A Uniform Model Selection Test for Semi/Nonparametric Models [*]

Zhipeng Liao [†] Xiaoxia Shi [‡]

June 8, 2017

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 uniform 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, Post Model Selection Inference, Semi/Nonparametric Models, Model Selection Test

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

[*] 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, Duke, Harvard/MIT, Chinese University of Hong Kong, UCLA, UCSD, UC-Riverside, USC, Yale, UW-Madison, Northwestern, and SUFE. 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.

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.

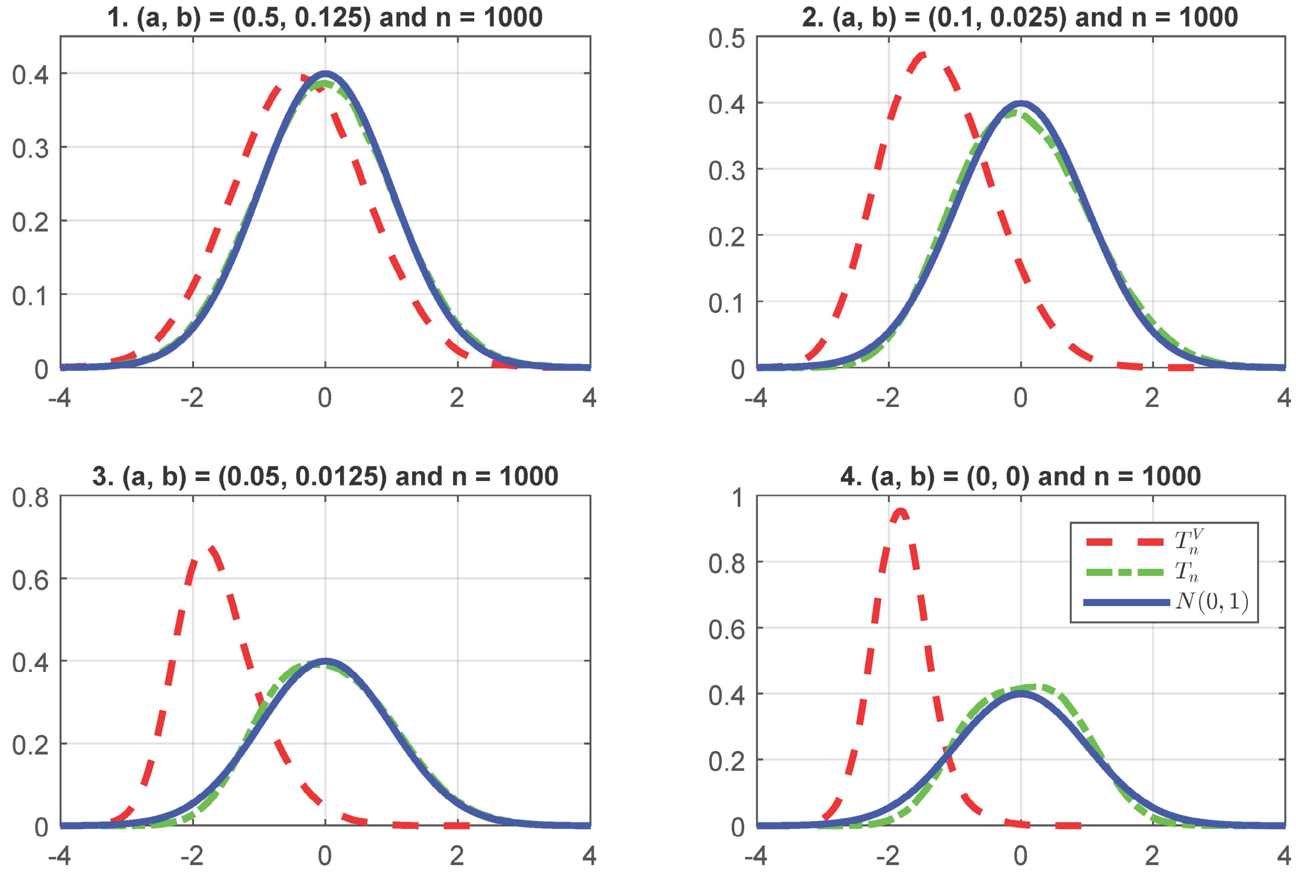
Depending on the structure of the two candidate models, the quasi-likelihood ratio (QLR) statistic used in Vuong (1989) may have different asymptotic (normal or weighted chi-square) distributions under the null hypothesis that the KLIC difference of two likelihood models is zero. The model structure is unknown when the models compared are overlapping nonnested. In such case, a pretest for the latent model structure could be performed to determine which asymptotic distribution to use for the model selection test. But the two-step procedure may (a) not be uniformly valid if the pretest does not use a conservative critical value, or (2) not be powerful because the pretest makes rejection difficult, especially when the pretest employs a conservative critical value to guarantee uniform size control. 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 model, or adding arbitrary bias correction term to the QLR statistic.

The first contribution of this paper is a revised QLR statistic that has the same asymptotic (standard normal) distribution regardless the latent model structure. The revised QLR statistic has the model complexity-related bias term mentioned above removed. Thus, the resulting model selection test has uniform asymptotic size control and is less susceptible to manipulation. The asymptotic distribution of the revised QLR statistic is established assuming that the number of unknown parameters of either candidate model or both increase with sample size. This is a natural condition when one or both of the models involve infinite dimensional parameters, and is a good approximation to the situation of parametric models with a moderately large number of parameters (e.g. cube root of the sample size number of parameters). The latter case is analogous to the many IV asymptotic theory of Bekker (1994) and subsequent works, and is a particularly important case because large parameter models are frequently used in empirical work.

The properties of the standardized QLR statistic (as used in Vuong (1989)) and our revised

[1] See, e.g., Fafchamps (1993), Moon and Stotsky (1993), Palfrey and Prisbrey (1997), Bonnal et al. (1997), Cameron and Heckman (1998), Caballero and Engel (1999), Heath et al. (1999), Nyamko and Schotter (2002), Coate and Conlin (2004), Bisin et al. (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.

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 1000000 simulated samples.

QLR test are illustrated in Figure 1. The simulation study in Figure 1 compares two parametric linear regression models based on their mean-squared error. Here, model 1 has two regressors and model 2 has 17 regressors. The red dashed line represents the finite sample density of the QLR statistic T_n^V defined in (3.5) below. Under the null hypothesis, T_n^V has asymptotic standard normal distribution when the latent parameters (a, b) are not zero, and it has asymptotic weighted chi-square distribution when (a, b) are zero. Suppose that one conducts model selection test using the critical value from the standard normal distribution. Although such test is justified by the asymptotic distribution of T_n^V when (a, b) are not zero, we see that it is over-rejecting under the null 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. Because the standard normal distribution is a poor approximation to the finite sample density of T_n^V when (a, b) are close to zero, Figure 1 also shows why pre-testing the latent model structure may be harmful for valid model selection test. The finite sample properties of the test proposed in this paper are also investigated in the simulation study in Figure 1. The green dash-dotted line represents the finite sample density of the revised QLR statistic T_n defined in (3.14) below. It is clear that 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 robust to the complexities of the two compared models.

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 unavoid in most 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 such the confidence intervals may have coverage probabilities very different from the nominal level. In this paper, we provide uniformly asymptotically valid confidence intervals for the parameters in the selected model. The uniform confidence intervals use critical values calculated from a hybrid conditional cumulative distribution function (CDF) of normal random variables that is easy to compute in practice.

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. The model selection part of this paper is the most closely related to Shi (2015b) in its basic idea. Shi (2015b) proposes a simulation based procedure that achieves uniform size control. The asymptotic size

[2] Extensions include Lavergne and Vuong (1996), Rivers and Vuong (2002), Kitamura (2000), Chen and Fan (2005),Chen et al. (2007), among others.

control of her procedure, when the procedure is applied to the semi/non-parametric setting with sieve approximation, is trivially justified by our asymptotic results because her test statistic is smaller and her critical value bigger than those proposed in this paper by construction. We recommend our test in the semi/non-parametric setting because (1) it requires no simulation, and thus are computationally easier, and (2) it is more powerful. Our asymptotic analysis is considerably more complex than hers due to the presence of infinite dimensional parameters. A few other papers in the literature also achieve uniform asymptotic size control. These include Li (2009), Schennach and Wilhelm (2016), Hsu and Shi (2014) 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 (2011) employ a sophisticated split-sample technique. Hsu and Shi (2014) 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 on nonnested cases, it can be applied to and has uniform asymptotic similarity in nested cases as well. In these cases, our test 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, 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 the Vuong's (1989) 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 test inference procedures. Section 7 shows Monte Carlo results of a mean-regression example. Section 8 applies the proposed uniform test and conditional confidence interval to a schooling choice example, and Section 9 concludes. Proofs of our main theorems 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, $\|A\|$ 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. Let $N(\mu, \Sigma)$ stand for a normal random vector with mean μ and variance-covariance matrix Σ. For any (possibly random) positive sequences $\{a_n\}_{n=1}^{\infty}$ 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 $1-p$ quantile of the standard normal distribution.

[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.

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 model comparison test 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:

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

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 $f(\mathcal{M}_j, F_0)$, $j = 1, 2$:

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)*

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

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

[4] A related but different fit measure is $f(\mathcal{M}_j, F_0) = E_{F_0}[m_j(Z, \alpha_j^\dagger)]$ for $\alpha_j^\dagger$ that is not a maximizer of $E_{F_0}[m_j(Z, \alpha_j)]$. Our analysis does not apply to this fit measure. A similar analysis may be done, but is left for future research.

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 this likelihood context if α for both models are finite-dimensional, although Shi (2015) shows that it may have size distortion. Shi 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*

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

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 (Absolute Deviation) *Consider $Z = (Y, X')'$, where Y is a dependent variable, X is a vector of regressors. A median-regression model may be written as*

$$\mathcal{M} = \{P : Q_{0.5,F}(Y|X = x) = g(x; \alpha),\ \forall x,\ \textit{for some } \alpha \in \mathcal{A}\}, \tag{2.8}$$

where $Q_{0.5,F}(Y|X)$ is the conditional median of Y given X under F, $g(\cdot, \cdot)$ 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 absolute deviation of Y from the best conditional median function in the model, which is $f(\mathcal{M}, F_0)$ defined in equation (2.3) with

$$m(Z, \alpha) = -|Y - g(X; \alpha)|/2. \tag{2.9}$$

[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.

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 $m_1(z; \alpha_1) = m_2(z; \alpha_2)$ $\forall$ $z \in \mathcal{Z}$.*

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]

The model relationship determines whether the random variable $m_1(Z; \alpha_1^*) - m_2(Z; \alpha_2^*)$ is always, never, or sometimes almost surely zero under H_0.[7] [8] For strictly nonnested cases, this variable is never almost surely zero. For nested models, this variable is almost surely zero under H_0 as long as α_1^* and α_2^* are the unique maximizers of $Em(Z_i; \alpha_1)$ and $Em(Z_i; \alpha_2)$.[9] For overlapping nonnested models, this variable may or may not be almost surely zero under H_0, depending on the *unknown* data distribution F_0. Note that to test H_0 in (2.1), we need to estimate $E[m_1(Z; \alpha_1^*) - m_2(Z; \alpha_2^*)]$, for which purpose, we need to estimate both the expectation $E[\cdot]$, and the pseudo-true values (α_1^*, α_2^*). The estimation errors of both parts should be taken into account when constructing a valid test. However, the relative asymptotic order (or finite sample magnitude) of the two estimation errors depends on the variance of $m_1(Z; \alpha_1^*) - m_2(Z; \alpha_2^*)$. This is because the asymptotic order of the estimation error of $E[\cdot]$ depends on this variance, while that of the error caused by estimating (α_1^*, α_2^*) does not. Thus, the relative asymptotic order is unknown in the overlapping nonnested case, which is the main challenge for constructing a uniformly asymptotically valid test, not to mention a uniformly asymptotically exact and similar one.

[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.

[9] Suppose not. That is, suppose $Em_1(Z_i; \alpha_1^*) = Em_2(Z_i; \alpha_2^*)$, but $m_1(Z_i; \alpha_1^*) \neq m_2(Z_i; \alpha_2^*)$ *a.s.*. By the definition of nestedness above, there exists $\alpha_1^{**} \in \mathcal{A}_1$ different from α_1^* such that $m_1(Z_i; \alpha_1^{**}) = m_2(Z_i; \alpha_2^*)$ *a.s.*. Then $Em_1(Z_i; \alpha_1^{**}) = Em_2(Z_i; \alpha_2^*) = Em_1(Z_i; \alpha_1^*)$. This contradicts the uniqueness of α_1^* as the maximizer of $Em_1(Z_i; \alpha_1)$.

As we will see, the test that we construct is valid uniformly over the data distribution in the overlapping nonnested case. It is also valid uniformly over all types of model relationship. Both types of uniformity are 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, nonuniform tests 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.

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 parameters, and use sieve M-estimator for estimation.[10] The specific procedure is explained now. For $j = 1, 2$, let $\mathcal{A}_{j,k_j}$ denote a finite dimensional approximation of the parameter space $\mathcal{A}_j$, which satisfies

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

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 basis functions, k_j is a positive integer which may diverge with the sample size n. We give examples of the basis functions in the illustration section below. We assume that the sieve coefficients $\beta_{k_j}^*$ are in the interior of their spaces $B_{k_j,j}$ for any k_j. In the rest of the paper, we write $\alpha_{k_j} = \alpha_{j,k_j}$ 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

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

[10] Many properties of the sieve M-estimator, including consistency (see, e.g., White and Wooldridge (1991)), rate of convergence (see, e.g., Shen and Wong (1994) and Chen and Shen (1998)), and asymptotic normality (see, e.g., Shen (1997) and Chen and Shen (1998)) are established in the literature. In addition to these properties, we also use a second-order expansion of the empirical criterion function in the sieve M-estimation. We derive this expansion for a mean-regression and a median-regression example in Supplemental Appendices C and D. Sufficient conditions for the second-order expansion in general semi/nonparametric model are also available upon request.

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

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

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 $\widehat{\alpha}_{\mathbf{k},n} = (\widehat{\alpha}_{k_1,n}, \widehat{\alpha}_{k_2,n})$.

Because 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

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

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

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

Then one would construct tests of the form: $\varphi_n^{V,\text{2-sided}}(p) = 1\{|T_n^V| > z_{p/2}\}$ or $\varphi_n^{V,\text{1-sided}}(p) = 1\{T_n^V > z_p\}$. 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 $n^{1/2}\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}) = n^{1/2}\bar{\ell}_n(\alpha^*_{F_0}) + o_p(1) \to_d N(0, \omega^2_{F_0,*})$ 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. [11]

The intuition of the failure of the naive extension test can be seen from the following heuristic second order expansion of the QLR statistic. Let $\ell_{\alpha,\mathbf{k}}(Z; \alpha)$ be the "score" function of $\ell(Z; \alpha_{\mathbf{k}})$

[11] A pretest for whether $\ell(\cdot; \alpha^*_{F_0}) = 0$ could be performed before the naive extension test. But the two-step procedure may (a) not be uniformly asymptotically valid if the pretest does not use a conservative critical value, or (2) not be powerful because the pretest makes rejection difficult.

evaluated at $\alpha \in \mathcal{A}$. Suppose for now that $\ell(Z;\alpha)$ is differentiable in α, we have[12] $\ell_{\alpha,\mathbf{k}}(Z;\alpha) = \frac{\partial \ell(Z;\alpha_{\mathbf{k}})}{\partial \beta_{\mathbf{k}}}\Big|_{\alpha_{\mathbf{k}}=\alpha}$.[13] Let $\bar{\ell}_{\alpha,\mathbf{k},n}(\alpha^*_{F_0}) = n^{-1/2}\sum_{i=1}^n \ell_{\alpha,\mathbf{k}}(Z_i;\alpha)$. Then a second order Taylor expansion of $\bar{\ell}_n(\alpha^*_{F_0})$ around $\widehat{\alpha}_{\mathbf{k},n}$ gives:

$$\begin{aligned} & n\left\{\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}) - E_{F_0}[\ell(Z;\alpha^*_{F_0})]\right\} \\ & \quad \approx n\left\{\bar{\ell}_n(\alpha^*_{F_0}) - E_{F_0}[\ell(Z;\alpha^*_{F_0})]\right\} - (n/2)\bar{\ell}_{\alpha,\mathbf{k},n}(\alpha^*_{F_0})' H^{-1}_{F_0,\mathbf{k}} \bar{\ell}_{\alpha,\mathbf{k},n}(\alpha^*_{F_0}), \end{aligned} \tag{3.7}$$

where $H_{F,\mathbf{k}}(\alpha_{\mathbf{k}}) = \frac{\partial^2 E_F[\ell(Z;\alpha_{\mathbf{k}})]}{\partial \beta_{\mathbf{k}} \partial \beta'_{\mathbf{k}}}\Big|_{\alpha_{\mathbf{k}}=\alpha^*_F}$. Appropriate conditons and the central limit theorem imply that $n^{1/2}\left\{\bar{\ell}_n(\alpha^*_{F_0}) - E_{F_0}[\ell(Z;\alpha^*_{F_0})]\right\} \to_d N(0,\omega^2_{F_0,*})$, and $n\bar{\ell}_{\alpha,\mathbf{k},n}(\alpha^*_{F_0}) \to_d N(0, D_{F_0,\mathbf{k}})$, where

$$D_{F,\mathbf{k}} = E_F[\ell_{\alpha,\mathbf{k}}(Z;\alpha^*_F)\ell_{\alpha,\mathbf{k}}(Z;\alpha^*_F)']. \tag{3.8}$$

The latter implies that $n\bar{\ell}_{\alpha,n}(\alpha^*_{F_0})' H^{-1}_{F_0,\mathbf{k}} \bar{\ell}_{\alpha,n}(\alpha^*_{F_0})$ is approximately $\sum_{j=1}^{|\mathbf{k}|} \lambda_j \chi^2_j$, where $|\mathbf{k}| = k_1 + k_2$, $\{\chi^2_j\}_{j=1}^{|\mathbf{k}|}$ 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,

$$n\left\{\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}) - E_{F_0}[\ell(Z;\alpha^*_{F_0})]\right\} \approx n\left\{\bar{\ell}_n(\alpha^*_{F_0}) - E_{F_0}[\ell(Z;\alpha^*_{F_0})]\right\} - \sum_{j=1}^{|\mathbf{k}|} \lambda_j \chi^2_j/2, \tag{3.9}$$

In particular, the mean of $\sum_{j=1}^{|\mathbf{k}|} \lambda_j \chi^2_j$ is $tr(D_{F_0,\mathbf{k}} H^{-1}_{F_0,\mathbf{k}})$, which is typically nonzero, and can be of comparable scale as $\omega_{F_0,*}$, the standard deviation of $n\bar{\ell}_n(\alpha^*_{F_0})$. This means that, even when H_0 hold ($E_{F_0}[\ell(Z;\alpha^*_{F_0})] = 0$), 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 $n\widehat{\omega}_n(\widehat{\alpha}_{\mathbf{k},n})^2$ is a biased estimator of $\omega^2_{F_0,*}$ as well, and the dominating term of the bias is coincidentally $\sum_{j=1}^{|\mathbf{k}|} \lambda^2_j$. 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 uniform model selection test corrects the two biases by estimating and removing them. From the heuristic discussion above, we see that it is sufficient to estimate $H_{F_0,\mathbf{k}}$ and $D_{F_0,\mathbf{k}}$. We

[12] This definition does not require that the evaluation point α be in the sieve space. Because we use linear sieve, $\partial \ell(Z;\alpha_{\mathbf{k}}(\beta_{\mathbf{k}}))/\partial \beta_{\mathbf{k}}$ depends on $\beta_{\mathbf{k}}$ only through $\alpha_{\mathbf{k}}(\beta_{\mathbf{k}})$. Thus, it makes sense to consider $\partial \ell(Z;\alpha_{\mathbf{k}}(\beta_{\mathbf{k}}))/\partial \beta_{\mathbf{k}}$ as a function of $\alpha_{\mathbf{k}}(\beta_{\mathbf{k}})$ on $\mathcal{A}_{\mathbf{k}}$. As such, the function can be extended to the whole space $\mathcal{A}$ and can be evaluated at any point on $\mathcal{A}$. See equation (5.5) below for an example. This comment applies to $H_{F_0,\mathbf{k}}$ defined below as well.

[13] The form of $\ell_{\alpha,\mathbf{k}}(Z;\alpha)$ in the median-regression example is available in Supplemental Appendix D.

let

$$\widehat{D}_n = n^{-1} \sum_{i=1}^{n} \ell_{\alpha,\mathbf{k}}(Z_i; \widehat{\alpha}_{\mathbf{k},n}) \ell_{\alpha,\mathbf{k}}(Z_i; \widehat{\alpha}_{\mathbf{k},n})'. \tag{3.10}$$

The second derivative matrix $H_{F_0,\mathbf{k}}$ can be estimated by

$$\widehat{H}_n = n^{-1} \sum_{i=1}^{n} \frac{\partial^2 \ell(Z_i; \widehat{\alpha}_{\mathbf{k},n})}{\partial \beta_{\mathbf{k}} \partial \beta_{\mathbf{k}}'} \tag{3.11}$$

when $\ell(Z; \cdot)$ is differentiable. When $\ell(Z; \cdot)$ is not differentiable, it is useful to note that $H_{F_0,\mathbf{k}}$ is a block diagonal matrix:

$$H_{F_0,\mathbf{k}}(\alpha_{\mathbf{k}}) \equiv \left. \frac{\partial^2 E_{F_0}[\ell(Z; \alpha_{\mathbf{k}})]}{\partial \beta_{\mathbf{k}} \partial \beta_{\mathbf{k}}'} \right|_{\alpha_{\mathbf{k}} = \alpha^*_{F_0}} = \begin{pmatrix} H_{F_0,k_1} & 0 \\ 0 & -H_{F_0,k_2} \end{pmatrix}. \tag{3.12}$$

where H_{F_0,k_j} is the Hessian matrix of model j: $H_{F_0,k_j} \equiv \left. \frac{\partial^2 E_{F_0}[m_j(Z;\alpha_{k_j})]}{\partial \beta_{k_j} \partial \beta_{k_j}'} \right|_{\alpha_{k_j} = \alpha^*_{F_0,j}}$. Hessian matrices for non-smooth M-estimation problems are available case by case in the literature. For example a suitable choice of $\widehat{H}_n$ in the median-regression example is available in Belloni et al. (2011)

With the estimators $\widehat{H}_n$ and $\widehat{D}_n$, we can construct a test statistic:

$$T_n^0 = \frac{n\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}) + 2^{-1} tr(\widehat{D}_n \widehat{H}_n^{-1})}{\sqrt{n\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k},n}) - 2^{-1} tr((\widehat{D}_n \widehat{H}_n^{-1})^2)}}. \tag{3.13}$$

We formally show in the next section that $T_n^0 \to_d N(0,1)$ under H_0 and regularity conditions. Intuitively, the standard normality comes from the asymptotic normality of the demeaned mixed chi-squared sequence $\sum_{j=1}^{|\mathbf{k}|} \lambda_j \chi_j^2$ and the asymptotic independence between this sequence and $n\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}) - E_{F_0}[\ell(Z; \alpha^*_{F_0})]$.

Result (3.13) holds regardless of the relationship between $m_1(\cdot, \alpha^*_{F_0,1})$ and $m_2(\cdot, \alpha^*_{F_0,2})$, and thus holds uniformly over the unknown data distribution when the models are overlapping nonnested, and uniformly over all types of model relationships when the relationship is unknown. As a consequence, a test based on T_n^0 is uniformly asymptotically exact and similar. However, there is one minor issue left. Note that the denominator involves the difference of two estimated quantities. In a finite sample, the difference can turn out to be zero or negative, even though the probability of that goes to zero asymptotically. To avoid division by zero or by the square root of a negative number or zero, we recommend a slight regularization:

$$T_n = \frac{n\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}) + 2^{-1} tr(\widehat{D}_n \widehat{H}_n^{-1})}{n^{1/2} \widehat{\sigma}_n}, \tag{3.14}$$

where

$$\widehat{\sigma}_n^2 = \max\{\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k},n}) - (2n)^{-1}tr((\widehat{D}_n\widehat{H}_n^{-1})^2), (2n)^{-1}tr((\widehat{D}_n\widehat{H}_n^{-1})^2)\}. \tag{3.15}$$

Hence $\widehat{\sigma}_n^2$ is a regularized version of the difference in the denominator T_n^0. The regularization term $(2n)^{-1}tr((\widehat{D}_n\widehat{H}_n^{-1})^2)$ is also a consistent estimator of the variance of $n^{1/2}\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}) + 2^{-1}n^{-1/2}tr(\widehat{D}_n\widehat{H}_n^{-1})$ in the extreme case that $\mathcal{M}_1$ and $\mathcal{M}_2$ are nested, and is asymptotically less than this variance in nonnested cases. Therefore, the modification does not affect the asymptotic limit of the denominator.

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

$$\varphi_n^{\text{2-sided}}(p) = 1\{|T_n| > z_{p/2}\}. \tag{3.16}$$

Analogously, the one-sided test of H_0 against $H_1^{\text{1-sided}}$ in (2.2) is

$$\varphi_n^{\text{1-sided}}(p) = 1\{T_n > z_p\}. \tag{3.17}$$

The test does not favor one model or the other when it does not reject the null hypothesis. The 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:

$$\lim_{n\to\infty} \inf_{F_0\in\mathcal{F}_0} E_{F_0}[\varphi_n(p)] = p, \tag{3.18}$$

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. In fact, it also is shown that

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

which combined with (3.18) shows that the tests proposed are asymptotically 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 mean-regression example and in a median-regression example in Supplemental Appendices C and D.

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 a sequence of 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 $\beta_{\mathbf{k}}$ on $B_{\mathbf{k}}$;*

(c) *for every $\mathbf{k}$, there exists $\alpha^*_{F_0}$ such that $E_{F_0}\left[\ell_{\alpha,\mathbf{k}}(Z,\alpha^*_{F_0})\right] = \mathbf{0}_{|\mathbf{k}|}$;*

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

(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) *$C^{-1} \leq \rho_{\min}(H_{F_0,\mathbf{k}}) \leq \rho_{\max}(H_{F_0,\mathbf{k}}) \leq C$ and $\rho_{\max}(D_{F_0,\mathbf{k}}) \leq C$ for all $\mathbf{k}$;*

(g) *H_{F_0,k_1} and H_{F_0,k_2} are negative definite.*

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 (3.12) is well defined. Assumption 4.1(c) generally follows from the first order optimality condition of α^*. 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 $\omega^2_{F_0,U,\mathbf{k}} \equiv \sum_{j=1}^{|\mathbf{k}|} \lambda^2_{F_0,j} \equiv tr((D_{F_0,\mathbf{k}} H^{-1}_{F_0,\mathbf{k}})^2)$. Assumption 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 sequence $\xi_{\mathbf{k}}$ depends on the models as well as the basis function used. For example, in the mean-regression example studied in later sections, it is the order of $|\mathbf{k}|^2$ if power series is used, and it is the order of $|\mathbf{k}|$ 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 (3.18) and (3.19), it is sufficient to consider all sequences of DGPs $\{F_n\}_{n\geq 1}$ in $\mathcal{F}$. For any $F_n \in \mathcal{F}$, we let $\alpha^*_{j,n}$ abbreviate α^*_{j,F_n}, and let α^*_n abbreviate $\left(\alpha^*_{1,n}, \alpha^*_{2,n}\right)$. Let $\overline{\ell}_{\alpha,n}(\alpha) = n^{-1}\sum_{i=1}^n \ell_{\alpha,\mathbf{k}}(Z_i;\alpha)$ 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) *$\overline{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}) = n^{-1}\sum_{i=1}^n \ell(Z_i;\alpha^*_n) - 2^{-1}\overline{\ell}_{\alpha,n}(\alpha^*_n)' H^{-1}_{F_n,\mathbf{k}} \overline{\ell}_{\alpha,n}(\alpha^*_n) + o_p(n^{-1/2}\sigma_{F_n,n})$;*

(b) *$\frac{1}{n\sigma^2_{F_n,n}} = o(1)$ and $\frac{|\mathbf{k}|\xi_{\mathbf{k}}}{n^2\sigma^2_{F_n,n}} = o(1)$.*

Assumption 4.3(a) is a second order expansion of $\overline{\ell}_n(\widehat{\alpha}_{\mathbf{k},n})$ around α^*_n. We illustrate this expansion in the mean-squared error example (Section 5) and the mean absolute deviation example

(Supplemental Appendix D).[14] With the formula of this expansion, we can add more details to the heuristic discussion in Section 2.2. The variance of the leading term, $n^{-1}\omega^2_{F_n,*}$, in the expansion comes from estimating the expectation, and the variance of the second term, approximately $2^{-1}n^{-2}\omega^2_{F_n,U,\mathbf{k}}$, comes from estimating α^*_n. The quantity $\omega^2_{F_n,*}$ 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 $\omega^2_{F_n,U,\mathbf{k}}$ typically is nonzero.[15] The relative magnitude of the two terms is proportional to $\frac{n\omega^2_{F_n,*}}{\omega^2_{F_n,U,\mathbf{k}}}$, 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.[16]

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 $2^{-1}\overline{\ell}_{\alpha,n}(\alpha^*_n)'H^{-1}_{F_n,\mathbf{k}}\overline{\ell}_{\alpha,n}(\alpha^*_n)$. To understand this assumption, note that $\sigma^2_{F_n,n} = \omega^2_{F_n,*} + (2n^2)^{-1}(n-1)\omega^2_{F_n,U,\mathbf{k}}$. If $\omega^2_{F_n,*}$ is bounded below by a positive constant (as is typical for strictly nonnested models), Assumption 4.3(b) is satisfied as long as $|\mathbf{k}|\xi_{\mathbf{k}}/n^2 = o(1)$. Otherwise, Assumption 4.3(b) imposes restriction on the U-statistic variance. Specifically, it requires, as $n \to \infty$, that

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

Recall that $\omega^2_{F_n,U,\mathbf{k}} = tr\left((H^{-1}_{F_n,\mathbf{k}}D_{F_n,\mathbf{k}})^2\right)$. Thus, what Assumption 4.3(b) requires are that $|\mathbf{k}|$ grows with n, and that there are not too many zero eigenvalues for the matrix $H^{-1}_{F_n,\mathbf{k}}D_{F_n,\mathbf{k}}$. Both can be assessed in practice because $\mathbf{k}$ is user-chosen and $H^{-1}_{F_n,\mathbf{k}}D_{F_n,\mathbf{k}}$ can be consistently estimated. Moreover, the requirement that $|\mathbf{k}|$ grows with n is natural and necessary in the literature of series estimation of semi/nonparametric models.[17]

Under the above assumptions, the following intermediate result holds.

[14] Sufficient conditions for Assumption 4.3(a) in general semi/nonparametric model are also available upon request.

[15] 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, $2^{-1}\overline{\ell}_{\alpha,n}(\alpha^*_n)'H^{-1}_{F_n,\mathbf{k}}\overline{\ell}_{\alpha,n}(\alpha^*_n) = 2^{-1}n^{-2}\sum_{i=1}^n\sum_{j=1}^n u_iu_jX_{2,i}'X_{2,j}$ which is clearly not degenerate. See Hong and White (1995) for more sophisticated examples.

[16] 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.

[17] 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 (which are not included in the paper but available upon request) show that our test works well even when $|\mathbf{k}|$ is only 10.

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*

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

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

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* $\sqrt{n}\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n})$ *is non-degenerate (*$F_n = F$ *for some* F *and for all* n*, and* $\omega_{F,*}^2 > 0$*);* (ii) *the statistic* $\sqrt{n}\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n})$ *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.*

Remark 2 *The asymptotic normal distribution in Theorem 4.1 makes both the inference of model comparison and the post model selection inference easy in practice. An alternative method is to generalize the fixed* $|\mathbf{k}|$ *asymptotic theory in Shi (2015) and then conduct inference using the approach provided in that paper. This may have the advantage of possibly not requiring Assumption 4.3(b). However, when* $|\mathbf{k}|$ *is large, there are a large number of nuisance parameters for Shi's approach to consider, which makes it difficult to use. It is also less powerful than the test proposed in this paper because her test statistic is smaller and critical value bigger by construction.*

In order to use the intermediate result in Theorem 4.1, we need to construct consistent estimators of $\widehat{D}_n(\alpha_n^*)$, $H_{F_n,\mathbf{k}}$, 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) $\|\widehat{H}_n - H_{F_n,\mathbf{k}}\| = o_p(\delta_n)$, $\|\widehat{D}_n - \widehat{D}_n(\alpha_n^*)\| = o_p(\delta_n)$ *and* $\|\widehat{D}_n(\alpha_n^*) - D_{F_n,\mathbf{k}}\| = o_p(\delta_n)$*;*
(b) $n^{-1}\sum_{i=1}^n |\ell(Z_i,\widehat{\alpha}_n) - \ell(Z_i,\alpha_n^*)|^2 = \bar{\ell}_{\alpha,n}(\alpha_n^*)'(H_{F_n,\mathbf{k}}^{-1}D_{F_n,\mathbf{k}}H_{F_n,\mathbf{k}}^{-1})\bar{\ell}_{\alpha,n}(\alpha_n^*) + o_p(\sigma_{F_n,n}^2)$*;*
(c) $n^{-1}\sum_{i=1}^n(\ell(Z_i,\alpha_n^*) - \ell_{F_n}(\alpha_n^*))\left[\ell(Z_i,\widehat{\alpha}_n) - \ell(Z_i,\alpha_n^*)\right] = o_p(\sigma_{F_n,n}^2)$*;*
(d) $|\mathbf{k}|\, n^{-1} = o(1)$.

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*

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

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

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

Remark 3 *Note that* $\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k},n})$ *may be viewed as a sample-analogue estimator of* $\omega_{F_n,*}^2$. *Lemma* 4.2 *shows that, in general,* $\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k},n})$ *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* $2^{-1}n^{-2}(n+1)\omega_{F_n,U,\mathbf{k}}^2$. *The upward bias is due to the estimation error in* $\widehat{\alpha}_{\mathbf{k},n}$.

Lemma 4.2 suggests that $\sigma_{F_n,n}^2$ can be consistently estimated by estimating and then removing the large-sample bias $2^{-1}n^{-2}(n+1)\omega_{F_n,U,\mathbf{k}}^2$ from $\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k},n})$. 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) $n^{-1}\left[tr((\widehat{D}_n\widehat{H}_n^{-1})^2) - \omega_{F_n,U,\mathbf{k}}^2\right] = o_p(\sigma_{F_n,n}^2)$, *and*

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

In the test statistic defined in (3.14), we use $\widehat{\sigma}_n^2$, instead of $\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k},n}) - 2^{-1}n^{-1}tr((\widehat{D}_n\widehat{H}_n^{-1})^2)$, to estimate $\sigma_{F_n,n}^2$. This is because the difference $\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k},n}) - 2^{-1}n^{-1}tr((\widehat{D}_n\widehat{H}_n^{-1})^2)$ may be equal to or less than zero in finite samples, which is not a desirable property for a variance estimator. To avoid estimating variance by a nonpositive number, we let $\widehat{\sigma}_n^2$ be simply $\max\{\widehat{\omega}_n^2(\widehat{\alpha}_{\mathbf{k},n}) - (2n)^{-1}tr((\widehat{D}_n\widehat{H}_n^{-1})^2), (2n)^{-1}tr((\widehat{D}_n\widehat{H}_n^{-1})^2)\}$. This modification does not affect consistency because using Lemma 4.3, we deduce that, for any $\varepsilon > 0$,

$$\Pr{}_{F_n}((2n)^{-1}tr((\widehat{D}_n\widehat{H}_n^{-1})^2) - \sigma_{F_n,n}^2 > \varepsilon) = \Pr{}_{F_n}(-\omega_{F_n,*}^2 + o_p(\sigma_{F_n,n}^2) > \varepsilon) \to 0, \qquad (4.4)$$

where the equality holds by the definition of $\sigma_{F_n,n}^2$, and the convergence holds by Assumption 4.3(b).

Theorem 4.1 and Lemmas 4.1–4.3 immediately lead to the uniform asymptotic size control and asymptotic similarity result for the tests proposed in the previous section. 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) *Both* (3.18) *and* (3.19) *hold 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* $\sqrt{n}E_{F_n}\ell(Z;\alpha_n^*)/\sigma_{F_n,n} \to c$ *for some constant* $c \in R$, *we have*

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

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

Remark 4 *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* $\sqrt{n}$, *and against alternatives with* $E_{F_n}\ell(Z;\alpha_n^*)$ *converging to 0 at a rate faster than* $\sqrt{n}$ *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 argument the variance either through sample splitting or other means.*

5 Illustration: Nonparametric Mean-Regression

In this section we illustrate the construction of our test using the mean-regression example. The verifications of the high-level assumptions in this example are in Supplemental Appendix C. Another illustrating example—median-regression—is given in Supplemental Appendix D.

For $j = 1, 2$, let model j be

$$\{F : E_F[Y - \alpha_j(X_j)|X_j] = 0, \ \alpha_j \in \mathcal{A}_j\}, \tag{5.1}$$

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 $\alpha_1(X_1) = \alpha_2(X_2) = Constant$.[18]

[18] A restriction on the parameter space that rules out constant regression functions would make models with strictly nonnested sets of regressors strictly nonnested, which might give false hope for the validity of the naive extension test. The hope is false because, although any artificial bound from constancy makes the models strictly nonnested, they may not change the performance of the tests in finite samples when the bounds are not far enough from constancy. In any given finite sample, it is difficult to know what bounds are far enough.

The model (5.1) 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 the correct specification of the models, or in other words, 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 basis function rfor this case has to do with the structure of $\mathcal{A}_j$. For example, suppose that we have a partial linear model $\alpha_j(X_j) = \beta_0 + \beta_1 X_{j,1} + g(X_{j,2})$. Then, we should let $P_{j,k_j}(X_j) = (p_{j,1}(X_j), p_{j,2}(X_j), \ldots, p_{j,k_j}(X_j))'$ such that $p_{j,1}(X_j) = 1$, $p_{j,2}(X_j) = X_{j,1}$, 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 a spline series on the space of $X_{j,2}$.

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

$$\widehat{\alpha}_{k_j,n}(\cdot) = P_{j,k_j}(\cdot)'\widehat{\beta}_{k_j,n} \text{ with } \widehat{\beta}_{k_j,n} = (\mathbf{P}'_{j,k_j,n}\mathbf{P}_{j,k_j,n})^{-1}\mathbf{P}'_{j,k_j,n}\mathbf{Y}_n, \tag{5.2}$$

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

$$\ell(Z;\alpha) = -|Y - \alpha_1(X_1)|^2/2 + |Y - \alpha_2(X_2)|^2/2. \tag{5.3}$$

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

The pseudo-true value of the parameter can be written as the limit of a sequence of sieve approxiamtion: $\alpha_j^*(x_j) = \sum_{l=1}^{\infty} p_{j,l}(x_j)\beta_{j,l}^*$, where $(\beta_{j,l}^*)_{\ell=1}^{\infty} = \underset{\beta_{j,l}\in R,\forall l}{\arg\min} E_{F_0}\left[|Y - \sum_{l=1}^{\infty} p_{j,l}(x_j)\beta_{j,l}|^2\right]$. Let $u_j = Y - \alpha_j^*(X_j)$. The definition of $\alpha_j^*(\cdot)$ implies the following first order condition

$$E_{F_0}\left[u_j p_{j,l}(X_j)\right] = 0, \tag{5.4}$$

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:

$$\ell_{\alpha,\mathbf{k}}(Z;\alpha) = \begin{pmatrix} (Y - \alpha_1(X_1))P_{1,k_1}(X_1) \\ -(Y - \alpha_2(X_2))P_{2,k_2}(X_2) \end{pmatrix}. \tag{5.5}$$

Then, the expectation of the outer product of the score function evaluated at α^* is

$$D_{F_0,\mathbf{k}} = \begin{pmatrix} E_{F_0}[u_1^2 P_{1,k_1}(X_1)P_{1,k_1}(X_1)'] & -E_{F_0}[u_1u_2 P_{1,k_1}(X_1)P_{2,k_2}(X_2)'] \\ -E_{F_0}[u_1u_2 P_{2,k_2}(X_2)P_{1,k_1}(X_1)'] & E_{F_0}[u_2^2 P_{2,k_2}(X_2)P_{2,k_2}(X_2)'] \end{pmatrix}, \tag{5.6}$$

and the population Hessian matrix is:

$$H_{F_0,\mathbf{k}} = \begin{pmatrix} -E_{F_0}[P_{1,k_1}(X_1)P_{1,k_1}(X_1)'] & \mathbf{0}_{k_1\times k_2} \\ \mathbf{0}_{k_2\times k_1} & E_{F_0}[P_{2,k_2}(X_2)P_{2,k_2}(X_2)'] \end{pmatrix}. \tag{5.7}$$

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

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

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

$$\widehat{H}_{n,\mathbf{k}} = \begin{pmatrix} -n^{-1}\sum_{i=1}^{n} P_{1,k_1}\left(X_{1,i}\right) P_{1,k_1}(X_{1,i})' & \mathbf{0}_{k_1\times k_2} \\ \mathbf{0}_{k_2\times k_1} & n^{-1}\sum_{i=1}^{n} P_{2,k_2}\left(X_{2,i}\right) P_{2,k_2}\left(X_{2,i}\right)' \end{pmatrix}. \tag{5.9}$$

Finally, the test statistic may be constructed easily using the above quantities following (3.14) and (3.15).

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 $\psi_1(\widehat{\alpha}_{k_1,n})$, and $\psi_2(\widehat{\alpha}_{k_2,n})$. Both subsections 6.2 and 6.3 rely on the joint normal limiting distribution of $(\psi_1(\widehat{\alpha}_{k_1,n}), \psi_2(\widehat{\alpha}_{k_2,n}), \bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}))'$ (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 $D_{F_0,k_j} = E[\ell_{\alpha,k_j}(Z;\alpha_j^*)\ell_{\alpha,k_j}(Z;\alpha_j^*)']$ for $j = 1, 2$. Also define

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

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

We will first derive the asymptotic distribution of

$$G_{n,F_0} \equiv \begin{pmatrix} \frac{n\left[\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}) - E_{F_0}[\ell(Z;\alpha^*)]\right] + (1/2)tr(\widehat{D}_n(\alpha^*)H^{-1}_{F_0,\mathbf{k}})}{n^{1/2}\sigma_{F_0,n}} \\ \frac{n^{1/2}\left[\psi_1(\widehat{\alpha}_{1,n}) - \psi_1(\alpha_1^*)\right]}{v^*_{\psi,k_1}} \\ \frac{n^{1/2}\left[\psi_2(\widehat{\alpha}_{2,n}) - \psi_2(\alpha_2^*)\right]}{v^*_{\psi,k_2}} \end{pmatrix}. \tag{6.3}$$

Finally, define the correlation coefficients

$$\rho_{0j,F_0} = \frac{\psi_{\alpha,k_j}(\alpha_j^*)' H_{F_0,k_j}^{-1}}{v_{\psi,k_j}^* \sigma_{F_0,n}} E_{F_0}\left[\ell_{\alpha,k_j}(Z;\alpha_j^*)\ell(Z;\alpha^*)\right] \text{ for } j = 1, 2,$$

$$\rho_{12,F_0} = \frac{\psi_{\alpha,k_1}(\alpha_1^*)' H_{F_0,k_1}^{-1} D_{F_0,k_1,k_2} H_{F_n,k_2}^{-1} \psi_{\alpha,k_2}(\alpha_2^*)}{v_{\psi,k_1}^* v_{\psi,k_2}^*}, \tag{6.4}$$

where $D_{F_0,k_1,k_2} = E_{F_0}[\ell_{\alpha,k_1}(Z;\alpha_1^*)\ell_{\alpha,k_2}(Z;\alpha_2^*)']$.

We make the following assumptions regarding the plug-in estimator $\psi_j(\widehat{\alpha}_{k_j,n})$ for $j = 1, 2$. Sufficient Conditions can be found in Chen et al. (2014) and Chen and Liao (2015).

Assumption 6.1 *Under any sequence $\{F_n\}_{n\geq 1}$ such that $F_n \in \mathcal{F}$ for all n, we have for $j = 1, 2$,*

$$\frac{n^{\frac{1}{2}}\left[\psi_j(\widehat{\alpha}_{j,n}) - \psi_j(\alpha_{j,n}^*)\right]}{v_{\psi,k_j}^*} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\psi_{\alpha,k_j}(\alpha_{j,n}^*)' H_{F_n,k_j}^{-1}}{v_{\psi,k_j}^*} \ell_{\alpha,k_j}(Z_i;\alpha_{j,n}^*) + o_p(1) \tag{6.5}$$

and $E_{F_n}\left[\left|\psi_{\alpha,k_j}(\alpha_j^*)' H_{F_n,k_j}^{-1} \ell_{\alpha,k_j}(Z_i;\alpha_j^*) \Big/ v_{\psi,k_j}^*\right|^4\right] = o(n)$.

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 G_{n,F_n} under an arbitrary sequence $F_n \in \mathcal{F}_0$. The proof of the lemma extends that for the results in Section 4 to joint convergence and is relegated to Supplemental Appendix E.

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

$$G_{u_n,F_{u_n}} \to_d N\left(\mathbf{0}_3, \Sigma_G\right), \text{ where } \Sigma_G = \begin{pmatrix} 1 & \rho_{01} & \rho_{02} \\ \rho_{01} & 1 & \rho_{12} \\ \rho_{02} & \rho_{12} & 1 \end{pmatrix}.$$

Lemma 6.1 cannot be used directly in practice not only because because the limit depends on the unknown nuisance parameter Σ_G but also because G_{n,F_n} itself involves the unknown quantites $\sigma_{F_n,n}^2$, $\widehat{D}_n(\alpha_n^*)H_{F_n,\mathbf{k}}^{-1}$, and v_{ψ,k_j}^{*2}. The consistent estimators of $\sigma_{F_n,n}^2$ and $\widehat{D}_n(\alpha^*)H_{F_n,\mathbf{k}}^{-1}$ have already been given in the previous section. Consistent estimators of v_{ψ,k_j}^{*2}, $\rho_{0j,n}$, and $\rho_{12,n}$ can be constructed using their sample analogs:

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

$$
\begin{aligned}
\widehat{\rho}_{0j,n} &= \frac{\psi_{\alpha,k_j}(\widehat{\alpha}_{j,n})'\widehat{H}_{k_j,n}^{-1}}{n\widehat{v}_{\psi,k_j}^{*}\widehat{\sigma}_n}\sum_{i=1}^{n}\ell_{\alpha,k_j}(Z_i;\widehat{\alpha}_{j,n})\ell(Z_i;\widehat{\alpha}_n), \text{ for } j=1,2\ ; \\
\widehat{\rho}_{12,n} &= \frac{\psi_{\alpha,k_1}(\widehat{\alpha}_{1,n})'\widehat{H}_{k_1,n}^{-1}\widehat{D}_{k_1,k_2,n}\widehat{H}_{k_2,n}^{-1}\psi_{\alpha,k_2}(\widehat{\alpha}_{2,n})}{\widehat{v}_{\psi,k_1}^{*}\widehat{v}_{\psi,k_2}^{*}},
\end{aligned}
\tag{6.6}
$$

where $\widehat{D}_{k_1,k_2,n} = n^{-1}\sum_{i=1}^{n}\ell_{\alpha,k_1}(Z_i;\widehat{\alpha}_{1,n})\ell_{\alpha,k_2}(Z_i;\widehat{\alpha}_{2,n})'$.

Let $\ell_{\alpha\alpha,k_j}(Z;\alpha_{j,n}^{*}) = \left.\frac{\partial\ell_{\alpha,k_j}(Z;\alpha_j)}{\partial\beta_{k_j}'}\right|_{\alpha_j=\alpha_{j,n}^{*}}$. Define

$$
D_{\ell,k_j,n} = E_{F_n}\left[\ell^2(Z;\alpha_n^{*})\ell_{\alpha,k_j}(Z;\alpha_{j,n}^{*})\ell_{\alpha,k_j}(Z;\alpha_{j,n}^{*})'\right] \text{ and } H_{\ell,k_j,n} = E_{F_n}\left[\ell(Z;\alpha_n^{*})\ell_{\alpha\alpha,k_j}(Z;\alpha_{j,n}^{*})\right].
\tag{6.7}
$$

The following conditions are needed for showing the consistency of $\widehat{v}_{\psi,k_j}^{*2}$, $\widehat{\rho}_{0j,n}$, and $\widehat{\rho}_{12,n}$.

Assumption 6.2 *For any $\{F_n\}_{n\geq 1}$ such that $F_n \in \mathcal{F}$ for all n, we have for $j=1,2$:*

(a) $\frac{1}{v_{\psi,k_j}^{*}}\left\|\psi_{\alpha,k_j}(\widehat{\alpha}_{j,n}) - \psi_{\alpha,k_j}(\alpha_{j,n}^{*})\right\| = o_p(|\mathbf{k}|^{-1/2})$;

(b) $\rho_{\max}(D_{\ell,k_j,n}) \leq C$ *and* $\rho_{\max}(H_{\ell,k_j,n}) \leq C$;

(c) *the following expansion holds*

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\left[\ell_{\alpha,k_j}(Z_i;\widehat{\alpha}_{j,n})\ell(Z_i;\widehat{\alpha}_n) - \ell_{\alpha,k_j}(Z_i;\alpha_{j,n}^{*})\ell(Z_i;\alpha_n^{*})\right] \\
&= n^{-1}\sum_{i=1}^{n}\ell_{\alpha,k_j}(Z_i;\alpha_{j,n}^{*})\ell_{\alpha,\mathbf{k}}(Z_i;\alpha_n^{*})'H_{F_n,\mathbf{k}}^{-1}\overline{\ell}_{\alpha,n}(\alpha_n^{*}) \\
&+ n^{-1}\sum_{i=1}^{n}\ell(Z_i;\alpha_n^{*})\ell_{\alpha\alpha,k_j}(Z_i;\alpha_{j,n}^{*})H_{F_n,k_j}^{-1}\overline{\ell}_{\alpha_j,n}(\alpha_{j,n}^{*}) + o_p(\sigma_{F_n,n});
\end{aligned}
$$

(d) $\left\|n^{-1}\sum_{i=1}^{n}\ell(Z_i;\alpha_n^{*})\ell_{\alpha\alpha,k_j}(Z_i;\alpha_{j,n}^{*}) - H_{\ell,k_j,n}\right\| = o_p(|\mathbf{k}|^{-1/2})$;

(e) $\left\|\psi_{\alpha,k_j}(\alpha_{j,n}^{*})/v_{\psi,k_j}^{*}\right\| \leq C$;

(f) $\|\widehat{H}_n - H_{F_n,\mathbf{k}}\| = o_p(|\mathbf{k}|^{-1/2})$ *and* $\|\widehat{D}_n - D_{F_n,\mathbf{k}}\| = o_p(|\mathbf{k}|^{-1/2})$.

The following lemma shows the consistency of the estimated variances and correlations. Like Lemma 6.1, the proof of this lemma is also relegated to Supplemental Appendix E.

Lemma 6.2 *Suppose that Assumptions* 4.1(b)-(g), 4.3(b), *and* 6.2 *hold. Then under all sequences $\{F_n\}_{n\geq 1}$ such that $F_n \in \mathcal{F}$ for all n, we have*

(a) $\max_{j=1,2}\left[\widehat{v}_{\psi,k_j}^{*2}\Big/ v_{\psi,k_j}^{*2} - 1\right] = o_p(1)$;

(b) $\widehat{\rho}_{12,n} - \rho_{12,n} = o_p(1)$;

(c) $\max_{j=1,2}|\widehat{\rho}_{0j,n} - \rho_{0j,n}| = o_p(1)$.

Lemmas 4.2, 4.3, 6.1, and 6.2 together immediately imply the following result, the proof of which is omitted.

Lemma 6.3 *Suppose that Assumptions* 4.1, 4.3, *and* 6.1-6.2 *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*

$$\widehat{G}_{n,F_n} \equiv \left(\begin{array}{c} \frac{n\left[\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}) - E_{F_n}\ell(Z;\alpha_n^*)\right] + (1/2)tr(\widehat{D}_n \widehat{H}_n^{-1})}{n^{1/2}\widehat{\sigma}_n} \\ \frac{n^{1/2}\left[\psi_1(\widehat{\alpha}_{1,n}) - \psi_1(\alpha_{1,n}^*)\right]}{\widehat{v}_{\psi,k_1}^*} \\ \frac{n^{1/2}\left[\psi_2(\widehat{\alpha}_{2,n}) - \psi_2(\alpha_{2,n}^*)\right]}{\widehat{v}_{\psi,k_2}^*} \end{array} \right) \to_d N(\mathbf{0}_3, \Sigma_G).$$

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.[19] Specifically, we construct a level $1-p$ conditional confidence interval, $CI_{\psi_1}(1-p)$ such that

$$\liminf_{n\to\infty} \inf_{F_0 \in \mathcal{F}_n} \Pr{}_{F_0}(\psi_1(\alpha_1^*) \in CI_{\psi_1}(1-p) | T_n \geq c) = 1-p, \tag{6.8}$$

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 $\Psi : R \times (-\infty, \infty] \times [-1,1] \to R$:

$$\Psi(\theta, t, \rho) = \begin{cases} \frac{\Phi(\theta) - \Phi\left(\theta - \frac{t}{\rho}\right)}{1 - \Phi\left(\theta - \frac{t}{\rho}\right)} & \text{if } \rho > 0 \text{ and } t \in R \\ \Phi(\theta) & \text{if } \rho = 0 \text{ or } t = \infty \\ \frac{\Phi(\theta)}{\Phi(\theta - \frac{t}{\rho})} & \text{if } \rho < 0 \text{ and } t \in R. \end{cases} \tag{6.9}$$

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

$$\Psi(\theta_{1,p}, T_n - c, \widehat{\rho}_{01,n}) = 1-p. \tag{6.10}$$

As $\Psi(\theta, t, \rho)$ is a strictly increasing function with range $(0,1)$ in θ for any $t > 0$ and any $\rho \in [-1,1]$, the solution $\theta_{1,p}$ of the above equation is unique whenever $T_n > c$ and easy to numerically obtain.

[19] 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.

Our conditional confidence interval is of the form:

$$CI_{\psi_1}(1-p) = \left[\psi_1(\widehat{\alpha}_{1,n}) - \theta_{1,p/2}\frac{\widehat{v}^*_{\psi,k_1}}{\sqrt{n}}, \psi_1(\widehat{\alpha}_{1,n}) - \theta_{1,1-p/2}\frac{\widehat{v}^*_{\psi,k_1}}{\sqrt{n}}\right]. \tag{6.11}$$

These critical values depend on T_n and hence are not approximations of the conditional quantiles of $\sqrt{n}(\psi_1(\widehat{\alpha}_{1,n}) - \psi_1(\alpha_1^*))/\widehat{v}^*_{\psi,k_1}$ given $T_n > c$. Therefore, the validity of our construction is 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.[20] 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 $\sqrt{n}E_{F_n}\ell(Z;\alpha_n^*)/\widehat{\sigma}_n \to -\infty$ 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

$$\mathcal{F}_n = \{F_0 \in \mathcal{F} : \sqrt{n}E_{F_0}\ell(Z;\alpha_n^*)/\sigma_n - c \geq -C\}, \tag{6.12}$$

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

Theorem 6.1 *Suppose that Assumptions* 4.1, 4.3 *and* 6.1-6.2 *hold. Then equation* (6.8) *holds with $\mathcal{F}_n$ defined in* (6.12).

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

$$\theta^* = \psi_1(\alpha_1^*)1(\ell_0 \geq 0) + \psi_2(\alpha_2^*)1(\ell_0 < 0). \tag{6.13}$$

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

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

vector $(\psi_1(\alpha_1^*), \psi_2(\alpha_2^*), \ell_0)$ has an asymptotically jointly normal estimator by Lemma 6.3. 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

$$\widehat{G}_n(x_1, x_2, x_0)'\widehat{\Sigma}_G^{-1}\widehat{G}_n(x_1, x_2, x_0) \leq \chi_3^2(1-p), \tag{6.14}$$

where $\chi_3^2(1-p)$ is the $1-p$ quantile of the chi-squared distribution with three degrees of freedom,

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

Then the projected confidence set of confidence level $1-p$ for θ^* is

$$CI_\theta(1-p) = \{\theta = x_1 1(x_0 \geq 0) + x_2 1(x_0 < 0) : \widehat{G}_n(x_1, x_2, x_0)'\widehat{\Sigma}_G^{-1}\widehat{G}_n(x_1, x_2, x_0) \leq \chi_3^2(1-p)\}. \tag{6.15}$$

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

Theorem 6.2 *Suppose that Assumptions* 4.1, 4.3, *and* 6.1-6.2 *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 uniform model selection test and the conditional confidence interval $CI_\psi(1-p)$.

Consider the following two models,

$$\begin{aligned} \mathcal{M}_1 &: E[Y|X_1] = \beta_{10} + \beta_{11}X_1 \\ \mathcal{M}_2 &: E[Y|X_2, X_3] = \beta_{20} + \beta_{21}X_2 + g(X_3), \end{aligned} \tag{7.1}$$

where $\beta_1 \equiv (\beta_{10}, \beta_{11})' \in R^2$, $\beta_2 \equiv (\beta_{20}, \beta_{21})' \in R^2$, and $g(\cdot) \in C^\infty([0,1])$. This example readily fits into the framework of (5.1) with $\alpha_1(x_1) = \beta_{10} + \beta_{11}x_1$, $\mathcal{A}_1 = \{b_0 + b_1x_1 : (b_0, b_1)' \in R^2\}$, $\alpha_2(x_2, x_3) = \beta_{20} + \beta_{21}x_2 + g(x_3)$, and $\mathcal{A}_2 = \{b_0 + b_1x_2 + g(x_3) : (b_0, b_1)' \in R^2, g \in C^\infty([0,1])\}$.

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, X_2. Let ε be standard normal and independent of X_1, X_2, X_3. Let

$$Y = 1 + aX_1 + bX_2 + c\sqrt{2}\sin(10\pi X_3) + \varepsilon. \tag{7.2}$$

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 $u_1 = bX_2 + c\sqrt{2}\sin(10\pi X_3) + \varepsilon$, and $u_2 = aX_1 + \varepsilon$. Thus,

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

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, $\omega^2_{F_0,*} > 0$.

7.1 Uniform Post Selection Test

To evaluate the performance of the uniform model selection 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 grids. 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 grids. This is the alternative collection in which, as η runs from 0 to 0.2, model 2 gets worse and worse relative to model $\mathcal{M}_1$.

We implement the uniform model selection 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.[21] The number of series terms is chosen by cross-validation with the search range being set between $[2\ln(\ln(n))]$ and 15.[22]

For comparison, we also investigate the performance of two bootstrap tests based on the naive extension statistic. Let $\widehat{\alpha}^*_{\mathbf{k},n}$, $\bar{\ell}^*_n(\cdot)$, and $\widehat{\omega}^*_n(\cdot)$ be the bootstrap analogue of $\widehat{\alpha}_{\mathbf{k},n}$, $\bar{\ell}_n(\cdot)$, and $\widehat{\omega}_n(\cdot)$. The first bootstrap test ($\varphi_n^{\text{boot-t}}$) uses T_n^V as the test statistic, and uses the conditional quantile of $n^{1/2}(\bar{\ell}^*_n(\widehat{\alpha}^*_{\mathbf{k},n}) - \bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n}))/\widehat{\omega}^*_n(\widehat{\alpha}^*_{\mathbf{k},n})$ as the critical value. The second bootstrap test ($\varphi_n^{\text{boot-LR}}$) uses $\bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n})$ as the test statistic and the conditional quantile of $\bar{\ell}^*_n(\widehat{\alpha}^*_{\mathbf{k},n}) - \bar{\ell}_n(\widehat{\alpha}_{\mathbf{k},n})$ as the critical value. Neither bootstrap test enjoys uniform asymptotic similarity or even validity when the models compared are overlapping nonnested. We compute their rejection rates for comparison because they have been suggested to us as alternatives.

[21] Fourier series yields similar results.

[22] 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 use it in this simulation example. The performance of our test with the cross-validated series terms is encouraging. The expanding lower bound of the search range ensures that the selected number grows with n, avoiding a violation of Assumption C.4(i).

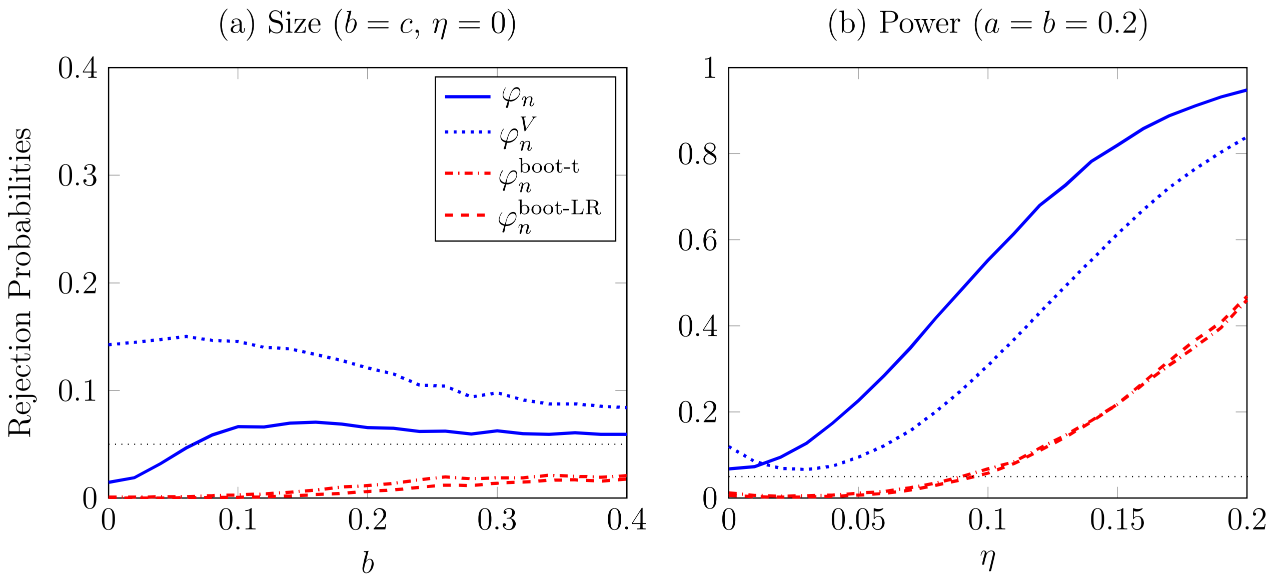


Figure 2: Rejection Rates of Two-sided Tests ($p = 0.05$, $n = 500$)

Figure 2 shows the rejection rates of the symmetric two-sided version of the tests at $n = 500$. The graph on the left shows the rejection rates under the first collection of DGPs—the collection of null DGPs. As the graph shows, 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 uniform model selection 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. The two bootstrap tests (dashed and dash-dotted lines) behave similarly to each other and both show severe under-rejection. The graph on the right shows the power properties. As we can see, the uniform model selection test has the best power across most of the range of η.[23]

The theory in previous sections suggests that the naive extension test biases toward the less parsimonious model (model $\mathcal{M}_2$ in this example). To see if this is true in finite samples, we plot the rejection rate under the null DGPs for the one-sided version of the tests in Figure 3. The left graph shows the tests for H_0 against $H_1 : f(\mathcal{M}_1, F_0) > f(\mathcal{M}_2, F_0)$, while the right graph shows the tests for H_0 against $H_1 : f(\mathcal{M}_1, F_0) < f(\mathcal{M}_2, F_0)$. Recall that model 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 over-rejection when the alternative is in favor of model 2 and severe under-rejection when the alternative is in favor of model 1. This behavior is in line with our theoretical derivation. Interestingly, the bootstrap tests have the opposite (albeit less severe) problem of the naive extension test.

[23] The power shown is not size-corrected. Size correction would bring the power of the naive extension test down more than it does the power of our test because there is more over-rejection in the former.

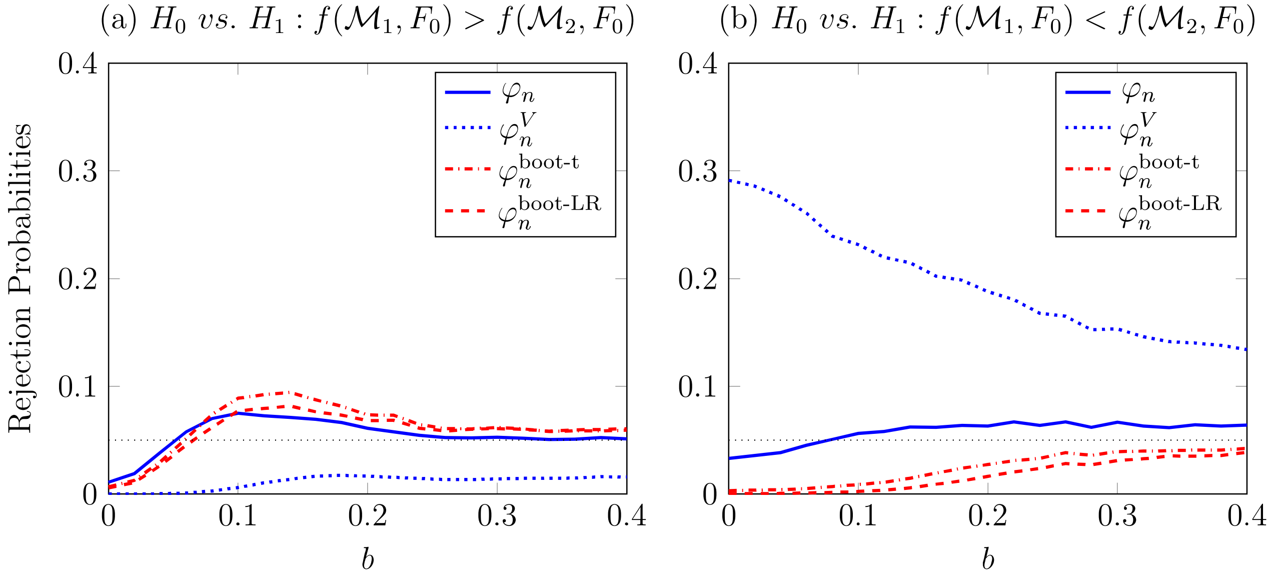


Figure 3: Null Rejection Rates of One-sided Tests ($p = 0.05$, $n = 500$)

7.2 Conditional Confidence Interval

In this subsection, we evaluate the performance of the conditional confidence interval $CI_{\psi_1}(1-p)$. Consider the parameters of interest β_{11} and β_{21}. Let model $\mathcal{M}_1$ be selected if $T_n > z_{0.05}$ 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 $j = 1, 2$,

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

Note that the conditional CI is only different from the naive CI in the critical values.

Figure 4 shows the results for β_{11}, and Figure 5 shows those for β_{21}. 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.

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 Figures 4 and 5 show the variability of the length of the conditional CI.[24] As we can see, the variability is small when the

[24] The variability of the length of the naive CI is negligible.

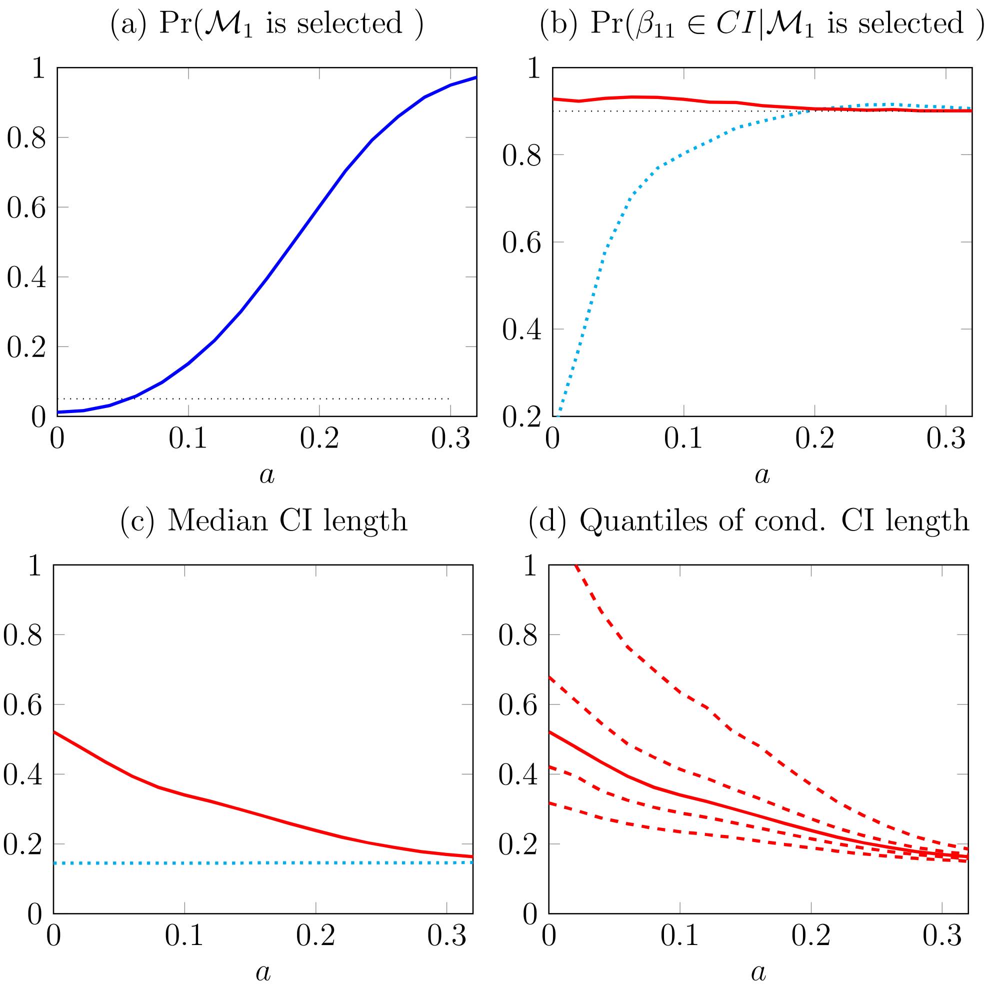


Figure 4: Performance of 90% Conditional Confidence Interval for β_{11}. (In graphs (b) and (c), the cyan dotted lines are for the naive CIs and the red solid lines are for our conditional CIs; in graph (d), the five lines are respectively the 25%, 40%, 50%, 60%, and 75% quantile of the length of the conditional CI.)

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

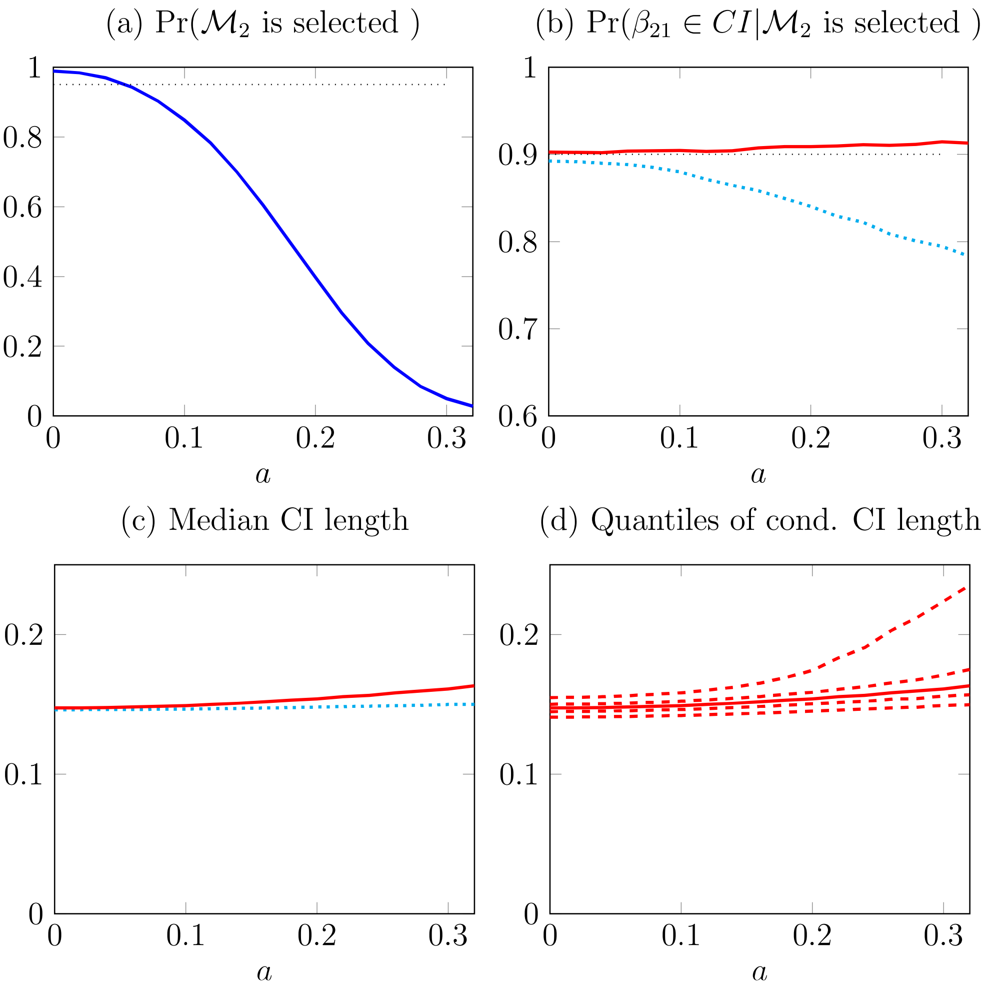


Figure 5: Performance of 90% Conditional Confidence Interval for β_{21}. (In graphs (b) and (c), the cyan dotted lines are for the naive CIs and the red solid lines are for our conditional CIs; in graph (d), the five lines are respectively the 25%, 40%, 50%, 60%, and 75% quantile of the length of the conditional CI.)

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. Our theory presented for the semi/non-parametric environment provides good approximation to this context due to the large number of parameters in each model.

8.1 Life-time Schooling Choice Example

We now 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 $s+1$:

$$\Pr(D_{s+1} = 1 | D_s = 1, X) = \frac{\exp(X'\beta_s)}{1 + \exp(X'\beta_s)}, \tag{8.1}$$

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

$$P(s|X) = \frac{1}{1 + \exp(X'\beta_s)} \times \frac{\exp(X'\beta_{s-1})}{1 + \exp(X'\beta_{s-1})} \times \cdots \times \frac{\exp(X'\beta_1)}{1 + \exp(X'\beta_1)}. \tag{8.2}$$

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

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

where $F(t) = \frac{\exp(t)}{1+\exp(t)}$, $\alpha_{\bar{s}} = -\infty$ 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.

We first 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.[25]

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 $\mathcal{M}_1$.[26]

Table 1: Model Selection Tests Based on NLSY 97

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

We then compare the models in terms of the Kullback-Leibler distance (that is, $f(\mathcal{M}_j, F_0) = \max_{\theta_j} E_{F_0} \log P(S|X, \theta_j))$. We implement the two-sided version of both our new robust test and the Vuong (1989) test.[27] Table 1 shows the value of the test statistics as well as p-values of both tests. The first-order asymptotic 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

[25] Results using reconstructed sample from the NLSY 79 are reported below.

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

[27] The Vuong (1989) test is the the strict nonnested test proposed in Vuong (1989).

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 $\mathcal{M}_1$ 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. The biased Vuong (1989) test tells a different story and can be misleading.

To illustrate our conditional confidence interval, we computed these intervals for the parameters in Model $\mathcal{M}_2$ conditional on the event that $T_n < z_{0.025} \approx 1.96$. 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 $\theta_{2,p}$ to be the same as z_p up to at least the sixth digit.

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. 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 shall 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 and will be reported in later papers.

References

Aït-Sahalia, Y., Bickel, P. J., and Stoker, T. M. (2001). Goodness-of-fit tests for kernel regression with an application to option implied volatility. *Journal of Econometrics*, 105:363–412.

Atkinson, A. (1970). A method for discriminating between models. *Journal of Royal Statistical Society*, B 32:323–353.

Barseghyan, L., Molinari, F., ODonoghue, T., and Teitelbaum, J. C. (2013). The nature of risk preferences: Evidence from insurance choices. *American Economic Review*, 103:2499–2529.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, 62(3):6570681.

Belloni, A., Chernozhukov, V., and Fernández-val, I. (2011). Conditional quantile process based on series or many regressors. unpublished manuscript, Department of Economics, Duke University.

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

Bisin, A., Topa, G., and Verdier, T. (2004). Religious intermarriage and socialization in the united states. *Journal of Political Economy*, 112:615–664.

Bonnal, L., Fougère, D., and Sérandon, A. (1997). Evaluating the impact of french employment policies on individual labour market histories. *The Review of Economic Studies, Special Issue (Oct., 1997)*, 64:683–713.

Bontemps, C., Florens, J.-P., and Richard, J.-F. (2008). Parametric and non-parametric encompassing procedures. *Oxford Bulletin of Economics and Statistics*, 70:751–780.

Caballero, R. J. and Engel, E. M. R. A. (1999). Explaining investment dynamics in u.s. manufacturing: A generalized (s,s) approach. *Econometrica*, 67:783–826.

Cameron, S. V. and Heckman, J. J. (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. and Fan, Y. (2005). Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection. *The Canadian Journal of Statistics*, 33:389–414.

Chen, X., Hong, H., and Shum, M. (2007). Nonparametric likelihood ratio model selection tests between parametric likelihood and moment condition models. *Journal of Econometrics*, 141:109–140.

Chen, X. and Liao, Z. (2015). Semiparametric two-step gmm estimation with weakly dependent data. *Journal of Econometrics*, 189:70–86.

Chen, X., Liao, Z., and Sun, Y. (2014). Sieve m inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics*, 178:639–658.

Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 66:289–314.

Coate, S. and Conlin, M. (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.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24:406–424.

Fafchamps, M. (1993). Sequential labor decisions under uncertainty: An estimable household model of west-african farmers. *Econometrica*, 61:1173–1197.

Fan, Y. and Li, Q. (1996). Consisstent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica*, 64:865–890.

Gandhi, A. K. and Serrano-Padial, R. (2015). Does belief heterogeneity explain asset prices: the case of the longshot bias. *Review of Economic Studies*, 82:156–186.

Gourieroux, C. and Monfort, A. (1995). Testing, encompassing, and simulating dynamic econometric models. *Econometric Theory*, 11:195–228.

Gowrisankaran, G. and Rysman, M. (2012). Dynamics of consumer demand for new durable goods. *Journal of Political Economy*, 120:1173–1219.

Heath, C., Huddart, S., and Lang, M. (1999). Psychological factors and stock option exercise. *Quarterly Journal of Economics*, 114:601–627.

Hong, Y. and White, H. (1995). Consistent specification testing via nonparametric series regression. *Econometrica*, 63:1133–1159.

Karaivanov, A. and Townsend, R. M. (2014). Dynamic financial constraints: Distinguishing mechanism design rom exogenously incomplete regimes. *Econometrica*, 82:887–959.

Kendall, C., Nannicini, T., and Trebbi, F. (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 Vuong, Q. H. (1996). Nonparametric selection of regressors: The nonnested case. *Econometrica*, pages 207–219.

Lavergne, P. and Vuong, Q. H. (2000). Nonparametric significance testing. *Econometric Theory*, 16:576–601.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21:21–59.

Leeb, H. and Pötscher, B. M. (2006). Can one estiamte the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34:2554–2591.

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 Richard, J. F. (1986). The encompassing principle and its applications to testing nonnested hypothesis. *Econometrica*, 3:657–678.

Moines, S. and Pouget, S. (2013). The bubble game: An experimental study of speculation. *Econometrica*, 81:1507–1539.

Moon, C.-G. and Stotsky, J. G. (1993). The effect of rent control on housing quality change: A longitudinal analysis. *Journal of Political Economy*, 101:1114–1148.

Nyamko, Y. and Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica*, 70:910–1005.

Palfrey, T. R. and Prisbrey, J. E. (1997). Anomalous behavior in public goods experiments: How much and why? *American Economic Review*, 87:829–846.

Paulson, A. L., Townsend, R. M., and Karaivanov, A. (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 Deaton, A. S. (1978). Testing nonnested nonlinear regression models. *Econometrica*, 76:677–694.

Pesaran, M. H. and Ulloa, M. R. D. (2008). Non-nested hypotheses. Palgrave Macmillan, 2nd edition.

Ramalho, J. J. S. and Smith, R. J. (2002). Generalized empirical likelihood non-nested tests. *Journal of Econometrics*, 107:99–125.

Rivers, D. and Vuong, Q. (2002). Model selection tests for nonlinear dynamic models. *Econometrics Journal*, 5:1–39.

Schennach, S. M. and Wilhelm, D. (2011). A simple parametric model selection test. unpublished manuscript, University of Chicago.

Shen, X. (1997). On methods of sieve and penalization. *The Annals of Statistics*, 25:2555–2591.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, 22.

Shi, X. (2015). A nondegenerate vuong test. *Quantitative Economics*, 6:85–121.

Tian, X. and Taylor, J. (2015). Asymptotics of selective inference. unpublished manuscript, Department of Statistics, Stanford University.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. unpublished manuscript, Carnegie Mellon University.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (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.

White, H. and Wooldridge, J. M. (1991). Some results on sieve estimation with dependent observations. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics.* Cambridge University Press.