j = 1, 2$. Also define

$$\psi_{\alpha,k_j}(\alpha_j) = \frac{\partial \psi_j(\alpha_j(\beta_{k_j}))}{\partial \beta_{k_j}} \text{ and } v^*_{\psi,k_j} = \left(\psi_{\alpha,k_j}(\alpha^*_{k_j})' H^{-1}_{F_0,k_j} D_{F_0,k_j} H^{-1}_{F_0,k_j} \psi_{\alpha,k_j}(\alpha^*_{k_j})\right)^{1/2}, \tag{6.2}$$

where v^*_{ψ,k_j} is the well-established formula for the asymptotic standard deviation of functionals of sieve-M estimator.

Let $\widehat{v}^*_{\psi,k_j}$ denote the estimator of v^*_{ψ,k_j} which is defined as

$$\widehat{v}^{*2}_{\psi,k_j} = \psi_{\alpha,k_j}(\widehat{\alpha}_{k_j})' \widehat{H}^{-1}_{k_j,n} \widehat{D}_{k_j,n} \widehat{H}^{-1}_{k_j,n} \psi_{\alpha,k_j}(\widehat{\alpha}_{k_j})$$

where $\widehat{H}_{k_j,n}$ and $\widehat{D}_{k_j,n}$ are the leading $k_j \times k_j$ submatrices of $\widehat{H}_n$ and $\widehat{D}_n$ respectively for $j = 1$, and the last $k_j \times k_j$ submatrices of $-\widehat{H}_n$ and $\widehat{D}_n$ respectively for $j = 2$.

We shall derive the asymptotic distribution of

$$\widehat{G}_{n,F_n} \equiv \begin{pmatrix} \frac{n\left[\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k}}) - E_{F_n}(\ell(Z;\alpha^*_n))\right] + (1/2)tr(\widehat{D}_n \widehat{H}^{-1}_n)}{n^{1/2}\widehat{\sigma}_n} \\ n^{1/2}\left[\psi_1(\widehat{\alpha}_{k_1}) - \psi_1(\alpha^*_{1,n})\right](\widehat{v}^*_{\psi,k_1})^{-1} \\ n^{1/2}\left[\psi_2(\widehat{\alpha}_{k_2}) - \psi_2(\alpha^*_{2,n})\right](\widehat{v}^*_{\psi,k_2})^{-1} \end{pmatrix}. \tag{6.3}$$

For this purpose, define the correlation coefficients

$$\begin{aligned} \rho_{0j,F_0} &= \psi_{\alpha,k_j}(\alpha^*_{k_j})' H^{-1}_{F_0,k_j} E_{F_0}\left[\ell_{\alpha,k_j}(Z;\alpha^*_{k_j})\ell(Z;\alpha^*_n)\right](v^*_{\psi,k_j}\sigma_{F_0,n})^{-1} \text{ for } j = 1, 2, \\ \rho_{12,F_0} &= \psi_{\alpha,k_1}(\alpha^*_{k_1})' H^{-1}_{F_0,k_1} D_{F_0,k_1,k_2} H^{-1}_{F_n,k_2} \psi_{\alpha,k_2}(\alpha^*_{k_2})(v^*_{\psi,k_1} v^*_{\psi,k_2})^{-1}, \end{aligned} \tag{6.4}$$

where $D_{F_0,k_1,k_2} = E_{F_0}\left[\ell_{\alpha,k_1}(Z;\alpha^*_{k_1})\ell_{\alpha,k_2}(Z;\alpha^*_{k_2})'\right]$.

For any sequence $\{F_n\}_{n\geq 1}$, we write $\rho_{0j,n} = \rho_{0j,F_n}$ and $\rho_{12,n} = \rho_{12,F_n}$ for ease of notation. The following lemma gives the limiting distribution of $\widehat{G}_{n,F_n}$ under an arbitrary sequence $F_n \in \mathcal{F}$, which extends the asymptotic distribution result in Section 4 to joint convergence.

Lemma 6.1 *Suppose that Assumptions* 4.1, 4.3, *and* B.1-B.2 *in Supplemental Appendix* B *hold. Then under any sequence* $\{F_n\}_{n\geq 1}$ *and any subsequence* $\{u_n\}$ *of* $\{n\}$ *such that with* $F_n \in \mathcal{F}$ *for all* n, $\rho_{0j,u_n} \to \rho_{0j}$ *and* $\rho_{12,u_n} \to \rho_{12}$ *for some* ρ_{0j} *and* $\rho_{12} \in [-1, 1]$, *we have*

$$\widehat{G}_{n,F_n} \to_d N(\mathbf{0}_3, \Sigma_G)$$

where Σ_G *is symmetric with* $\Sigma_G(i,i) = 1$ *for* $i = 1,2,3$, $\Sigma_G(i,i+1) = \rho_{ij}$ *for* $i = 0,1$ *and* $\Sigma_G(1,3) = \rho_{02}$.

Lemma 6.1 follows immediately from Lemmas 4.2 and 4.3 in Section 4, and Lemmas B.1-B.2 in Supplemental Appendix B and hence is omitted.

6.2 Conditional Inference for Model-Specific Parameters

In this subsection, we consider the conditional inference of a functional – denoted $\psi_1(\alpha_1^*)$ – of the parameter in model $\mathcal{M}_1$ given that $\mathcal{M}_1$ is selected.[15] Specifically, we construct a level $1-p$ conditional confidence interval, $CI_{\psi_1}(1-p)$ such that

$$\liminf_{n\to\infty} \inf_{F_0 \in \mathcal{F}_n} \Pr{}_{F_0}(\psi_1(\alpha_1^*) \in CI_{\psi_1}(1-p) | T_n \geq t) = 1-p, \tag{6.5}$$

where $\mathcal{F}_n$ is a sequence of subsets of $\mathcal{F}$ defined below. Note that we allow c to be an arbitrary number, which the user can choose according to her interpretation of the event that $\mathcal{M}_1$ is selected.

To describe our conditional confidence interval, first define a function $\Psi : R \times (-\infty, \infty] \times [-1,1] \to R$:

$$\Psi(c,h,\rho) = \begin{cases} \left[\Phi(c) - \Phi\left(c - h/\rho\right)\right] / \left[1 - \Phi\left(c - h/\rho\right)\right] & \text{if } \rho > 0 \text{ and } h \in R \\ \Phi(c) & \text{if } \rho = 0 \text{ or } h = \infty \\ \Phi(c)/\Phi(c - h/\rho) & \text{if } \rho < 0 \text{ and } h \in R. \end{cases} \tag{6.6}$$

For any $t \in R$ and $p \in (0,1)$, let $c_{1,p}$ be the solution to the equation:

$$\Psi(c_{1,p}, T_n - t, \widehat{\rho}_{01,n}) = p, \tag{6.7}$$

where $\widehat{\rho}_{0j,n} = \psi_{\alpha,k_j}(\widehat{\alpha}_{k_j})' \widehat{H}_{k_j,n}^{-1} (\widehat{v}^*_{\psi,k_j} \widehat{\sigma}_n)^{-1} n^{-1} \sum_{i=1}^n \ell_{\alpha,k_j}(Z_i; \widehat{\alpha}_{k_j}) \ell(Z_i; \widehat{\alpha}_{\mathbf{k}})$, for $j = 1,2$. This equation only needs to be solved when $T_n \geq t$ because the confidence interval is only needed then. The equation always has a unique solution when $T_n \geq t$ because $\Psi(c,h,\rho)$ is a strictly increasing function in θ with range $(0,1)$, for any $h \geq 0$ and any $\rho \in [-1,1]$. Our conditional confidence interval is of the form:

$$CI_{\psi_1}(1-p) = \left[\psi_1(\widehat{\alpha}_{k_1}) - n^{-1/2} c_{1,1-p/2} \widehat{v}^*_{\psi,k_1}, \psi_1(\widehat{\alpha}_{k_1}) - n^{-1/2} c_{1,p/2} \widehat{v}^*_{\psi,k_1}\right]. \tag{6.8}$$

These critical values depend on T_n and hence are not approximations of the conditional quantiles of $\sqrt{n}(\psi_1(\widehat{\alpha}_{k_1}) - \psi_1(\alpha_1^*))/\widehat{v}^*_{\psi,k_1}$ given $T_n > t$. Therefore, the validity of our construction is

[15] Conditional inference for a functional of the parameter in model $\mathcal{M}_2$ given that $\mathcal{M}_2$ is selected is analogous and thus omitted.

not contradictory to the impossibility results in Leeb and Pötscher (2006). The construction of the critical values is inspired by the construction in Tibshirani et al. (2016) of valid p-values and confidence intervals for post Lasso inference in a linear regression context with known Gaussian noise.[16] We generalize Tibshirani et al. (2016) to post model selection test inference for general semi-nonparametric models, and provide asymptotically exact confidence intervals without imposing special structure on the models compared or requiring knowledge about the variance-covariance Σ_G of the statistics $\widehat{G}_{n,F_n}$.

The formal justification of the above construction requires us to rule out the case where $n^{1/2}E_{F_n}\left[\ell(Z;\alpha_n^*)\right]/\widehat{\sigma}_n \to -\infty$ because in that case the conditioning event occurs with diminishing probability, and the conditional distribution of our test statistic becomes difficult to characterize. We rule out this troublesome case by considering

$$\mathcal{F}_n = \{F_0 \in \mathcal{F} : n^{1/2}E_{F_0}\left[\ell(Z;\alpha_n^*)\right]\sigma_n^{-1} - t \geq -C\}, \tag{6.9}$$

for some large $C > 0$. The formal validity result is stated as Theorem 6.1 below. The proof of this theorem is given in Supplemental Appendix B.

Theorem 6.1 *Suppose that Assumptions* 4.1*,* 4.3*, and* B.1-B.2 *in Supplemental Appendix* B *hold. Then equation* (6.5) *holds with* $\mathcal{F}_n$ *defined in* (6.9)*.*

6.3 Inference for Common Parameters

In this subsection, we consider the inference for the parameter θ that equals $\psi_1(\alpha_1)$ in model $\mathcal{M}_1$ and $\psi_2(\alpha_2)$ in model $\mathcal{M}_2$. Let $\ell_0 = f(\mathcal{M}_1, F_0) - f(\mathcal{M}_2, F_0)$. Then the pseudo-true value of θ is

$$\theta^* = \psi_1(\alpha_1^*)1(\ell_0 \geq 0) + \psi_2(\alpha_2^*)1(\ell_0 < 0). \tag{6.10}$$

Note that θ^* is a function of $(\psi_1(\alpha_1^*), \psi_2(\alpha_2^*), \ell_0)$. Because this function is discontinuous, we cannot obtain uniformly asymptotically valid inference via the Delta method even though the vector $(\psi_1(\alpha_1^*), \psi_2(\alpha_2^*), \ell_0)$ has an asymptotically jointly normal estimator by Lemma 6.1. Instead, we construct a confidence interval for θ^* by projecting a joint confidence set for $(\psi_1(\alpha_1^*), \psi_2(\alpha_2^*), \ell_0^*)$.

We let the joint confidence set of $(\psi_1(\alpha_1^*), \psi_2(\alpha_2^*), \ell_0)$ of confidence level $1-p$ to be all (x_1, x_2, x_0) such that

$$\widehat{G}_n(x_1, x_2, x_0)'\widehat{\Sigma}_G^{-1}\widehat{G}_n(x_1, x_2, x_0) \leq \chi^2_{1-p}(3), \tag{6.11}$$

[16] Asymptotically conservative one-sided inference is also available in Tibshirani et al. (2016) when the variance of the noise is unknown.

where $\chi^2_{1-p}(3)$ is the $1-p$ quantile of the chi-squared distribution with three degrees of freedom,

$$\widehat{\Sigma}_G = \begin{pmatrix} 1 & \widehat{\rho}_{01,n} & \widehat{\rho}_{02,n} \\ \widehat{\rho}_{01,n} & 1 & \widehat{\rho}_{12,n} \\ \widehat{\rho}_{02,n} & \widehat{\rho}_{12,n} & 1 \end{pmatrix}, \text{ and } \widehat{G}_n(x_1, x_2, x_0) = \begin{pmatrix} T_n - n^{1/2} x_0 / \widehat{\sigma}_n \\ n^{1/2}(\psi_1(\widehat{\alpha}_{1,n}) - x_1)/\widehat{v}^*_{\psi,k_1} \\ n^{1/2}(\psi_2(\widehat{\alpha}_{2,n}) - x_2)/\widehat{v}^*_{\psi,k_2} \end{pmatrix}$$

where $\widehat{\rho}_{0j,n}$ is defined in the previous subsection for $j = 1, 2$ and

$$\widehat{\rho}_{12,n} = \psi_{\alpha,k_1}(\widehat{\alpha}_{k_1})' \widehat{H}^{-1}_{k_1,n} \widehat{D}_{k_1,k_2,n} \widehat{H}^{-1}_{k_2,n} \psi_{\alpha,k_2}(\widehat{\alpha}_{k_2}) (\widehat{v}^*_{\psi,k_1} \widehat{v}^*_{\psi,k_2})^{-1}$$

where $\widehat{D}_{k_1,k_2,n} = n^{-1} \sum_{i=1}^n \ell_{\alpha,k_1}(Z_i; \widehat{\alpha}_{k_1}) \ell_{\alpha,k_2}(Z_i; \widehat{\alpha}_{k_2})'$. Then the projected confidence set of confidence level $1-p$ for θ^* is

$$CI_\theta(1-p) = \{\theta = x_1 1(x_0 \geq 0) + x_2 1(x_0 < 0) : \widehat{G}_n(x_1, x_2, x_0)' \widehat{\Sigma}_G^{-1} \widehat{G}_n(x_1, x_2, x_0) \leq \chi^2_{1-p}(3)\}. \quad (6.12)$$

Theorem 6.2 below shows the uniform asymptotic validity of this confidence interval. The proof of this theorem is given in Supplemental Appendix B.

Theorem 6.2 *Suppose that Assumptions* 4.1, 4.3, *and* B.1-B.2 *in Supplemental Appendix* B *hold. In addition, suppose that there is a constant $C > 0$ such that under all $F_0 \in \mathcal{F}$, we have $\rho_{\min}(\Sigma_G) > C^{-1}$. Then $\liminf_{n\to\infty} \inf_{F_0 \in \mathcal{F}} \Pr_{F_0}(\theta^* \in CI_\theta(1-p)) \geq 1-p$.*

7 Simulation Studies

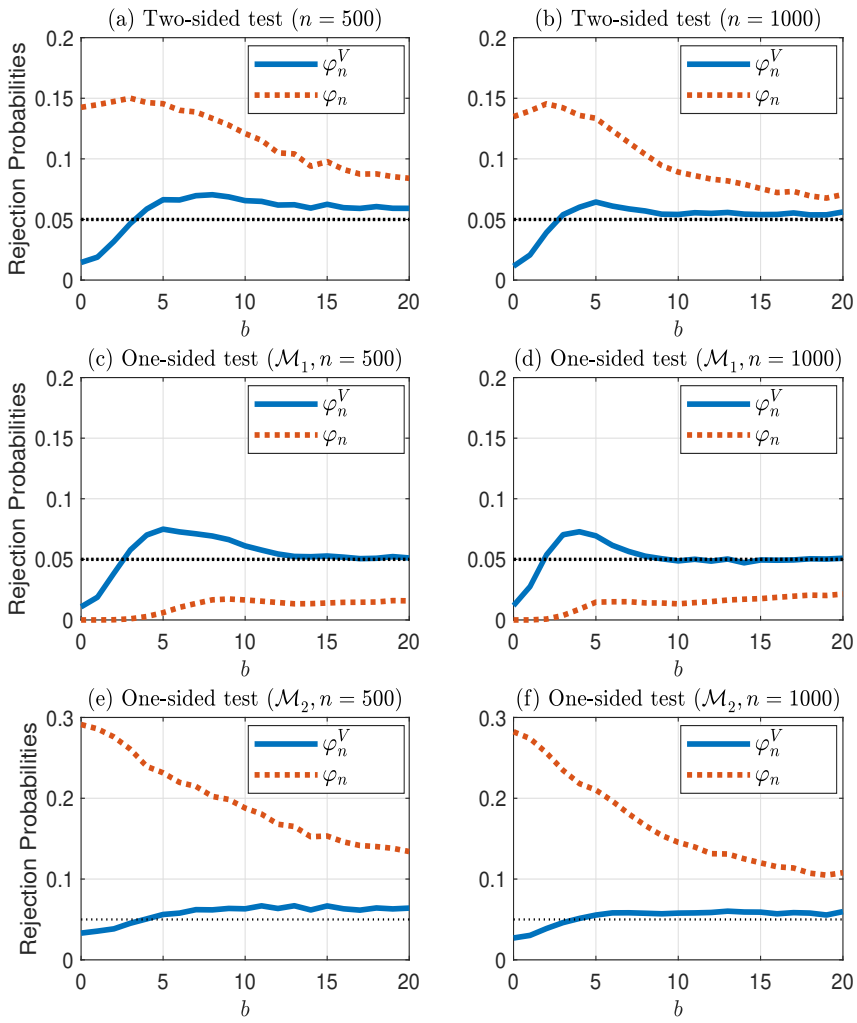
In this section, we report Monte Carlo simulation results on the finite sample performance of the nondegenerate test and the conditional confidence interval $CI_\psi(1-p)$.

Consider the following two models,

$$\mathcal{M}_1 : \ E[Y|X_1] = \beta_{10} + X_1 \beta_{11} \text{ and } \mathcal{M}_2 : \ E[Y|X_2, X_3] = X_2 \beta_{21} + g(X_3), \quad (7.1)$$

where $(\beta_{10}, \beta_{11})' \in R^2$, $\beta_{21} \in R$ and $g(\cdot) \in C^\infty([0,1])$. This example readily fits into the framework of regression model studied in Section 5 with $\alpha_1(x_1) = \beta_{10} + \beta_{11} x_1$, $\mathcal{A}_1 = \{b_0 + x_1 b_1 : (b_0, b_1)' \in R^2\}$, $\alpha_2(x_2, x_3) = x_2 \beta_{21} + g(x_3)$, and $\mathcal{A}_2 = \{x_2 b_2 + g(x_3) : b_2 \in R, g(\cdot) \in C^\infty([0,1])\}$.

Figure 2: Null Rejection Rates of the Tests



To generate the data, let X_1, X_2 be independent standard normal random variables, and let X_3 be a uniform random variable independent of X_1 and X_2. Let ε be standard normal and independent of X_1, X_2 and X_3. Let

$$Y = 1 + X_1 a + X_2 b + c\sqrt{2}\sin(10\pi X_3) + \varepsilon. \tag{7.2}$$

7.1 Uniform Model Selection Test

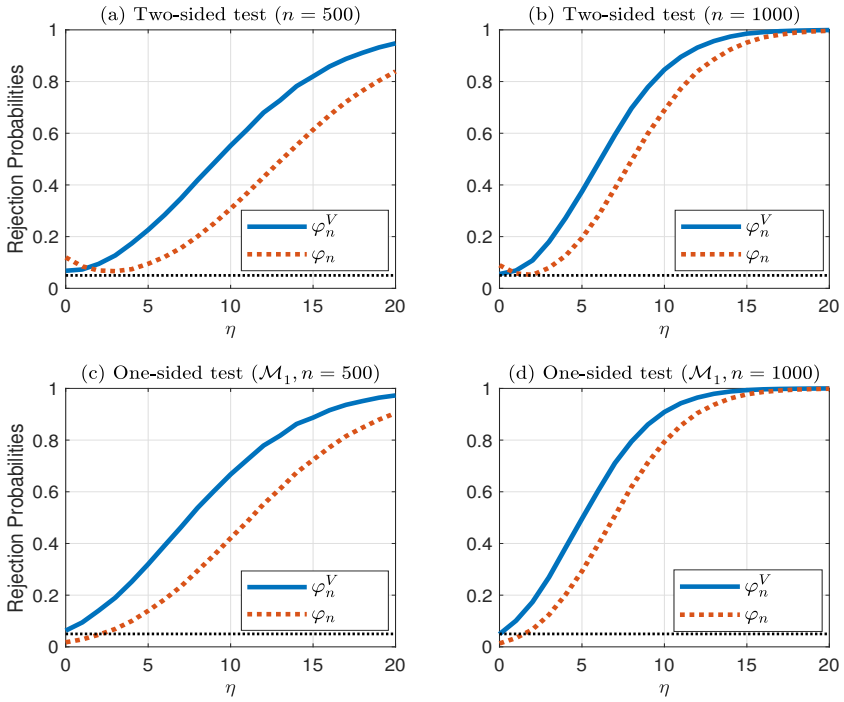
Independence between the regressors and the additive structure in the generation process of Y are not important for the performance of our test, but they allow us to derive an analytical form of

the fit measures and hence to conveniently characterize the null hypothesis. By exploiting them, we see that $u_1 = X_2 b + c\sqrt{2}\sin(10\pi X_3) + \varepsilon$, and $u_2 = X_1 a + \varepsilon$. Thus,

$$-2f(\mathcal{M}_1, F_0) = E_{F_0}[u_1^2] = b^2 + 1 + c^2 \text{ and } -2f(\mathcal{M}_2, F_0) = E_{F_0}[u_2^2] = a^2 + 1. \tag{7.3}$$

Therefore, the null hypothesis holds if and only if $a^2 = b^2 + c^2$, and when $a^2 > b^2 + c^2$, $f(\mathcal{M}_1, F_0) > f(\mathcal{M}_2, F_0)$. When $a^2 = b^2 + c^2 = 0$, $u_1 = u_2$, in which case, $\omega^2_{F_0,*} = 0$. Otherwise, $\omega^2_{F_0,*} > 0$.

Figure 3: Null and Alternative Rejection Rates of the Tests



To evaluate the performance of the nondegenerate test, we consider two collections of DGPs. One collection sets $a^2 = b^2 + c^2$, $b = c$, and b (and c) to grid points in $[0, 0.4]$ with the spacing of 0.02 between adjacent points. This is the null collection in which, as b runs from 0 to 0.4, $\omega^2_{F_0,*}$ grows from zero up. The other collection sets $b = c = 0.2$, $a^2 = b^2 + c^2 + \eta$, and η to grid points in $[0, 0.2]$ with the spacing of 0.01 between adjacent points. This is the alternative collection in which, as η runs from 0 to 0.2, model $\mathcal{M}_2$ gets worse and worse relative to model $\mathcal{M}_1$. We implement the nondegenerate test as well as the naive extension test as they are defined in Section 3. We use cubic spline to approximate $g(\cdot)$ in model 2.[17]

Selection of the number of series terms on approximating $g(\cdot)$ is important for the implementation of our nondegenerate test and conditional confidence intervals. For regression examples like

[17]Fourier series yields similar results.

the one considered in this section, we recommend using cross-validation with a slowly diverging lower bound imposed on the number of sieve terms. Cross-validation is a commonly used method in the semi/nonparametric regression literature for selecting smoothing parameters and has been shown to yield optimal rate of convergence in nonparametric series regression (ref. Li (1987) and Andrews (1991)) as well as in nonparametric series quantile regression (ref. Chetverikov and Liao (2019)). The slowly diverging lower bound – we use $2\log(\log(n))$ – ensures that the dimension of at least one model to diverge to infinity which is needed for our Assumption 4.3(b).[18] [19]

The finite sample rejection rates of the tests are calculated using 50,000 simulated samples. Figure 2 presents the rejection rates of the two-sided and one-sided tests under the first collection of DGPs—the collection of null DGPs. Graphs (a) and (b) show the tests for H_0 against $H_1 : f(\mathcal{M}_1, F_0) \neq f(\mathcal{M}_2, F_0)$ with sample size $n = 500$ and $n = 1000$ respectively. In graph (a), the naive extension test (dotted line) over-rejects noticeably when $\omega^2_{F_0,*}$ is zero or close to zero. On the other hand, the rejection rate of the nondegenerate test (solid line) never exceeds the nominal level by much, although there is some under-rejection at very small b's and slight over-rejection at bigger b's. When the sample size is increased from 500 to 1000, the rejection rate of the nondegenerate test gets closer to the nominal level while the naive extension test maintains overall over-rejection and under-rejection respectively. Graphs (c) and (d) show the one-sided tests for H_0 against $H_1 : f(\mathcal{M}_1, F_0) > f(\mathcal{M}_2, F_0)$ with sample sizes $n = 500$ and $n = 1000$ respectively, and graphs (e) and (f) show the one-sided tests for H_0 against $H_1 : f(\mathcal{M}_1, F_0) < f(\mathcal{M}_2, F_0)$ with sample size $n = 500$ and $n = 1000$ respectively. Recall that model $\mathcal{M}_1$ is the more parsimonious one. As we can see, our robust test has a rejection rate of approximately 5% against both one-sided alternative hypotheses. The naive extension test has severe under-rejection when $\mathcal{M}_1$ is better under the alternative (graphs (c) and (d)) and severe over-rejection when $\mathcal{M}_2$ is better under the alternative (graphs (e) and (f)). This behavior is in line with our theoretical derivation.

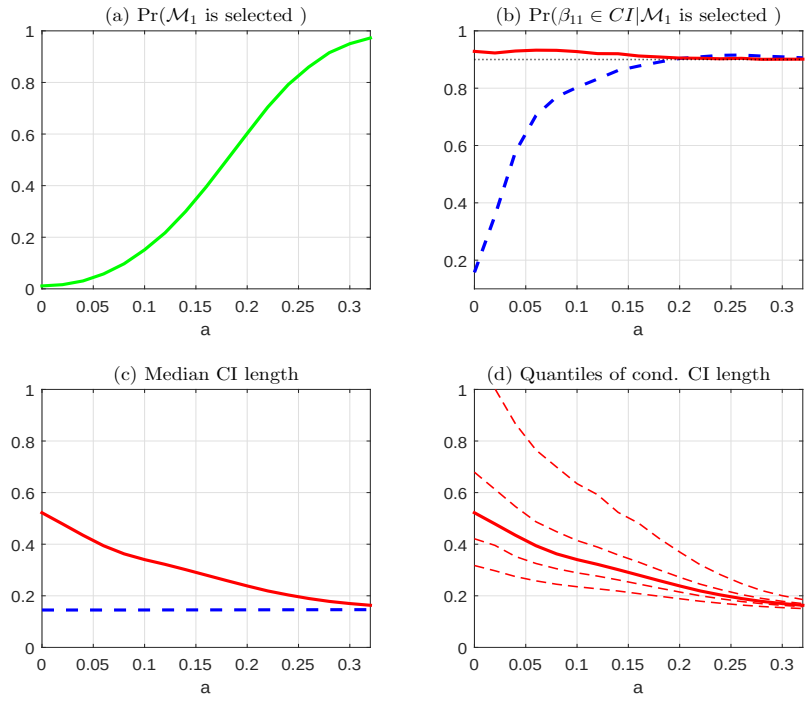
The rejection rates of the two-sided and one-sided tests under the second collection of DGPs—the collection of null and alternative DGPs are included in Figure 3. In this set of DGPs, the null hypothesis H_0 holds when $\eta = 0$ and the alternative hypothesis $H_1 : f(\mathcal{M}_1, F_0) > f(\mathcal{M}_2, F_0)$ holds when $\eta \neq 0$. The model $\mathcal{M}_2$ becomes worse when the magnitude of η becomes large. Moreover, in this set of DGPs, $\omega^2_{F_0,*} > 0$ since $b = c = 0.2$ for all different values of η. In Figure 3, we see that the nondegenerate test has rejection rates close to the nominal level 5% under the null H_0 (when $\eta = 0$), while the naive extension test over-rejects for the two-sided alternative (graphs (a) and (b)) and under-rejects for the one-sided alternative (graphs (c) and (d)). This is again in line with

[18] In our simulations, we also impose an upper bound of 15 on the cross-validation search range.

[19] Strictly speaking, the theory presented in earlier sections applies only to non-data-dependent choices of series terms. However, in practice, cross-validation is often employed, which is why we suggest it for empirical implementation of our tests and why we use it in this simulation example. The performance of our test with the cross-validated series terms is encouraging.

our theoretical results that the naive extension test favors large models. For the power properties, the nondegenerate test has the best power across most of the range of η in the two-sided test. It also has better power than the naive extension test in the one-sided test.

Figure 4: Performance of Conditional Confidence Interval for β_{11}.



7.2 Conditional Confidence Interval

In this subsection, we evaluate the performance of the conditional confidence interval $CI_{\psi_1}(1-p)$ with $p = 0.1$. Consider the parameters of interest β_{11} and β_{21}. Let model $\mathcal{M}_1$ be selected if $T_n > z_{0.95}$ and model $\mathcal{M}_2$ be selected otherwise. Consider the DGPs with $b = 0$, $c = 0$ and a running from 0 to 0.32. We report the probability of the model being selected, as well as the coverage probability, the median length, and other quantiles of the length of the conditional confidence interval. For comparison, we also report the performance of the naive confidence interval that ignores the model selection step, that is, for $j = 1, 2$,

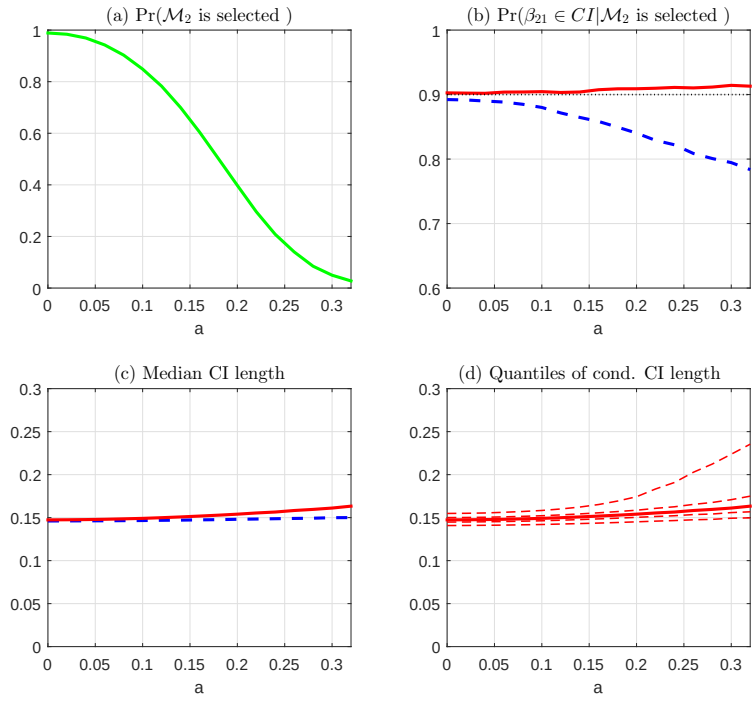
$$CI_j^{\text{naive}}(1-p) = [\psi_j(\widehat{\alpha}_{k_j}) - n^{-1/2} z_{1-p/2} \widehat{v}^*_{\psi,k_j}, \psi_j(\widehat{\alpha}_{k_j}) - n^{-1/2} z_{p/2} \widehat{v}^*_{\psi,k_j}], \tag{7.4}$$

where z_p stands for the p quantile of the standard normal distribution. Note that the conditional CI is only different from the naive CI in that it uses the critical value $c_{j,p}$ instead of z_p.

Figure 4 shows the results for β_{11}, and Figure 5 shows those for β_{21}. In graphs (b) and (c)

of both figures, the blue dotted lines are for the naive CIs and the red solid lines are for our conditional CIs; in graph (d), the five lines are respectively the 25%, 40%, 50%, 60%, and 75% quantile of the length of the conditional CI. As we can see, the naive CI may severely under-cover when the probability that the model is selected is small. On the other hand, the coverage probability of our conditional CI is always very close to the nominal level. In terms of length, our conditional CI is longer than the naive CI when the naive CI under-covers, and is about the same as the naive CI when the latter has good coverage properties.

Figure 5: Performance of Conditional Confidence Interval for β_{21}



By definition, the critical values of the conditional CI depends on T_n, and thus is random. As a result, the length of the conditional CI is also random. Part (d) of Figure 4 shows the variability of the length of the conditional CI. As we can see, the variability is small when the probability that the model under consideration is selected is large, and can be big otherwise. In light of the difficulties of post model selection inference pointed out by Leeb and Pötscher (2005), we view the variability and the extra length of the conditional CI as an inevitable price to pay for its good coverage property. It is encouraging to see that the conditional CI has similar length as the naive CI when the latter does not under-cover.

8 An Empirical Example

In this section we illustrate the use of our robust model selection test and the conditional confidence interval in the study of life-cycle schooling choices. We compare two models considered in Cameron and Heckman (1998) using our model selection test, and also report the conditional confidence intervals of some of the model specific parameters. The two models considered are parametric likelihood models. We consider our theory presented for the semi/non-parametric environment as reasonable approximation to this context since the number of parameters in each model is large.

8.1 Model Description

We apply our test on the comparison of two life cycle schooling models taken from Cameron and Heckman (1998). The paper is a classic piece of structural modeling, which is why we use it to illustrate our model selection and post model selection inference tools.

Consider an individual deciding how much schooling (S, number of years of schooling) to complete, and consider a vector of individual characteristics X that may be relevant for this decision. The first model (Model $\mathcal{M}_1$) is the *logit transition* model that Cameron and Heckman (1998) set up to formalize the statistical model prevalent in the political science literature at the time. To describe this model, define the binary variable $D_s = 1\{S \geq s\}$. This variable indicates whether or not the individual completed grade s or not. The model imposes a logit form on the transition probability from completing grade s to completing grade $s+1$:

$$\Pr(D_{s+1} = 1 | D_s = 1, X) = \frac{\exp(X'\beta_s)}{1 + \exp(X'\beta_s)},$$

where β_s is the grade-specific effect of X on the transition probability. This implies that the probability of s being the highest grade completed is given by

$$P_1(s|X, \theta_1) = \frac{1}{1 + \exp(X'\beta_s)} \times \frac{\exp(X'\beta_{s-1})}{1 + \exp(X'\beta_{s-1})} \times \cdots \times \frac{\exp(X'\beta_1)}{1 + \exp(X'\beta_1)} \tag{8.1}$$

where $\theta_1 = (\beta_1', \beta_2', \ldots, \beta_{\bar{s}}')'$ with $\bar{s}$ being the highest grade available. Note that this model contains many parameters since β_s is allowed to be different across s. However, it allows no individual heterogeneity other than the logit error, and thus effectively assumes that the population making the transition decision at different grade levels are the same. In technical terms, it rules out dynamic selection as the population move up grades. This is an important drawback of the model as discussed in Cameron and Heckman (1998).

The second model (Model $\mathcal{M}_2$) is an *ordered logit* model. Cameron and Heckman (1998) set

up this model as an economically well-grounded yet parsimonious contestant to the first model. In this model, the probability of s being the highest grade completed is given by

$$P_2(s|X,\theta_2) = \int_\Omega F(\alpha_{s-1} + y + X'\beta) - F(\alpha_s + y + X'\beta) dF_\omega(y), \tag{8.2}$$

where $\theta_2 = (\alpha_1, \ldots, \alpha_{\bar{s}-1}, \beta')'$, $F(t) = \exp(t)(1+\exp(t))^{-1}$, $\alpha_0 = +\infty$ for the highest possible grade $\bar{s}$, and ω is an unobservable individual type that has support Ω and distribution $F_\omega(\cdot)$. From the statistical point of view, the ordered logit aspect is not fundamentally different from the logit transition model since an ordered logit model can be written as a transition model with some (albeit non-logit) shocks in the transition decisions. However, this model adds the unobservable type ω, which makes sure that the dynamic selection effect is accounted for. The model further specifies that $\Omega = \{0, \omega_2\}$, and $F_\omega(y) = p_1 1(y \geq 0) + (1-p_1)1(y \geq \omega_2)$ for unknown parameters $\omega_2 > 0$ and $p_1 \in (0,1)$. The model uses a parsimonious specification for the effect of X on the ordered logit cutoffs — the β is not indexed by s.

8.2 Implementation Details

We compare the models in terms of their population log-likelihood. We implement the two-sided version of both our robust test and the Vuong (1989) test. The detailed implementation steps are as follows:

1. Given the data set $(S_i, X_i)_{i=1}^n$, define the log-density functions for the two models respectively as $m_j(S_i, X_i, \theta_j) = \log P_j(S_i|X_i, \theta_j)$ for $j = 1, 2$.

2. Define the log-likelihoods of the two models as $\widehat{f}(\mathcal{M}_j, \theta_j) = n^{-1}\sum_{i=1}^n m_j(S_i, X_i, \theta_j)$ for $j = 1, 2$.

3. Respectively for $j = 1, 2$, compute $\widehat{\theta}_{n,j} = \arg\max_{\theta_j} \widehat{f}(\mathcal{M}_j, \theta_j)$ using a suitable maximization algorithm, like the **fminunc** function in Matlab, or the **ml** package in Stata.

4. Compute $\bar{\ell}_n(\widehat{\theta}_n) = \widehat{f}(\mathcal{M}_1, \widehat{\theta}_{n,1}) - \widehat{f}(\mathcal{M}_2, \widehat{\theta}_{n,2})$ and $\widehat{\omega}_n^2(\widehat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n (\ell_i(\widehat{\theta}_n) - \bar{\ell}_n(\widehat{\theta}_n))^2$, where $\ell_i(\theta) = m_1(S_i, X_i, \theta_1) - m_2(S_i, X_i, \theta_2)$ and $\widehat{\theta}_n = (\widehat{\theta}'_{n,1}, \widehat{\theta}'_{n,2})'$.

5. Compute the score $\partial m_j(S_i, X_i, \widehat{\theta}_{n,j})/\partial\theta_j$ for each i and $j = 1, 2$ either by deriving and using the analytical formula for the first derivative function, or by numerical differentiation of the log-density function. Let $\widehat{\ell}_{\theta,i} = \partial m_1(S_i, X_i, \widehat{\theta}_{n,1})/\partial\theta_1 - \partial m_2(S_i, X_i, \widehat{\theta}_{n,2})/\partial\theta_2$.

6. Compute $\widehat{D}_n = n^{-1}\sum_{i=1}^n \widehat{\ell}_{\theta,i}\widehat{\ell}'_{\theta,i}$.

7. Compute $\widehat{H}_{n,j} = \partial^2 f(\mathcal{M}_j, \widehat{\theta}_{n,j})/\partial\theta_j\partial\theta_j'$ for $j = 1, 2$ either numerically or using analytical formula of the second derivative. Let $\widehat{H}_n = diag(\widehat{H}_{n,1}, -\widehat{H}_{n,2})$.

8. Let $T_n^V = n^{1/2}\bar{\ell}_n(\widehat{\theta}_n)(\widehat{\omega}_n^2(\widehat{\theta}_n))^{-1/2}$ and let $T_n = \frac{n\bar{\ell}_n(\widehat{\theta}_n)+2^{-1}tr(\widehat{D}_n\widehat{H}_n^{-1})}{\sqrt{\max\{n\widehat{\omega}_n^2(\widehat{\theta}_n)-\frac{1}{2}tr((\widehat{D}_n\widehat{H}_n^{-1})^2),\, 2^{-1}tr((\widehat{D}_n\widehat{H}_n^{-1})^2)\}}}$.

9. Compute the p-value of our robust test as p-value $= 2(1 - \Phi(T_n))$ and of the Vuong (1989) test as p-value$^V = 2(1 - \Phi(T_n^V))$.

8.3 Data and Results

We compare these models using data from the 1997 wave of the National Longitudinal Survey (NLSY 97). This is a newer wave of the NLSY 79 used in Cameron and Heckman (1998) that covers a sample of young men and women born between 1980 and 1984. Following Cameron and Heckman (1998), we use the male sample only and drop observations with missing values on the relevant variables. Our final sample contains 1938 individuals.[20]

The X variables for models $\mathcal{M}_1$ contain a constant and 15 nonconstant variables including the number of siblings, highest grade completed by father, that by mother, broken family dummy, log family income, urban/rural residence dummy, etc. and interaction terms. The X variable for model $\mathcal{M}_2$ contains all those 15 nonconstant variables, but does not contain a constant term. We aggregate the grades (S) into four, following Cameron and Heckman (1998): completed high school ($s = 1$), attended college ($s = 2$), graduated college ($s = 3$) and attended 17 or more years of school ($s = 4$). As a result, Model $\mathcal{M}_1$ contains $4 \times 16 = 64$ parameters and Model $\mathcal{M}_2$ contains $4 + 15 + 2 = 21$ parameters. Clearly, Model $\mathcal{M}_2$ is much more parsimonious than Model $\mathcal{M}_1$.[21]

Table 1: Model Selection Tests Based on NLSY 97

	Test Statistic	p-value
Robust Test	1.856	.063
Vuong (1989) Test	3.924	.000

Table 1 shows the value of the test statistics as well as p-values of both tests. The Vuong (1989) test strongly rejects the null in favor of the less parsimonious models $\mathcal{M}_1$. However, we believe that the strong rejection is partly due to the bias in favor of large models. Indeed, the robust test that corrects the bias presents much weaker evidence against the parsimonious Model $\mathcal{M}_2$. In particular, according to the robust test, we cannot reject the null that $\mathcal{M}_2$ is as good as

[20] Results using reconstructed sample from the NLSY 79 are reported in Supplemental Appendix G.

[21] Parameter estimates are irrelevant for our analysis and thus are omitted. They are available upon request.

$\mathcal{M}_1$ at significance level 5%. Cameron and Heckman (1998) advocate for $\mathcal{M}_2$ for its simplicity and interpretability. Our robust test shows that it achieves the simplicity without sacrificing too much of its fit to the data. In contrast, the Vuong (1989) test tells a different story and can be misleading.[22]

To illustrate our conditional confidence interval, we computed these intervals for the parameters in Model $\mathcal{M}_2$ conditional on the event that $T_n < z_{0.975} \approx 1.96$. It turns out that the conditional confidence intervals are the same as the naive CI's computed using the sandwich standard error formula. Upon further inspection, we find that the correlation coefficients of T_n and the parameter estimates of Model $\mathcal{M}_2$ are nearly zero, which causes $c_{2,p}$ to be the same as z_p up to at least the sixth digit. We believe that this is a special feature of this application and does not have general implication.

9 Conclusion

This paper studies the statistical comparison of semi/nonparametric models when the competing models are overlapping nonnested, strictly nonnested, or nested. We provide a new model selection test that achieves uniform asymptotic size control in all testing scenarios. The new test uses a critical value from standard normal distribution and employs a bias-corrected quasi-likelihood ratio statistic that is easy to compute in practice. This makes our test convenient for empirical implementation. Moreover, uniformly valid post model selection test inference procedures of model parameters are also provided. Simulation results show that our test and our post model selection test confidence interval perform well in finite samples.

At least two future research directions arise from the findings of this paper. First, the theory of this paper is established under the i.i.d. assumption of the data. It is important to extend it for the comparison of time series models with dependent data. Second, when there are many competing models to be compared, it can be interesting to construct a model confidence set that covers the best model with valid asymptotic size. These directions of research form part of our ongoing work, during the course of which some preliminary results have been obtained.

[22]Cameron and Heckman (1998) implemented the Vuong (1989) test with the Bayesian information criterion (BIC) penalty, and thus were effectively testing the null hypothesis that

$$H_0 : f(\mathcal{M}_1, F_0) - \frac{k_1 \log(n)}{2n} - \frac{tr(D_{F_0,k_1} H_{F_0,k_1}^{-1})}{2n} = f(\mathcal{M}_2, F_0) - \frac{k_2 \log(n)}{2n} - \frac{tr(D_{F_0,k_2} H_{F_0,k_2}^{-1})}{2n},$$

where $f(\mathcal{M}_j, F_0) \equiv \max_{\theta_j} E_{F_0} \log P_j(S|X;\theta_j)$ is the Kullback-Leibler distance from model $\mathcal{M}_j$ to the data. Their test result strongly rejects the null in favor of the ordered logit model. The penalty would not matter asymptotically in the asymptotic framework assuming strict nonnestedness, as argued in Vuong (1989). Yet it clearly leads to a different testing conclusion here.

References

Ai, C. and X. Chen (2007): "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables," *Journal of Econometrics*, 141, 5–43.

Aït-Sahalia, Y., P. J. Bickel, and T. M. Stoker (2001): "Goodness-of-fit Tests for Kernel Regression with an Application to Option Implied Volatility," *Journal of Econometrics*, 105, 363–412.

Andrews, D. W. (1991): "Asymptotic optimality of generalized CL, cross-validation, and generalized cross-validation in regression with heteroskedastic errors," *Journal of Econometrics*, 47, 359–377.

Atkinson, A. (1970): "A Method for Discriminating Between Models," *Journal of Royal Statistical Society*, B 32, 323–353.

Barseghyan, L., F. Molinari, T. O'Donoghue, and J. C. Teitelbaum (2013): "The Nature of Risk Preferences: Evidence from Insurance Choices," *American Economic Review*, 103, 2499–2529.

Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernández-Val (2011): "Conditional quantile processes based on series or many regressors," *arXiv preprint arXiv:1105.6154*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28, 29–50.

Bontemps, C., J.-P. Florens, and J.-F. Richard (2008): "Parametric and Non-parametric Encompassing Procedures," *Oxford Bulletin of Economics and Statistics*, 70, 751–780.

Cameron, S. V. and J. J. Heckman (1998): "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political Economy*, 106, 262–333.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, North Holland, vol. 6, 5369–5468.

Chetverikov, D. and Z. Liao (2019): "On the Optimality of Cross-Validated Series Quantile Estimators," *Working Paper*, 47, 359–377.

Coate, S. and M. Conlin (2004): "A Group Rule - Utilitarian Approach to Voter Turnout: Theory and Evidence," *The American Economic Review*, 94, 1476–1504.

Cox, D. R. (1961): "Tests of Separate Families of Hypotheses," Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability, University of California Press: Berkeley.

——— (1962): "Further Results on Tests of Separate Families of Hypotheses," *Journal of the Royal Statistical Society. Series B (Methodological)*, 24, 406–424.

Donald, S. G., G. W. Imbens, and W. K. Newey (2003): "Empirical likelihood estimation and consistent tests with conditional moment restrictions," *Journal of Econometrics*, 117, 55–93.

Fan, Y. and Q. Li (1996): "Consisstent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms," *Econometrica*, 64, 865–890.

Gandhi, A. K. and R. Serrano-Padial (2015): "Does Belief Heterogeneity Explain Asset Prices: the Case of the Longshot Bias," *Review of Economic Studies*, 82, 156–186.

Gourieroux, C. and A. Monfort (1995): "Testing, Encompassing, and Simulating Dynamic Econometric Models," *Econometric Theory*, 11, 195–228.

Gowrisankaran, G. and M. Rysman (2012): "Dynamics of Consumer Demand for New Durable Goods," *Journal of Political Economy*, 120, 1173–1219.

Hall, P. (1984): "Central limit theorem for integrated square error of multivariate nonparametric density estimators," *Journal of multivariate analysis*, 14, 1–16.

Hong, Y. and H. White (1995): "Consistent Specification Testing via Nonparametric Series Regression," *Econometrica*, 63, 1133–1159.

Horowitz, J. L. and W. Härdle (1994): "Testing a parametric model against a semiparametric alternative," *Econometric theory*, 10, 821–848.

Hsu, Y.-C. and X. Shi (2017): "Model-selection tests for conditional moment restriction models," *The Econometrics Journal*, 20, 52–85.

Jun, S. J. and J. Pinkse (2012): "Testing Under Weak Identification with Conditional Moment Restrictions," *Econometric Theory*, 28, 1229–1282.

Karaivanov, A. and R. M. Townsend (2014): "Dynamic Financial Constraints: Distinguishing Mechanism Design rom Exogenously Incomplete Regimes," *Econometrica*, 82, 887–959.

Kendall, C., T. Nannicini, and F. Trebbi (2015): "How Do Voters Respond to Information: Evidence from a Randomized Campaign," *American Economic Review*, 105, 322–353.

Kitamura, Y. (2000): "Comparing Misspecified Dynamic Econometric Models Using Nonparametric Likelihood," unpublished manuscript, Department of Economics, University of Pennsylvania.

Lavergne, P. and Q. H. Vuong (1996): "Nonparametric Selection of Regressors: The Nonnested Case," *Econometrica*, 207–219.

——— (2000): "Nonparametric Significance Testing," *Econometric Theory*, 16, 576–601.

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

Leeb, H. and B. M. Pötscher (2006): "Can One Estiamte the Conditional Distribution of Post-Model-Selection Estimators?" *The Annals of Statistics*, 34, 2554–2591.

Li, K.-C. (1987): "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set," *The Annals of Statistics*, 15, 958–975.

Li, T. (2009): "Simulation Based Selection of Competing Structural Econometric Models," *Journal of Econometrics*, 148, 114–123.

Loh, W. Y. (1985): "A New Method for Testing Separate Families of Hypothesis," *Journal of the American Statistical Association*, 80, 362–368.

Mizon, G. and J. F. Richard (1986): "The Encompassing Principle and Its Applications to Testing Nonnested Hypothesis," *Econometrica*, 3, 657–678.

Moines, S. and S. Pouget (2013): "The Bubble Game: An Experimental Study of Speculation," *Econometrica*, 81, 1507–1539.

Paulson, A. L., R. M. Townsend, and A. Karaivanov (2006): "Distinguishing Limited Liability from Moral Hazard in a Model of Entrepreneurship," *Journal of Political Economy*, 114, 100–144.

Pesaran, M. H. (1974): "On the General Problem of Model Selection," *Review of Economic Studies*, 41, 153–171.

Pesaran, M. H. and A. S. Deaton (1978): "Testing Nonnested Nonlinear Regression Models," *Econometrica*, 76, 677–694.

Pesaran, M. H. and M. R. D. Ulloa (2008): "Non-nested Hypotheses," Palgrave Macmillan, 2nd ed.

Ramalho, J. J. S. and R. J. Smith (2002): "Generalized Empirical Likelihood non-nested Tests," *Journal of Econometrics*, 107, 99–125.

Rivers, D. and Q. Vuong (2002): "Model Selection Tests for Nonlinear Dynamic Models," *Econometrics Journal*, 5, 1–39.

Schennach, S. M. and D. Wilhelm (2017): "A Simple Parametric Model Selection Test," *Journal of the American Statistical Association*, 112, 1663–1674.

Shi, X. (2015a): "Model Selection Tests for Nonnested Moment Inequality Models," *Journal of Econometrics*, 187, 1–17.

——— (2015b): "A Nondegenerate Vuong Test," *Quantitative Economics*, 6, 85–121.

Stone, C. J. (1985): "Additive regression and other nonparametric models," *The Annals of Statistics*, 13, 689–705.

Tian, X. and J. Taylor (2015): "Asymptotics of Selective Inference," unpublished manuscript, Department of Statistics, Stanford University.

Tibshirani, R. J., A. Rinaldo, R. Tibshirani, and L. Wasserman (2015): "Uniform Asymptotic Inference and the Bootstrap After Model Selection," unpublished manuscript, Carnegie Mellon University.

Tibshirani, R. J., J. Taylor, R. Lockhart, and R. Tibshirani (2016): "Exact Post-Selection Inference for Sequential Regression Procedures," *Journal of the American Statistical Association*, 40, 1198–1232.

Vuong, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica*, 57, 307–33.