Journal of Econometrics 194 (2016) 187–202

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Consistent model specification tests based on *k*-nearest-neighbor estimation method

Hongjun Li [a], Qi Li [b,a,*], Ruixuan Liu [c]

[a] ISEM, Capital University of Economics & Business, Beijing, 100070, China

[b] Department of Economics, Texas A&M University, College Station, TX 77843-4228, United States

[c] Department of Economics, Emory University, Atlanta, GA 30322, United States

ARTICLE INFO

Article history:
Received 1 July 2013
Received in revised form
19 August 2015
Accepted 24 March 2016
Available online 15 June 2016

JEL classification:
C14

Keywords:
k-nearest-neighbor method
Consistent test
Bootstrap
Empirical application

ABSTRACT

We propose a simple consistent test for a parametric regression functional form based on *k*-nearest-neighbor (*k*-nn) method. We derive the null distribution of the test statistic and show that the test achieves the minimax rate optimality against smooth alternatives. A wild bootstrap method is used to better approximate the null distribution of the test statistic. We also propose a *k*-nn statistic which tests for omitted variables nonparametrically. Simulations and an empirical application using US economics new Ph.D. job market matching data show that the *k*-nn method is more appropriate than the kernel method to analyze unevenly distributed data.

1. Introduction

Specification testing is an important research area for econometrics and statistics as inferences based on misspecified econometric models can lead to misleading conclusions. For example, active research areas where the correct specification of parts of a model plays a crucial role are Missing Data and Program Evaluation, respectively. Under a selection-on-observable assumption, average treatment effects can be consistently estimated based on the estimated propensity score function. However, propensity score is usually parametrically specified as a Linear Probability, Probit or Logit model, see e.g. Bravo et al. (2011) and references therein. Policy evaluation based on an incorrectly specified propensity score function can have serious consequences. Therefore, it is important to test for the correct model specifications in empirical analysis. There is a rich literature on constructing consistent model specification tests, see Bierens (1982, 1990), Azzalini and Bowman (1993), Hong and White (1995), Lewbel (1995), Fan and Li (1996), Zheng (1996), Lavergne and Vuong (1996), Andrews (1997), Bierens and Ploberger (1997), Delgado and Manteiga (2001), Guerre and Lavergne (2002), among others. Various nonparametric techniques are used in developing those tests such as kernel, series, splines and wavelet methods. While *k*-nearest-neighbor (*k*-nn) method is one of the popular nonparametric estimation techniques, its adaption to the model specification testing is rather limited in scope until the recent work of Jun and Pinkse (2009b, 2012).

The *k*-nn method is one of the earliest nonparametric recipes (Fix and Hodges, 1951), and it continues to serve as a building block in non/semi-parametric econometrics. Different nonparametric estimation methods have their own advantages in different situations. Undoubtedly, the *k*-nn method is preferable over the kernel smoothing when data points are highly unevenly distributed over the data support, as is often the case for nonexperimental data in social sciences, because the kernel method with a fixed smoothing parameter may contain too few data in the sparse ranges of the support and thus may lead to unreliable estimation and testing results. In contrast, the *k*-nn method always uses (the nearest) k data points in estimation and does not suffer the problem that too few data is used in the nonparametric estimation. Indeed, this advantage has been recognized and explored in subsequent non/semi-parametric econometric applications. Notable contributions on using *k*-nn method include Robinson (1987, 1995), Delgado (1992), Delgado and Stengos (1994) for the

* Corresponding author at: Department of Economics, Texas A&M University, College Station, TX 77843-4228, United States.
E-mail address: qi-li@tamu.edu (Q. Li).

http://dx.doi.org/10.1016/j.jeconom.2016.03.004

estimation of conditional heteroskedasticity, and Heckman et al. (1998), Abadie and Imbens (2006, 2011), Jacho-Chávez (2008) for matching type estimators in treatment effects literature. We refer readers to Newey (1990), Liu and Lu (1997), Jun and Pinkse (2009a), Chu et al. (2013) for applying nonparametric k-nn method to study a variety of interesting semiparametric models.

In this paper, we consider the problem of constructing consistent specification tests in regression models using the nonparametric k-nn method. Following the existing model specification testing literature using nonparametric kernel and series methods as in Härdle and Mammen (1993), Hong and White (1995), Fan and Li (1996), and Zheng (1996), we construct test statistics utilizing the fact that the conditional (on the covariates) mean of the error term from the null model equals zero and we estimate the conditional mean functions by the k-nn method. We prove that our standardized test statistic is asymptotically pivotal and it converges to the standard normal distribution weakly. Furthermore, this approach is shown to be extendable to testing null hypotheses where the null models are themselves nonparametric or semiparametric regression models (e.g., Fan and Li, 1996). Our analysis complements the work of Jun and Pinkse (2009b, 2012), and extends the existing k-nn based tests in several directions. First, we show the bootstrapped version of the k-nn based test is asymptotically valid. Second, we investigate the local power property in the minmax context (Horowitz and Spokoiny, 2001; Guerre and Lavergne, 2002). Namely, we prove that our proposed k-nn based test is rate optimal against a class of smooth alternatives. Third, we propose a k-nn based test checking for omitted variables, in the spirit of Fan and Li (1996), where the null hypothesis is not parametrically determined. Fourth, we conduct simulations to show that the k-nn based test has better power performance than the kernel based test when data are unevenly distributed in the data support. Finally, we provide an empirical example showing the advantage of using the k-nn method over the kernel method when data is highly unevenly distributed over the data support.

Technically, the k-nn method is more difficult to handle than other nonparametric approaches such as kernel and series methods, as it can be considered as a kernel approach but with a random bandwidth in the local neighborhood. Thus, it creates non-standard dependence among random variables even when the data are originally independent and identically distributed (i.i.d.). Here we explore the martingale structure of our test statistics, and utilize various useful results in Jun and Pinkse (2009b, 2012). The Monte-Carlo simulation results demonstrate that the proposed test based on k-nn is more powerful than the kernel based test when the data are highly unevenly distributed. In our empirical analysis of a job matching model, we found that kernel based estimation yields a very rough estimated curve, while the k-nn based approaches (both for estimation and testing) are consistent with the theoretical model in Gan and Li (2016). We do not pretend that we can argue a k-nn based test has theoretical advantages over other type of nonparametric tests such as kernel based tests. We view the k-nn based tests as complementary to other specification tests.

The paper is organized as follows. Section 2 reviews the existing literature focusing on the nonparametric smoothing based tests. Section 3 proposes a consistent test for a parametric regression functional form based on k-nn method. We derive the asymptotic null distribution, prove its rate optimality uniformly against smooth alternatives, and show the wild bootstrap method can be used to approximate the null distribution of the test statistic. Section 4 extends the k-nn based specification test to checking for omitted variables. Section 5 reports Monte Carlo simulation results and illustrates the advantage of our proposed k-nn based test over kernel based test when data are unevenly distributed. Section 6 provides an empirical application of using the proposed test method to the US new Economics Ph.D. ''Job Matching Market'' data. Section 7 concludes the paper and discusses some possible extensions. All the proofs and technical lemmas are regulated to three appendices.

2. Literature review

There are mainly two types of nonparametric approaches commonly used for specification testing: the approach based on nonparametric smoothing methods such as kernel, series and k-nn methods. The second approach is the so-called non-smoothing tests based on certain integrated moments or empirical processes, see Bierens (1982, 1990), Bierens and Ploberger (1997), Stute (1997), Neumeyer and Dette (2003), Neumeyer and Van Keilegom (2010), among others. Fan and Li (2000) point out the close relationship between these two types of tests, showing that many of the existing non-smoothing tests can be obtained from kernel based smoothing tests with the bandwidth parameters fixed (instead of shrinking to zero as sample size goes to infinity). Therefore, our discussion in this section will selectively focus on the smoothing tests. One can refer to Hart (1997) for a collective survey on the use of nonparametric smoothing methods for specification testing of a parametric model and González-Manteiga and Crujeiras (2013) for a comprehensive review of the recent developments.

Bierens (1982) is the first to propose a consistent model specification test without smoothing methods, whereas Ullah (1985) is the first to propose using nonparametric method to construct smoothing specification tests. Other early work on consistent specification tests include Eubank and Spiegelman (1990), Eubank and Hart (1992), Wooldridge (1992), Yatchew (1992) and Härdle and Mammen (1993), among others. The idea of constructing smoothing consistent model specification tests in the regression model framework is quite simple and intuitive, basically, one wants to compare the differences between a restricted null model curve (say, a parametric regression model) and an unrestricted nonparametric smoothing curve estimates. There are many different ways to construct the test statistics based on consistent estimators of different characteristics of the null models. The kernel based method seems to be the most popular choice in constructing test statistics for model check, a partial list includes Robinson (1989, 1991), Härdle and Mammen (1993), Fan and Li (1996), Li and Wang (1998), Zheng (1996), Dette (1999), Guerre and Lavergne (2002), among others. Fan and Li (1996), Lavergne and Vuong (1996) extend the analysis of testing a parametric null model to the case in which the null models themselves can be semiparametric or nonparametric models. Fan and Li (1996) construct their test statistics by using density functions as weights to avoid the random denominator problems in kernel based tests because these random denominators are not bounded away from zero under general conditions. Breunig (2015), Eubank and Spiegelman (1990), Eubank and Hart (1992), Hong and White (1995), Li et al. (2003), Stengos and Sun (2000), Sun and Li (2006) propose different test statistics using spline, series or wavelet regressions.

Unlike the literature of specification test based on kernel or series method, there are only a small number of papers that use the k-nn based method to construct test statistics. Delgado and Stengos (1994) first employ k-nn based method testing for a non-nested regression specification. Under the alternative hypothesis, they require there exist additional regressors besides those in the null parametric form. Stute and González-Manteiga (1996) consider a k-nn based test for a univariate regression model, they require the regressor and the error term are independent with each other, thus ruling out the possibility of conditional heteroskedastic errors. With the monotone transformation, Stute and González-Manteiga's (1996) k-nn test essentially becomes a kernel-based