cemmap
centre for microdata methods and practice

Nonparametric testing for exogeneity with discrete regressors and instruments

Katarzyna Bech
Grant Hillier

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP11/15

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

An ESRC Research Centre

Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE **Tel:** 44 (0)20 7291 4800 **Email:** info@cemmap.ac.uk **www.cemmap.ac.uk**

Nonparametric testing for exogeneity with discrete regressors and instruments

Katarzyna Bech and Grant Hillier
University of Southampton

March, 2015

Abstract

This paper presents new approaches to testing for exogeneity in non-parametric models with discrete regressors and instruments. Our interest is in learning about an unknown structural (conditional mean) function. An interesting feature of these models is that under endogeneity the identifying power of a discrete instrument depends on the number of support points of the instruments relative to that of the regressors, a result driven by the discreteness of the variables. Observing that the simple nonparametric additive error model can be interpreted as a linear regression, we present two test-statistics. For the point identifying model, the test is an adapted version of the standard Wu-Hausman approach. This extends the work of Blundell and Horowitz (2007) to the case of discrete regressors and instruments. For the set identifying model, the Wu-Hausman approach is not available. In this case the test-statistic is derived from a constrained minimization problem. The asymptotic distributions of the test-statistics are derived under the null and fixed and local alternatives. The tests are shown to be consistent, and a simulation study reveals that the proposed tests have satisfactory finite-sample properties.

1 Introduction

The possible presence of endogeneity is one of the common problems in econometric models. It occurs when the regressor is correlated with the model error term. Typically it is a result of omitting a relevant explanatory variable, of simultaneity in the model, or measurement error in the regressor. The presence of endogenous regressors in the nonparametric model produces bias in the identified case, and non-existence of any consistent estimator in the set identified case. Because of the potentially severe consequences of endogeneity, applied researchers need to check whether the explanatory variables used are exogenous, before providing an inference on the parameters of interest. Following the work of Hausman (1978), a vast literature on testing for exogeneity of the regressors has emerged.

Recently, with the expansion of interest in nonparametric models, new testing procedures have been developed. The problem of testing the correct specification of a nonparametric model of the form

$$Y = h(X) + \varepsilon \tag{1}$$

has been discussed by many authors including Fan and Li (1996), Zheng (1996), Lavergne and Vuong (2000), Lavergne and Patilea (2008) and Blundell and Horowitz (2007). These tests fit in a conditional moment restriction testing framework, and are based on the earlier work of Newey (1985) and Bierens (1990), among others.

All the nonparametric tests of this type assume that the regressors are continuously distributed. The aim of this paper is to provide a test for exogeneity in a nonparametric model with discrete explanatory variables. A model with discrete regressors arises in many economic problems. Variables such as gender, marital status or education levels typically take discrete values. When X is binary it may indicate the occurrence of the event. In empirical applications, such regressors are called 'dummy variables' taking values 0 or 1, for example, an individual is either male or female, working or unemployed. The discrete regressor with multiple categories might measure e.g. the number of children in a household, or give the position on an attitudinal scale. The nonparametric model with discrete regressors has been applied by Hu and Lewbel (2008) to identify and estimate the difference in average wages between individuals who falsely claim college experience and those who tell the truth about not completing college education. More recently, Iori, Kapar and Olmo (2014) use nonparametric methods to explain variation in the continuous variable (bank funding spreads) given a set of discrete regressors (bank characteristics, nationality, size and operating currency) in the European interbank money market.

The most popular method of dealing with endogeneity in econometric models is by instrumental variable (IV) estimation. Although IV methods are traditionally parametric in nature, the extension of the approach to a more flexible, non-parametric framework was introduced by Newey and Powell (2003). The method suggests that researchers should find a set of variables satisfying instrument relevance and exogeneity conditions and use them to consistently estimate the causal relationship between the dependent variable and endogenous regressors. However, the IV method involves some identification issues. The problem with identification is particularly noticeable in nonparametric models with additive errors when the regressors are discrete. Florens and Malavolti (2003) and Das (2005) show that the identification of the unknown function of interest depends on the support of instruments relative to the support of the endogenous regressor. If the identification condition is violated and point identification is not feasible, the model still has some partial identifying power. Partial identification can be achieved in models which cannot provide the exact value of the parameter or structure of interest, but contain enough information to bound these values to informative sets. Chesher (2004) discusses the estimation of the regression function $h(\cdot)$ in equation (1) with this framework.

One of the advantages of nonparametric models with discrete endogenous regressors is that they do not suffer from the ill-posed inverse problem that arises in nonparametric models with continuous endogenous regressors. The problem derives from the discontinuity of the mapping from the structural to the reduced form, when estimating an infinite dimensional function $h(\cdot)$ in continuous specifications (Newey and Powell (2003)). This means that $h(\cdot)$ cannot be estimated consistently by replacing the unknown population quantities with consistent estimators. In order to obtain a consistent estimator, it is necessary to regularize the mapping that identifies the unknown function of interest. Restricting the endogenous regressors to be discrete eliminates the ill-posed inverse problem. The discrete specification is well-posed, and no regularization of the problem is required.

The plan of this paper is as follows. Section 2 introduces the nonparametric model of interest and presents the notation that enables us to interpret equation (1) as a linear model. This section also explains the identification problems in the presence of endogenous regressors and shows some basic estimation results. Section 3 presents the test for models that point identify the unknown function of interest and establishes the asymptotic properties under the null and alternative hypothesis. Section 4 introduces the test for models that are set identified. The asymptotic distribution of the test statistic under the null and alternative hypothesis is also derived in this section. In Section 5, we present the results of the Monte Carlo investigation of the finite-sample properties of the proposed tests. Section 6 concludes. All proofs are in the appendix.

2 Model and assumptions

2.1 Notation

The upper case letters X, Y, Z will denote observed random variables, and x_i^s, y_i^s, z_i^s will denote sample (data) points. Symbols x_k for $k = 1, ..., K$ denote the points of support of a discrete random variable X. $I(A)$ stands for an indicator function, which takes value 1 if the event A occurs,and is 0 otherwise. The probability density function of a continuous random variable W is denoted by $f_W(w)$, and the probability mass function of a discrete random variable X is $p_X(x)$. The cumulative distribution function is denoted by $F_X(x)$. For a matrix A of full column rank we define $P_A = A\left(A'A\right)^{-1}A'$ and $M_A = I - P_A$, both of which depend only on the space spanned by the columns of A. For any r, l_r denotes an r-vector of ones and C_r denotes an $r \times (r-1)$ matrix with the properties $C_r' l_r = 0$ and $C_r' C_r = I_{r-1}$.

2.2 Model

We consider the simple additive error model in which a continuous outcome Y is determined by equation (1), with X a single discrete regressor, and ε denotes a continuously distributed error term. The interest of econometricians typically lies in estimating the unknown structural function $h(\cdot)$. Consistent nonparametric estimation of $h(\cdot)$ is feasible under the assumption that the regressors are exogenous. Numerous definitions of exogeneity have been provided in the literature, see Deaton (2010). The standard exogeneity condition is that of an absence of correlation between the regressor and the model error term. Here we employ the definition proposed by Blundell and Horowitz (2007) for nonparametric regressions: the explanatory variable X is exogenous if the conditional moment restriction $E[\varepsilon|X = x_k] = 0$ holds for all $k = 1, ..., K$. In that case $E[Y|X] = h(X)$, i.e. the conditional mean of the dependent variable given X coincides with $h(X)$. This definition has the advantage that the standard nonparametric regression of Y on X is then appropriate for the consistent estimation of the unknown function of interest $h(\cdot)$.

In the presence of endogeneity of regressors, further analysis needs to be conducted. The common strategy to deal with the endogeneity problem is to use instrumental variables. However, the choice of a consistent estimation method depends on a characteristic of the available instruments. The identifying power of the model varies with the number of the points of support of the instrumental variable (see Section 2.5). The complete model is characterized by the following set of assumptions:

Assumption 1. X is a discrete (scalar) random variable with support $\{x_1, ..., x_K\}$ and associated probabilities $p_k > 0$. ■

Assumption 2. There exists a discrete instrumental variable Z with support $\{z_1, ..., z_J\}$ and associated probabilities $q_j > 0$, with the property

$$E[\varepsilon|Z = z_j] = 0, \quad j = 1, ..., J \tag{2}$$

which defines the instrument exogeneity condition.[1] The matrix of joint probabilities P with elements

$$p_{jk} = \Pr[X = x_k \cap Z = z_j]; \; j = 1, ..., J; \; k = 1, ..., K$$

is of full rank K when $J \geq K$ and of full rank J when $J < K$. ■

Assumption 3. $E[X|Z = z_j]$ and $E[h(X)|Z = z_j]$ vary with z_j. The first condition (the instrument relevance condition) together with (2) ensures that Z is a valid instrument. ■

Assumptions 2 and 3 are analogous to the standard assumptions for the validity of instruments in single equation IV estimation (see, for example, Greene (1993), Section 20.4.3)

Assumption 4. The data consists of n *iid* observations on (Y, X, Z), denoted by (y_i^s, x_i^s, z_i^s) for $i = 1, ..., n$. Under exogeneity, for all j and k,

$$E[\varepsilon|X = x_k, Z = z_j] = 0 \text{ and } Var[\varepsilon|X = x_k, Z = z_j] = \sigma^2.$$

■

[1] Notice that we include in the support of X and Z only points for which p_k and q_j are strictly positive

The complete model consists of equations (1) and (2). We are interested in testing the null hypothesis of exogeneity of the regressor i.e. $E[\varepsilon|X = x_k] = 0$ for all k. Equivalently, in terms of observables,

$$H_0 : E[Y|X = x_k] = h(x_k), \quad k = 1, ..., K.$$

If this condition is satisfied the unknown function $h(\cdot)$ can be consistently estimated nonparametrically.

In equation (1) the function $h(\cdot)$ is unknown and if $h(x_k)$ is completely arbitrary, the null hypothesis would not constrain the conditional density function of Y given X, $f_{Y|X}(y|x)$, and would therefore be untestable. Thus, more information than just equation (1) is required for H_0 to become a testable hypothesis. This additional information is acquired by using the fact that there exists a valid instrument Z satisfying (2) for any admissible z_j.

Let $n_k^X = \sum_{i=1}^n I(x_i^s = x_k)$ and $n_j^Z = \sum_{i=1}^n I(z_i^s = z_j)$ denote the multiplicities of x_k and z_j in the sample, and also $n_{jk} = \sum_{i=1}^n I(x_i^s = x_k)I(z_i^s = z_j)$. Under Assumption 2, the unknown function $h(\cdot)$ satisfies the set of J linear equations

$$E[Y|Z = z_j] = \sum_{k=1}^{K} \Pr[X = x_k|Z = z_j]h(x_k), \quad j = 1, ..., J. \tag{3}$$

Let β denote the K-vector with $\beta_k = h(x_k)$, $k = 1, ..., K$, π be the J-vector with the elements $E[Y|Z = z_j]$, $j = 1, ..., J$, and Π be the $J \times K$ matrix of conditional probabilities $\Pr[X = x_k|Z = z_j]$, $j = 1, ..., J$, $k = 1, ..., K$. Then, (3) can be written as the system[2]:

$$\pi = \Pi\beta. \tag{4}$$

The nonparametric nature of the model is reflected in the fact that β, the vector of values of $h(\cdot)$ at the support points of X *is completely unknown*.

It is worth noting that equation (4) always has a solution (for β), since for each $j = 1, ..., J$, by definition

$$E[Y|Z = z_j] = \sum_{k=1}^{K} \Pr[X = x_k|Z = z_j]E[Y|X = x_k, Z = z_j]$$

so that π is certainly in the space spanned by the columns of Π. That Π has full rank $\min\{J, K\}$ is part of Assumption 2.

The hypothesis H_0 imposes the constraint that the vector of conditional means $E[Y|X = x_k]$ is a solution to a linear equations $\pi = \Pi\beta$, so in this case the null hypothesis imposes a restriction on the conditional density function $f_{Y|X}(y|x)$ and is therefore testable.

[2] In the continuous case, equation (4) corresponds to the integral equation for the structural function (2.2) in Blundell and Horowitz (2007).

Remark 1 *There might be other restrictions that can be imposed on $h(\cdot)$ to make the null hypothesis testable. In order to make sure that $h(\cdot)$ is not entirely arbitrary, one could impose some shape restrictions dictated by economic theory. Such restrictions are already in use in the literature of nonparametric estimation, for example by Hall and Huang (2001) who estimate the conditional mean function subject to a monotonicity constraint. Monotone estimates are required in many empirical applications, when the theory suggests that the outcome should be monotonic in explanatory variables e.g. wage increasing in the years of schooling. Blundell, Horowitz and Parey (2012) use different shape restriction and provide a nonparametric estimator of the demand function assuming that the unknown function $h(\cdot)$ satisfies the Slutsky condition of consumer theory. The literature suggests that imposing shape restrictions improves the precision of nonparametric estimates, but in our case it might also act as a tool to ensure that the hypothesis of exogeneity of regressors is testable.*

The elements of the vector π can be consistently estimated from the data, by averaging those y_i that correspond to the observations with $z_i^s = z_j$, i.e. by

$$\widehat{\pi}_j = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i I(z_i^s = z_j)}{\frac{1}{n}\sum_{i=1}^{n} I(z_i^s = z_j)} = \frac{1}{n_j^Z}\sum_{i=1}^{n} y_i I(z_i^s = z_j). \tag{5}$$

The elements of the matrix of conditional probabilities Π can be written as

$$\Pr[X = x_k | Z = z_j] = \frac{\Pr[X = x_k \cap Z = z_j]}{\Pr[Z = z_j]}$$

and can be consistently estimated by

$$\widehat{\Pi}_{jk} = \frac{\frac{1}{n}\sum_{i=1}^{n} I(x_i^s = x_k) I(z_i^s = z_j)}{\frac{1}{n}\sum_{i=1}^{n} I(z_i^s = z_j)} = \frac{n_{jk}}{n_j^Z}. \tag{6}$$

Thus, π and Π can (ultimately) be learned from the data, and the problem is to use this information to make inference on $h(\cdot)$.

Remark 2 *In the discussion here, and also in what follows, it is implicitly assumed that all K support points of X, and all J of Z, occur in the sample. That is, that both n_k^X and n_j^Z are non-zero for all $k = 1, .., K$ and $j = 1, .., J$. This will ultimately (for large enough n) be the case with probability one. The alternative would be to define estimates for the π_j and Π_{jk} only for those points x_k and z_j that occur in the sample, say $K_s \leq K$ and $J_s \leq J$ points, and allow these to increase to K and J respectively, as n increases. This would make the arguments and derivations to follow considerably more cumbersome, without materially affecting the results, so instead we will tacitly assume throughout that n is large enough to ensure that $K_s = K$ and $J_s = J$.*

There is no difficulty in extending the results by allowing for additional discrete exogenous regressors in the model as long as there is only one possibly endogenous explanatory variable. An unresolved issue is how to deal with multiple discrete endogenous regressors. Assuming that more than one regressor is endogenous is likely to affect the identification conditions and existing estimation and testing procedures. The model with multiple discrete endogenous regressors will be addressed in future research.

2.3 Linear Model Representation

The above setup can be represented compactly in terms of a linear model. To do so, define the $n \times K$ matrix L_X with (i, k) element

$$(L_X)_{ik} = I(x_i^s = x_k),$$

so that $(L_X)_{ik} = 1$ if observation i corresponds to a value x_k for X, and is 0 otherwise. Likewise, define the $n \times J$ matrix L_Z with elements

$$(L_Z)_{ij} = I(z_i^s = z_j).$$

Note that the row sums of both L_X and L_Z are 1, because each row of both contains exactly one element that is equal to 1. Both L_X and L_Z are random matrices, because the positions of the non-zero elements, and the multiplicities of each x_k and z_j, are determined randomly in the sample. Let x denote the K-vector with elements x_k, $k = 1, ..., K$, the support points of the regressor, and let $x^s = L_X x$ denote the n-vector of sample observations x_i^s, $i = 1, ..., n$. Finally, let y denote the $n-$vector of sample observations on Y, a realization of the random $n-$vector $\mathcal{Y}$.

Using the notation just introduced, (5) can be written as

$$\widehat{\pi} = \left(L_Z' L_Z\right)^{-1} L_Z' y \tag{7}$$

and (6) becomes

$$\widehat{\Pi} = \left(L_Z' L_Z\right)^{-1} L_Z' L_X. \tag{8}$$

The inverse in (7) and (8) exists almost surely for large enough sample size[3], since $\Pr[Z = z_j] = q_j > 0$. Note that

$$n^{-1} L_Z' L_Z \to^p \operatorname{diag}(q_j) := D_Z$$

because $\frac{1}{n}\sum_{i=1}^n I(z_i^s = z_j) \to^p E[I(z_i^s = z_j)] = \Pr[Z = z_j]$. Hence, by the Slutsky Theorem

$$\left(n^{-1} L_Z' L_Z\right)^{-1} \to^p D_Z^{-1}.$$

Similarly, the elements of $n^{-1} L_Z' L_X$ are consistent estimates of the joint probability matrix P. Therefore, $\widehat{\pi} \to^p \pi$ and $\widehat{\Pi} \to^p \Pi := D_Z^{-1} P$.

[3] Of course, for existence we require $n > K$ and $n > J$ here. And, as discussed in Remark 2, we are tacitly assuming that n is large enough to ensure that $K_s = K$ and $J_s = J$.

Letting $\mathcal{X}$ denote the random $n-$vector of observations on X, the model can be written in the familiar form $E[\mathcal{Y}|\mathcal{X} = L_X x] = L_X\beta + E[\varepsilon|\mathcal{X} = L_X x]$, a linear model for the vector $\mathcal{Y}$ with random regressor matrix L_X and unknown parameters $\beta_k = h(x_k)$, $k = 1, ..., K$. The null hypothesis then takes the form:

$$H_0 : E[\mathcal{Y}|\mathcal{X} = L_X x] = L_X\beta,$$

Thus, although the model is purely nonparametric, it can be interpreted as a linear model. Note that even though in the nonparametric specification there is only one discrete regressor X, L_X is $n \times K$ in the linear model specification. Also,observe that the support points x_k of X determine the points at which we can learn $h(\cdot)$, i.e., β, but do not appear elsewhere in the linear model. This familiar linear model specification allows us to connect the nonparametric estimators with well known regression estimators, particularly OLS and 2SLS.

2.4 A complication

There is a relationship between L_X and L_Z, which has an important implication for the further analysis. This is that every sample point is associated with exactly one support point of both X and Z. It follows that, for any regressor X and any instrument Z, the row sums of both L_X and L_Z are all equal to one. That is,

$$L_X l_K = L_Z l_J = l_n.$$

Algebraically, this says that the column spaces of L_X and L_Z always have the vector l_n in common, and this needs to be taken into account in adapting existing procedures to the present problem. Let us, for brevity, call this Property C.

Note that Property C implies, in particular,

$$M_{L_X} L_Z l_J = M_{L_X} l_n = 0.$$

As a consequence of Property C, some matrices involving both L_X and L_Z have reduced rank. Hence, special attention has to be paid when dealing with these matrices.

2.5 Identification

Newey and Powell (2003) and Das (2005) study identification of the unknown structural function $h(\cdot)$ in the presence of endogeneity of discrete regressors X. Florens and Malavolti (2003) and Das (2005) consider estimation in this framework. They show that nonparametric identification is achieved if the vector of instruments Z has at least as many points of support as the endogenous regressor X under a marginal covariation condition, i.e. $E[\varepsilon|Z = z] = c$, where c is a constant that is invariant with respect to Z.

Using this marginal covariation restriction, one can normalize $c = 0$, producing the system of linear equations (4). Since the conditional expectations on the left hand side and probabilities on the right hand side are observables, (4) forms a set of linear equations in the unknown $h(x_k)$, i.e., in β. Hence, the value of the vector β is identified if and only if the solution to these linear equations is unique. Assuming that equations in (4) represent the only information about $h(\cdot)$ that the data contains, point identification requires that the matrix Π has rank K.

Proposition 1 *(Newey and Powell (2003)) The necessary and sufficient condition for identification in the model $Y = h(X) + \varepsilon$, with discrete endogenous X and a discrete instrument Z satisfying $E[\varepsilon|Z] = 0$, both with finite support, is that the number of points of support of the instrument Z is at least as large as the number of points of support of endogenous X.*[4]

Hence, if $J \geq K$, β is point-identified for known (π, Π) and $\beta = (\Pi'\Pi)^{-1}\Pi'\pi$. Even if the identification condition fails, the model still has partial identifying power. Partial identification arises in models, which cannot provide the exact value of the parameter or structure of interest, but contain enough information to bound these values to informative sets. The literature on partial identification has been growing rapidly since the late 1980s. See Tamer (2010) for a detailed review.

Chesher (2004) presents the conditions under which the nonparametric model with discrete endogenous regressors partially identifies the conditional mean of the outcome by bounding its value in informative ways when the support of instruments is sparse relative to the support of endogenous regressor. If $J < K$, even though the exact value of the vector β in (4) remains unknown, we are able to bound its value by quantities which are easily estimated from the data.

2.6 Estimation

This section presents some basic estimation results under point identification. That is, we assume that $J \geq K$. Since $E[Y|X] = h(X) + E[\varepsilon|X]$, $\beta_k = h(x_k)$ can be nonparametrically estimated from the data by averaging the y_i corresponding to all x_i^s that equal x_k. Given the linear interpretation of the model, the standard OLS estimator for β is

$$\hat{\beta} = (L_X' L_X)^{-1} L_X' y, \tag{9}$$

which coincides with the standard nonparametric estimator. The important observation is that the value of the conditional mean of Y given X, does not depend on the values x_k of X and the configuration of x_k in the sample (the position of non-zero elements in the matrix L_X) does not matter. The only thing that matters is the multiplicity of each x_k in the sample. Since $n^{-1}n_k^X$ is a sample proportion, it converges in probability to p_k i.e. the probability mass on the support point x_k.

Substituting the linear model $y = L_X\beta + \varepsilon$ in (9) gives

$$\begin{aligned} \hat{\beta} &= (L_X' L_X)^{-1} L_X' y \\ &= \beta + (L_X' L_X)^{-1} L_X' \varepsilon \end{aligned}$$

[4] The result can also be found in Matzkin (2007), Chapter 73 in "Handbook of Econometrics"

and since $(n^{-1}L'_X L_X)^{-1} \to^p D_X^{-1}$ where D_X is diag(p_k), the matrix of probability masses on each point of support of X on the main diagonal with zero entries elsewhere, and $n^{-1}L'_X\varepsilon \to^p E_X[L'_X E[\varepsilon|X]] = 0$, under the null hypothesis, we have $\hat{\beta} \to^p \beta$, i.e., if X is exogenous the OLS estimator $\hat{\beta}$ is a consistent estimator of β. We can also readily establish the asymptotic distribution of the OLS estimator.

Theorem 1 *Under the assumptions above, if X is exogenous then the OLS estimator $\widehat{\beta}$ is consistent and*

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to^d N\left(0, \sigma^2 D_X^{-1}\right)$$

Remark 3 *The primitive components of the elements of $\hat{\beta}$ are sums of random numbers of i.i.d.random variables, since the multiplicities and positions of the x_k in the sample are random. At first sight, therefore, one might expect to need a central limit theorem adapted to this situation, such as those of, for example, Robbin's (1948), or Anscombe (1952), both of which deal with this case. However, the problem turns out to be more straightforward, and Theorem 1 can be proved by using a multivariate version of the Lindeberg-Feller central limit theorem (see Appendix).*

It can be shown that the covariance matrix $\sigma^2 D_X^{-1}$, under exogeneity, achieves the asymptotic Cramer-Rao bound and hence $\widehat{\beta}$ is asymptotically efficient. And, the unknown parameter σ^2 can be consistently estimated by the usual estimator in a linear regression model: $n^{-1}y'M_{L_X}y \to^p \sigma^2$.

Of course, if $E[\varepsilon|X = x_k] \neq 0$, i.e. X is endogenous, then

$$n^{-1}L'_X\varepsilon \to^p E_X[L'_X E[\varepsilon|X]] \neq 0$$

and $\widehat{\beta}$ is an inconsistent estimator for β. However, if X is endogenous the unknown function $h(\cdot)$ (or vector β) can be estimated using familiar IV methods. When the model point-identifies the structure of interest, the problem can be treated as a standard IV problem and the IV estimator for β is

$$\begin{aligned}
\hat{\beta}_{IV} &= \left(\hat{\Pi}'L'_Z L_Z\hat{\Pi}\right)^{-1}\hat{\Pi}'L'_Z L_Z\hat{\pi} \\
&= \left(L'_X L_Z\left(L'_Z L_Z\right)^{-1}L'_Z L_X\right)^{-1}L'_X L_Z\left(L'_Z L_Z\right)^{-1}L'_Z y \\
&= \left(L'_X P_{L_Z} L_X\right)^{-1}L'_X P_{L_Z} y.
\end{aligned}$$

This is the IV estimator for β in the null model $y = L_X\beta + \varepsilon$, in the presence of the instrument matrix L_Z. Even though in the nonparametric specification there is only one discrete instrument Z, we have J instrumental values $(I(Z = z_j),\ j = 1, ..., J)$ in the linear regression specification. The matrix of instruments corresponding to this interpretation of the model is L_Z, so the familiar requirements for the validity of the instruments are that $n^{-1}L'_Z L_X \to^p P$, a finite nonsingular matrix; $n^{-1}L'_Z\varepsilon \to^p 0$ and $n^{-1}L'_Z L_Z \to^p D_Z$, a positive definite matrix (Greene (1993), p.601). All these conditions are covered by Assumption 2. Note that the matrix P is the matrix of joint probabilities, and the full rank assumption for P requires: in the case $J \geq K$, that there is no non-zero $K \times 1$ vector x for which $Px = 0$, and in the case $J < K$, there is no non-zero $J \times 1$ vector z such that $P'z = 0$. The last condition follows from the fact that $D_Z = \text{diag}(q_1, ..., q_J)$ with $q_j > 0$.

The IV estimator is consistent in both scenarios: when X is exogenous and when it is endogenous, since $n^{-1}L_Z' L_X \to^p P$, $n^{-1}L_Z' L_Z \to^p D_Z$ and $n^{-1}L_Z' \varepsilon \to^p E_Z[L_Z' E[\varepsilon|Z]] = 0$. The last expression follows because of instrument exogeneity condition (2). The asymptotic normality of the IV estimator is established through:

Theorem 2 *Under assumptions above, the IV estimator $\hat{\beta}_{IV}$ is consistent and*

$$\sqrt{n}\left(\hat{\beta}_{IV} - \beta\right) \to^d N\left(0, \sigma^2 \left(P' D_Z^{-1} P\right)^{-1}\right).$$

It can be shown that the IV estimator defined for the linear representation of the nonparametric model is equivalent to the standard nonparametric estimator (see, for example Das (2005)). The advantage of our approach is that the estimator can be written in a compact matrix notation, which is easier to work with.

It is crucial to understand that because K and J are fixed, we cannot estimate the entire unknown function $h(\cdot)$, but can only learn about specific values of $h(\cdot)$ at the support points. Additional information about $h(\cdot)$, could possibly be acquired if the support of the regressor (and instrument) were assumed to be increasing with the sample size. Allowing for growing dimensions could be considered as an abstract way of generating asymptotic approximations to the distributions of estimators and might result in different limiting behaviour instead of Theorems 1 and 2. Additionally, letting both J and K grow at a rate that is proportional to n, would have an impact on the identification analysis. It is possible that a model that is only set identified $(J < K)$ in small samples, point identifies $h(\cdot)$ in large samples if K is fixed and J increases with n, or if J grows faster than K. Therefore, considering such increasing dimensions might be an interesting extension of our work, but this topic is left for further research.

3 Testing for exogeneity under point identification

Assume that $J \geq K$ and the model point identifies the unknown function of interest $h(\cdot)$ by Proposition 1. The OLS estimator $\widehat{\beta}$ is consistent and efficient if X is exogenous, but inconsistent otherwise. The IV estimator is consistent in both cases, but inefficient if X is exogenous. For this situation, then, the test is really just to decide which estimator to use (OLS or IV).

The standard Wu-Hausman-type statistic for testing exogeneity in this context is based on a quadratic form in the difference between the two estimators $\hat{\beta}_{IV}$ and $\hat{\beta}$, namely

$$\hat{\beta}_{IV} - \hat{\beta} = \left(L_X' P_{L_Z} L_X\right)^{-1} L_X' P_{L_Z} M_{L_X} y, \tag{10}$$

with the matrix of the quadratic form equal to the inverse of $Cov(\hat{\beta}_{IV} - \hat{\beta})$, in order to produce a χ^2 variable asymptotically (Hausman (1978)). The covariance matrix of the difference is given by

$$Cov(\hat{\beta}_{IV} - \hat{\beta}) = \left(L_X' P_{L_Z} L_X\right)^{-1} L_X' P_{L_Z} M_{L_X} P_{L_Z} L_X \left(L_X' P_{L_Z} L_X\right)^{-1}. \tag{11}$$

However, in this case, Property C implies that this covariance matrix is singular. To see this, observe that

$$\begin{aligned} l_K' \left(L_X' P_{L_Z} L_X\right)(\hat{\beta}_{IV} - \hat{\beta}) &= l_K' L_X' P_{L_Z} M_{L_X} y \\ &= l_n' P_{L_Z} M_{L_X} y \;\; (L_X l_K = l_n) \\ &= l_n' M_{L_X} y \;\; (P_{L_Z} l_n = l_n) \\ &= 0 \;\; (M_{L_X} l_n = 0). \end{aligned}$$

That is, for all L_X and L_Z there is an exact linear relation between the elements of $\hat{\beta}_{IV} - \hat{\beta}$, so its covariance matrix will always be singular.

We therefore need to adapt the Wu-Hausman test statistic to this situation. To do so we simply replace the inverse of the covariance matrix - the matrix that would normally be used in the quadratic form to produce an asymptotically χ^2 test statistic - by a generalized inverse of that matrix. The covariance matrix in (11) can be written as

$$S = \left(L_X' P_{L_Z} L_X\right)^{-1} C_K \left[C_K' L_X' P_{L_Z} M_{L_X} P_{L_Z} L_X C_K\right] C_K' \left(L_X' P_{L_Z} L_X\right)^{-1},$$

since $M_{L_X} P_{L_Z} L_X [l_K, C_K] = [0, M_{L_X} P_{L_Z} L_X C_K]$ and $[l_K, C_K]^{-1} = [K^{-1} l_K, C_K]'$ (see section 2.1 for notation). The middle matrix $C_K' L_X' P_{L_Z} M_{L_X} P_{L_Z} L_X C_K$ is a $(K-1)-$square matrix of full rank. Thus, the covariance matrix can be expressed as a matrix of the form $S = A^{-1} C B C' A^{-1}$, where C is $m \times p$, $C'C = I_p$, B is $p \times p$ nonsingular and symmetric, and A is $m \times m$ nonsingular and symmetric. The generalized inverse of a matrix with this form is $S^+ = A C B^{-1} C' A$. To verify this it is sufficient to check that the two conditions that define a generalized inverse, i.e. $SS^+S = S$ and $S^+SS^+ = S^+$, both hold.

Therefore, the generalized inverse of the covariance matrix is

$$S^+ = \left(L_X' P_{L_Z} L_X\right) C_K \left[C_K' L_X' P_{L_Z} M_{L_X} P_{L_Z} L_X C_K\right]^{-1} C_K' \left(L_X' P_{L_Z} L_X\right).$$

Using this matrix to define the test statistic, we have

$$\begin{aligned} T_n^* &= y' M_{L_X} P_{L_Z} L_X C_K \left[C_K' L_X' P_{L_Z} M_{L_X} P_{L_Z} L_X C_K\right]^{-1} C_K' L_X' P_{L_Z} M_{L_X} y \\ &= y' W_{XZ} \left(W_{XZ}' W_{XZ}\right)^{-1} W_{XZ}' y \end{aligned}$$

where $W_{XZ} = M_{L_X} P_{L_Z} L_X C_K$ is $n \times (K-1)$.

Scaling to eliminate σ^2, we propose the test-statistic

$$T_n = \frac{y' W_{XZ} \left(W_{XZ}' W_{XZ}\right)^{-1} W_{XZ}' y}{n^{-1} y' M_{L_X} y}. \tag{12}$$

Observe that the values x_k of X and z_j of Z do not appear in the test statistic, nor does their configuration in the sample matter. The only things that appear are the multiplicities of each value in the sample, the n_k^X and n_j^Z, and the multiplicity of the joint event $(X = x_k, Z = z_j)$, n_{jk}. Note that the numerator of the modified version of T_n is easily computed from a linear regression of y on W_{XZ}. Since W_{XZ} is easy to construct in practice, the value of the test-statistic might be efficiently calculated by any statistical software package.

Remark 4 *Using the generalized inverse is not the only way to deal with singularity of the covariance matrix. The naive approach would be to reduce the dimension of the test-statistic by eliminating for example the first element in the difference (10) and picking up the lower-right corner of the covariance matrix in (11). Then the Wu-Hausman test-statistic of reduced dimension would follow standard results. An alternative approach would be to use the Moore-Penrose inverse of the covariance matrix (built into all econometric software). All three approaches give similar values of the test-statistic, thus in applications, the researcher could choose the method that is most convenient.*

3.1 Asymptotic distribution under the null hypothesis

To discuss the asymptotic distribution of T_n, define the $(J-1) \times 1$ vector

$$z_n = C_J' L_Z' M_{L_X} y = C_J' L_Z' M_{L_X} \varepsilon. \tag{13}$$

The primitive components of z_n are the two vectors $u_n = L_Z' \varepsilon$ and $v_n = L_X' \varepsilon$. Thus, we first consider the asymptotic behaviour of these two vectors, i.e. the joint asymptotic distribution of

$$\frac{1}{\sqrt{n}} w_n = \frac{1}{\sqrt{n}} \begin{pmatrix} u_n \\ v_n \end{pmatrix}.$$

This is given in:

Lemma 1 *Under H_0 and the given assumptions,*

$$\frac{1}{\sqrt{n}} w_n \to^d N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{bmatrix} D_Z & P \\ P' & D_X \end{bmatrix} \right).$$

This result will also be useful in the set-identified model later. Now, z_n is a linear function of u_n and v_n,

$$\frac{1}{\sqrt{n}} z_n = \frac{1}{\sqrt{n}} C_J' \left(u_n - L_Z' L_X \left(L_X' L_X \right)^{-1} v_n \right)$$

with

$$p \lim_{n \to \infty} \frac{L_Z' L_X}{n} \left(\frac{L_X' L_X}{n} \right)^{-1} = P D_X^{-1}.$$

We therefore immediately obtain

Lemma 2 *Under H_0 and the given assumptions,*

$$\frac{1}{\sqrt{n}} z_n \to^d N(0, \sigma^2 \Sigma)$$

where $\Sigma = C_J'(D_Z - P D_X^{-1} P')C_J$ is positive definite.

The numerator of the proposed test-statistic in (12) is a quadratic form in z_n :

$$T_n^* = z_n' A_n B_n^{-1} A_n' z_n$$

where

$$A_n = C_J' \left(L_Z' L_Z\right)^{-1} L_Z' L_X C_K$$

is a $(J-1) \times (K-1)$ matrix with probability limit equal to

$$A = C_J' D_Z^{-1} P C_K,$$

and

$$B_n = A_n' \left(C_J' L_Z' M_{L_X} L_Z C_J\right) A_n$$

is $(K-1)-$square matrix.

Using these results we obtain the asymptotic distribution of T_n under the null hypothesis:

Theorem 3 *Under H_0, and the assumptions above,*

$$T_n \to^d \chi^2_{K-1}.$$

The asymptotic behaviour of the test-statistic under the null hypothesis is fully characterized by the χ^2 distribution. Therefore, for practical applications, the critical values can be easily obtained from statistical tables. The accuracy of this asymptotic result is examined in Section 5.

3.2 Asymptotics under the alternative hypothesis

In this section, we establish the asymptotic distribution of the test-statistic under a sequence of local alternatives, and in order to show that the proposed test is consistent, i.e. the power of the test approaches 1 as $n \to \infty$, we discuss the asymptotic behaviour of T_n under a fixed alternative hypothesis.

3.2.1 Local alternatives

Let $m(X, V)$ be a bounded function, depending on X and another variable V, which does not appear in the model and is independent of X. Assume now that the conditional expectation of the error term is given by:

$$E[\varepsilon|X = x, V = v] = E[\varepsilon|X] = m(x, v),$$

and define the K vector

$$m = \begin{bmatrix} m(x_1, v) \\ ... \\ m(x_K, v) \end{bmatrix}.$$

In the linear representation of the model, we have

$$E[\varepsilon|\mathcal{X} = L_X x] = L_X m. \tag{14}$$

To derive the asymptotic distribution of the test statistic under the alternative hypothesis, consider the sequence of local alternatives in which $E[\varepsilon|\mathcal{X} = L_X x] = n^{-\frac{1}{2}} L_X m$.

Theorem 4 *Under the sequence of local alternatives to (14) and the assumptions above, the test statistic T_n converges to a non-central $\chi^2_{K-1}(\delta_L)$ distribution, with the noncentrality parameter*

$$\delta_L = \frac{\mu' A \left(A'\Sigma A\right)^{-1} A'\mu}{\sigma^2}$$

where $\mu = -C'_J P m$, $A = C'_J D_Z^{-1} P C_K$ and $\Sigma = C'_J[D_Z - P D_X^{-1} P'] C_J$.

The proof of Theorem 4 is based on familiar results for quadratic forms in normal variables with non-zero mean. The asymptotic behaviour of the test-statistic is captured by the non-central χ^2 distribution. For a given size of test, the power increases with noncentrality parameter δ_L. The value of this parameter depends on the the distance between an inconsistent OLS and consistent IV estimators. Hence, the test is more powerful if the probability limit of the OLS estimator is far from the true value of the parameter of interest.

3.2.2 Fixed alternatives

Let us next consider fixed alternatives of form $H_1 : E(\varepsilon_i|X = x_k) = m(x_k, v_i)$. Building on the results used in the previous section, by a simple generalization of Lemma 1, we obtain

$$\frac{1}{\sqrt{n}} \begin{bmatrix} L'_Z\varepsilon \\ L'_X\varepsilon \end{bmatrix} \rightarrow^d N\left(\left(\begin{array}{c} 0 \\ \sqrt{n} D_X m \end{array}\right), \sigma^2 \left(\begin{array}{cc} D_Z & P \\ P' & D_X \end{array}\right)\right)$$

Additionally

$$\frac{1}{\sqrt{n}} z_n \rightarrow^d N(\mu_S, \sigma^2\Sigma)$$

with $\mu_S = -\sqrt{n}C_J'Pm := \sqrt{n}\mu$, i.e. the mean is proportional to the square root of the sample size.

Therefore, under fixed alternatives the test statistic in (12) converges to a noncentral Chi-square distribution with $(K-1)$ degrees of freedom and noncentrality parameter δ_F equal to

$$\delta_F = \frac{\mu_S' A \left(A'\Sigma A\right)^{-1} A'\mu_S}{\sigma^2} = n\delta_L$$

The following proposition establishes the consistency of the test against a fixed alternative hypothesis.

Proposition 2 *Under fixed alternatives and the earlier assumptions, the proposed test is consistent, i.e., for any fixed constant c_α,*

$$\Pr\left(T_n > c_\alpha\right) \to 1 \textit{ as } n \to \infty$$

Since μ_S is a multiple of $\sqrt{n}$, the noncentrality parameter is proportional to the sample size. This implies that if the alternative hypothesis holds, as $n \to \infty$, the chi-square distribution moves to the right and the probability of rejecting a false null hypothesis increases, i.e. $p\lim_{n\to\infty} \Pr(\chi^2_{K-1}(\delta_F) > c_\alpha) = 1$. Hence, as $n \to \infty$, the power of the test converges to 1 and the test is said to be consistent.

Remark 5 *Under the alternative hypothesis we will have $E[\varepsilon|X = x_k] \neq 0$ for at least one value of k. For some specifications of how these values are determined the tests proposed above will have no power. This occurs if, when the null hypothesis fails,*

$$E[Y|X = x_k] = h(x_k) + \mu(x_k)$$

where $\mu(x_k) = E[\varepsilon|X = x_k]$ depends only on x_k. In this case we will have the model

$$y = L_X(\beta + \mu) + \widetilde{\varepsilon}$$

where $\widetilde{\varepsilon} = \varepsilon - \mu$, which is identical to the original model with the unknown h replaced by the also-unknown $h + \mu$. Thus, it is not surprising that the test should have power equal to size in this circumstance.

4 Testing for exogeneity under set identification

In this situation $(J < K)$ there is no consistent estimator (in the conventional sense) for β if X is endogenous, so in this case the test is to decide whether point estimation of β is even possible. When $J < K$ the Wu-Hausman approach to testing H_0 is not available. However, assuming the existence of an instrument Z with the properties given above, β is constrained to satisfy the linear equations $\pi = \Pi\beta$, but is not point identified by them. That is, there is a set of vectors β, a subset of $\mathbb{R}^K$, that satisfy these equations, of dimension $K-J$. The model maintains that β belongs to this set, and H_0 says that $E[\mathcal{Y}|\mathcal{X}=L_X x]=L_X\beta$.

Now, consider the empirical counterpart of the system $\pi=\Pi\beta$, namely $\widehat{\pi}=\widehat{\Pi}\beta$, and the vector β that, among all solutions to this system, minimizes $(y-L_X\beta)'(y-L_X\beta)$. That is, define

$$\widehat{\beta}_Z=\arg\min_{\beta:\widehat{\pi}=\widehat{\Pi}\beta}(y-L_X\beta)'(y-L_X\beta).$$

Straightforward algebra gives

$$\begin{aligned}\widehat{\beta}_Z &= \widehat{\beta}+\left(L_X'L_X\right)^{-1}\widehat{\Pi}'\left(\widehat{\Pi}\left(L_X'L_X\right)^{-1}\widehat{\Pi}'\right)^{-1}\left(\widehat{\pi}-\widehat{\Pi}\widehat{\beta}\right)\\ &= \widehat{\beta}+\left(L_X'L_X\right)^{-1}L_X'L_Z\left(L_Z'P_{L_X}L_Z\right)^{-1}L_Z'M_{L_X}y,\end{aligned}$$

where $\widehat{\beta}$ is the OLS estimator defined earlier. The minimum achieved by this choice for β is therefore

$$\begin{aligned}Q_n &= (y-L_X\widehat{\beta}_Z)'(y-L_X\widehat{\beta}_Z)\\ &= y'M_{L_X}y+y'M_{L_X}L_Z\left(L_Z'P_{L_X}L_Z\right)^{-1}L_Z'M_{L_X}y.\end{aligned}$$

Intuitively, a large value for this minimum sum of squares is evidence against H_0, because it means that, among all solutions to $\widehat{\pi}=\widehat{\Pi}\beta$, none produces a small value of $(y-L_X\beta)'(y-L_X\beta)$. This suggests, not that $\pi\neq\Pi\beta$, because this is ruled out, but rather that $E[\mathcal{Y}|\mathcal{X}=L_X x]\neq L_X\beta$, i.e. that the null hypothesis is false. Normalizing Q_n by dividing by $n^{-1}y'M_{L_X}y$, this argument suggests rejecting H_0 when the statistic

$$R_n=\frac{y'M_{L_X}L_Z\left(L_Z'P_{L_X}L_Z\right)^{-1}L_Z'M_{L_X}y}{n^{-1}y'M_{L_X}y}$$

is large.

Now, in view of Property C,

$$M_{L_X}L_Z[l_J,C_J]=[M_{L_X}l_n,M_{L_X}L_ZC_J]=[0,M_{L_X}L_ZC_J]$$

and, the $(2,2)$ block of

$$[[l_J,C_J]'\left(L_Z'P_{L_X}L_Z\right)[l_J,C_J]]^{-1}=\left[\begin{array}{cc}n & l_n'L_ZC_J\\ C_J'L_Z'l_n & C_J'L_Z'P_{L_X}L_ZC_J\end{array}\right]^{-1}$$

is given by

$$\left(C_J'L_Z'\left[P_{L_X}-P_{l_n}\right]L_ZC_J\right)^{-1}.$$

Thus, after taking account of Property C, R_n reduces to

$$R_n=\frac{y'M_{L_X}L_ZC_J\left(C_J'L_Z'\left[P_{L_X}-P_{l_n}\right]L_ZC_J\right)^{-1}C_J'L_Z'M_{L_X}y}{n^{-1}y'M_{L_X}y}\qquad(15)$$

with the middle matrix being $(J-1)$ square. Thus, although at first sight a quadratic form involving J variables, the numerator of R_n in fact involves only $J-1$ terms.

4.1 Asymptotic distribution under the null hypothesis

The following theorem gives the asymptotic distribution of the test statistic under the null hypothesis.

Theorem 5 *Under H_0 and the assumptions above,*

$$R_n \to^d \sum_{j=1}^{J-1} \omega_j \chi_j^2(1)$$

where the ω_j are positive eigenvalues satisfying

$$\det[\Sigma - \omega\Omega] = 0$$

with

$$\Omega = p \lim_{n\to\infty} \frac{1}{n}[C_J' L_Z' (P_{L_X} - P_{l_n}) L_Z C_J]$$

and

$$\Sigma = p \lim_{n\to\infty} \frac{1}{n}[C_J' L_Z' M_{L_X} L_Z C_J]$$

and the $\chi_j^2(1)$ variables are independent copies of a χ_1^2 random variable.

The proposed test-statistic converges to a quadratic form in a normal vector z, and the distribution of that quadratic form is given by the distribution of a weighted sum of chi-square (1) random variables.

The asymptotic distribution of the proposed test with discrete regressors and instruments is similar to the distribution obtained by Blundell and Horowitz (2007) for the continuous case. Their test-statistic follows asymptotically the distribution of an infinite sum of weighted chi-square variables with 1 degree of freedom. When calculating the critical values, they face the additional problem of approximating an infinite sum by a finite number of terms. In the discrete case, the asymptotic distribution is more straightforward, since it is based on a finite sum of terms due to the discrete nature of variables. Nonetheless, the distribution theory for such variables is complicated, and there is an incentive to use approximations, and several have been discussed extensively in the literature. In Section 4.3 we discuss the approximation proposed by Hall (1983) and further explored by Buckley and Eagleson (1988), which allows us to compute the critical values in practical applications.

4.2 Asymptotics under the alternative hypothesis

This section obtains the asymptotic distribution of R_n under a sequence of local alternatives. The test is also shown to be consistent against fixed alternatives.

4.2.1 Local alternatives

Consider the sequence of local alternatives to (14). Using the vector z_n defined in (13), the numerator of the test statistic in (15) can be written as

$$R_n^* = z_n' \left(C_J' L_Z' (P_{L_X} - P_{l_n}) L_Z C_J\right)^{-1} z_n$$

with $n^{-\frac{1}{2}} z_n \to^d N(\mu, \sigma^2 \Sigma)$ with $\mu = -C_J' Pm$ and $\Sigma = C_J'[D_Z - PD_X^{-1}P']C_J$ as before.

The following theorem establishes the asymptotic distribution of the test-statistic under local alternatives.

Theorem 6 *Under the sequence of local alternatives to (14) and the assumptions above, the test statistic R_n converges to a distribution of a weighted sum of non-central chi-square random variables:*

$$R_n \to^d \sum_{j=1}^{J-1} \omega_j \chi_1^2(\delta_j^2)$$

with the noncentrality parameters

$$(\delta_1, ..., \delta_{J-1})' = S' \Sigma^{-\frac{1}{2}} \mu = -S' \Sigma^{-\frac{1}{2}} C_J' Pm$$

where S denotes the orthogonal matrix of the eigenvectors of $\Sigma^{-\frac{1}{2}} \Omega^{-1} \Sigma^{-\frac{1}{2}}$.

Under local alternatives, the test-statistic asymptotically follows the distribution of a weighted sum of non-central chi-square (1) variables. This result again corresponds to the distribution obtained by Blundell and Horowitz (2007) for the continuous case.

4.2.2 Fixed alternatives

Under fixed alternatives (14) the test statistic in (15) converges to a weighted sum of noncentral $\chi^2_{(1)}$ random variables, $\sum_{j=1}^{J-1} \omega_j \chi_1^2(\delta_j^2)$, with

$$(\delta_1, ..., \delta_{J-1})' = -\sqrt{n} S' \Sigma^{-\frac{1}{2}} C_J' Pm$$

Since the noncentrality parameter is again proportional to the sample size for each term, the power of the test goes to 1 as $n \to \infty$.

Proposition 3 *Under fixed alternatives, the proposed test is consistent, i.e., for any fixed constant c_α,*

$$\Pr(R_n > c_\alpha) \to 1 \text{ as } n \to \infty$$

4.3 Computation of critical values

The asymptotic distribution of the test-statistic is non-standard and depends on the weights ω_j, which, in practice, need to be estimated from the data. Since we cannot provide statistical tables with the appropriate tail probabilities and cut off points, it is essential to find a quick technique for calculating the critical values of the proposed test.

Although the distribution of the weighted sum of chi-square variables has been studied in the literature since 1960's and the explicit formulas for the probability density function and a cumulative distribution function have been derived, they are rather complicated and difficult to handle in empirical applications. From the practical point of view, in order to calculate the critical values for the proposed test, it is crucial to be able to approximate the process of interest by a well known structure. Alternatively, one could use the inverse interpolation procedure of finding the critical values proposed by Sheil and Muircheartaigh (1977). However, this method is computationally intensive and requires specifying the upper and lower bounds on the weights, which we would like to avoid.

There are numerous ways of computing the critical values in this case. Letting $\widehat{\omega}_j$ be consistent estimators of the weights ω_j under H_0, the distribution of $\sum_{j=1}^{J-1} \widehat{\omega}_j \chi^2_j(1)$ can be simulated and appropriate $1-\alpha$ quantiles can be used as critical values in the standard rejection rule. However, our experiments show that this approach is computationally intensive and time consuming. The second method involves simulating the quadratic form $z'\widehat{\Omega}^{-1}z$ with $z \sim N(0,\widehat{\Sigma})$ and computing the quantiles. This method delivers satisfactory results and reduces the simulation time significantly. The third method is based on using an approximation to the distribution of a weighted sum of chi-square variables.

Even though a linear combination of independent chi-squared variables is, under regularity conditions, known to be asymptotically normally distributed when the sample size tends to ∞ (Johnson, Kotz and Balakrishnan (1994), p.444), the simulations reveal the unsatisfactory performance of the normal approximation. Hence, we suggest applying the approximation proposed by Hall (1983) and further explored by Buckley and Eagleson (1988), where the distribution of a weighted sum of χ^2_1 random variables is approximated by the distribution of a variate $\tilde{R} = a\chi^2_v + b$ by choosing (a, b, v) so that the first three cumulants of R and $\tilde{R}$ agree.

The cumulants κ_l of a random variable are defined via the cumulant-generating function $K(t)$, which is the logarithm of the characteristic function $\phi(t)$ with the following expansion (Muirhead (1982), p.40)

$$K(t) = \log(\phi(t)) = \sum_{l=1}^{\infty} \kappa_l \frac{(it)^l}{l!}.$$

Since the characteristic function $\phi(t)$ of a chi-square random variable with r degrees of freedom is

$$\phi(t) = (1-2it)^{-\frac{r}{2}}$$

the cumulant generating function $K(t)$ of $\chi^2_{(r)}$ variable is

$$K(t) = -\frac{r}{2}\log(1-2it) = \frac{1}{2}r\sum_{l=1}^{\infty}\frac{(2it)^l}{l}$$

and the cumulants κ_l solve

$$\sum_{l=1}^{\infty}\kappa_l\frac{(it)^l}{l!} = \frac{1}{2}r\sum_{l=1}^{\infty}\frac{(2it)^l}{l}.$$

Let $R = \sum_{j=1}^{J-1}\omega_j\chi^2_j(1)$. The cumulants of this chi-squared-type mixture are given by[5]

$$\kappa_l(R) = 2^{l-1}(l-1)!\sum_{j=1}^{J-1}\omega_j^l.$$

Therefore, the first three cumulants of R are

$$\begin{aligned}
\kappa_1(R) &= E(R) = \sum_{j=1}^{J-1}\omega_j = trace(\Sigma\Omega^{-1}) \\
\kappa_2(R) &= Var(R) = 2\sum_{j=1}^{J-1}\omega_j^2 = 2trace\left(\left(\Sigma\Omega^{-1}\right)^2\right) \\
\kappa_3(R) &= E\left((R-E(R))^3\right) = 8\sum_{j=1}^{J-1}\omega_j^3 = 8trace\left(\left(\Sigma\Omega^{-1}\right)^3\right).
\end{aligned}$$

The cumulants of $\tilde{R} = a\chi^2_v + b$ are:

$$\kappa_1(\tilde{R}) = av + b,\ \ \kappa_2(\tilde{R}) = 2a^2v,\ \ \kappa_3(\tilde{R}) = 8a^3v.$$

To determine the parameters a, b and v we set $\kappa_m(\tilde{R}) = \kappa_m(R)$ for $m = 1,2,3$ which leads to

$$\begin{aligned}
a &= \frac{\kappa_3(R)}{4\kappa_2(R)} \qquad (16) \\
b &= \kappa_1(R) - \frac{2\kappa_2^2(R)}{\kappa_3(R)} \\
v &= \frac{8\kappa_2^3(R)}{\kappa_3^2(R)}.
\end{aligned}$$

Hence the approximate cumulative distribution of R is

$$F_R(t) = \Pr(R \le t) \approx \Pr(\tilde{R} \le t) = \Pr\left(\chi^2_v \le \frac{t-b}{a}\right).$$

[5] See Severini (2005), Theorem 8.5, p. 245

The critical value c_α solves

$$1 - \Pr\left(\chi^2_v \leq \frac{c_\alpha - b}{a}\right) = \alpha$$

for $\alpha = 1\%$, 5% or 10%.

Note that parameter v is typically not an integer and the χ^2_v distribution here is in fact a gamma distribution with parameters $\frac{1}{2}$ and $\frac{v}{2}$. In practice, the matrix $\Sigma\Omega^{-1}$ is unknown and, in order to calculate the values of parameters in (16), it has to be replaced by its consistent estimate:

$$C_J' L_Z' M_{L_X} L_Z C_J \left[C_J' L_Z' (P_{L_X} - P_{l_n}) L_Z C_J\right]^{-1}.$$

An alternative (and popular) procedure of obtaining the critical values, based on the numerical inversion of the characteristic function, was proposed by Imhof (1961). This procedure is much more computationally intensive, since it requires the knowledge of all eigenvalues of $\Sigma\Omega^{-1}$, while for the three-cumulants approximation only the traces of powers of this matrix are needed.

5 Monte Carlo simulations

In this section, we discuss the results of Monte Carlo simulations designed to examine the finite sample size and power properties of the proposed tests. We modify Blundell and Horowitz's (2007) setup by generating X and Z as discrete random variables.

5.1 Simulation design

In the experiments, realizations of (X, Z) are generated as $Z = Binomial(J-1, p_Z)$ with $p_Z = 0.5$ and X is a function of Z such that

$$X = x_k \text{ if } a < X^* \leq b$$

where a and b are constants, and $X^* = \alpha Z + (1-\alpha^2)^{1/2}\nu$ with $v \sim N(0,1)$ and $\alpha \in \{0.35, 0.7\}$. Note that α measures the strength of the relationship between X and Z. Weak instruments are characterized by $\alpha = 0.35$ and $\alpha = 0.7$ characterizes strong instruments. The realizations of a continuous outcome Y are generated from

$$Y = \theta_0 + \theta_1 X + \sigma_\varepsilon \varepsilon$$

where $\varepsilon = \eta v + (1-\eta^2)^{\frac{1}{2}} u$ with $u \sim N(0,1)$ and $\theta_0 = 0$, $\theta_1 = 0.5$ and $\sigma_\varepsilon = 0.2$. The parameter η measures the strength of the relationship between X and ε, and its value varies across experiments. The null hypothesis is true if $\eta = 0$ and false otherwise. The experiments use sample sizes of $n = 50, 100, 200, 400$ and 1000 observations and there are 2000 Monte Carlo replications in each experiment.

5.2 Size analysis $J \geq K$

Recall that under the null hypothesis $T_n \to^d \chi^2_{K-1}$, so the critical values are easily obtained from statistical tables. For the size analysis, $\eta = 0$ and the errors are generated as $N(0,1)$. The empirical size of the proposed test for different combinations of J and K (satisfying $J \geq K$) is presented in Tables 1 and 2.

K=2		J=2			J=3			J=4		
α	sample size	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.35	50	1.20	5.25	10.25	0.85	5.85	10.80	0.95	5.05	10.20
	100	0.90	4.95	10.05	1.00	4.95	9.35	0.95	5.35	10.50
	200	1.30	5.10	10.20	1.10	5.05	10.45	1.10	5.10	11.20
	400	0.85	5.20	10.20	1.10	5.05	10.45	1.10	5.10	10.25
0.7	50	1.40	5.70	10.95	1.25	5.05	10.05	1.05	5.60	10.50
	100	0.80	4.50	9.55	0.85	4.95	10.10	1.25	4.95	9.65
	200	1.10	4.55	10.80	1.25	4.85	10.10	0.95	5.40	9.70
	400	1.25	5.20	10.10	0.95	4.95	10.50	1.10	5.15	10.40

Table 1: Proportion of rejections under the null hypothesis; K=2

K=3		J=3			J=4			J=5		
α	sample size	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.35	50	0.85	5.50	11.50	0.75	4.75	9.35	1.10	5.40	10.20
	100	0.95	5.35	10.75	0.85	5.35	11.20	1.05	5.45	10.30
	200	0.80	4.75	10.10	1.25	5.25	10.95	1.25	5.50	10.45
	400	0.85	4.85	10.25	1.10	5.05	9.85	1.20	5.10	9.55
0.7	50	0.85	5.70	11.40	0.95	5.60	10.15	0.90	5.10	10.65
	100	1.05	5.90	11.20	1.15	5.15	10.70	0.95	5.15	10.30
	200	1.00	5.40	10.65	1.25	5.05	10.35	1.10	5.75	10.70
	400	1.45	5.80	10.65	0.95	5.45	10.55	1.05	5.10	9.65

Table 2: Proportion of rejections under the null hypothesis; K=3

The empirical size is reasonably close to the nominal values of 1%, 5% and 10%, even in small samples of 50 observations. The size seems not very sensitive to changes in the number of points of support of the endogenous regressor and instrument and do not vary with the strength of instrument.

5.3 Power analysis $J \geq K$

For the power analysis, the errors are generated as $\varepsilon = \eta v + (1-\eta^2)^{\frac{1}{2}} u$, $u \sim N(0,1)$. Recall that this specification excludes the alternatives with $E[\varepsilon | X = x_k] = \mu(x_k)$ in which the power is equal to the size of the test. The results of power analysis at 5% significance level for different sample sizes are summarized in Figures 1 and 2.

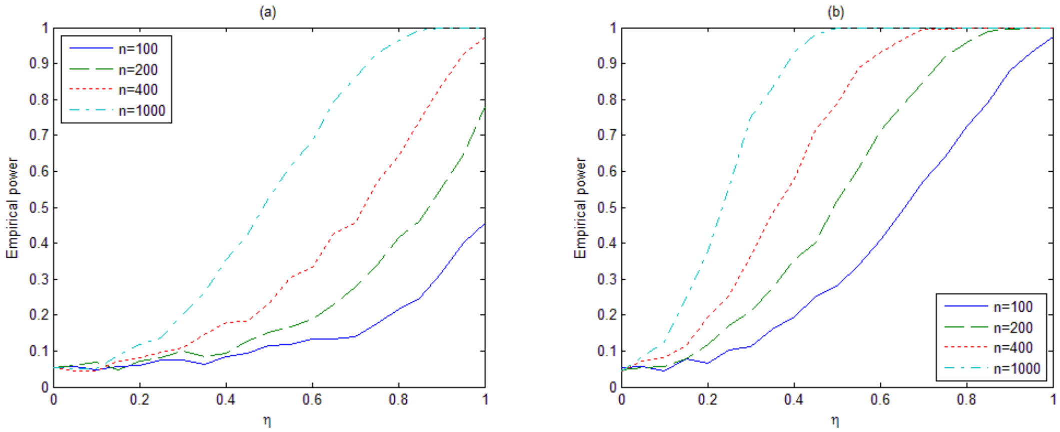

Figure 1: Empirical power for K=2 and J=2 with weak (a) and strong (b) instruments

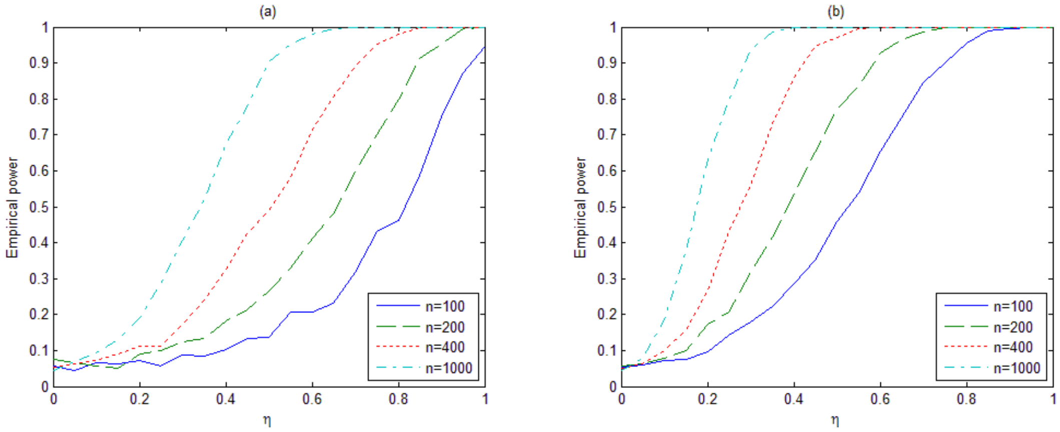

Figure 2: Empirical power for K=3 and J=3 with weak (a) and strong (b) instruments

The proposed test exhibits satisfactory power properties. The empirical power increases with a sample size and converges to 1 quickly. For a fixed number of support points of the endogenous regressor and instrument, the empirical power is higher if the instrument used in experiment are strong. The test has also higher power if the support of endogenous regressor is larger.

Figures 3 and 4 show how the empirical power changes with the number of points of support of the instrument.

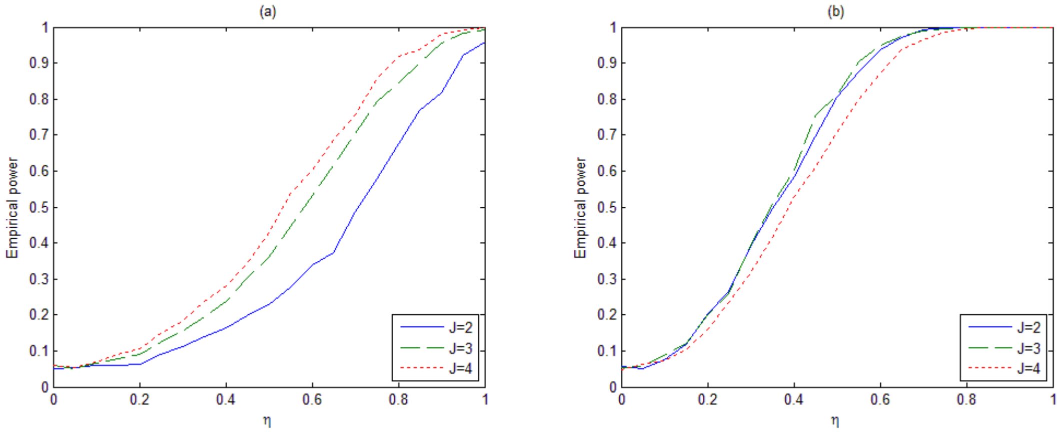

Figure 3: Empirical power for K=2 and n=400 with weak (a) and strong (b) instruments

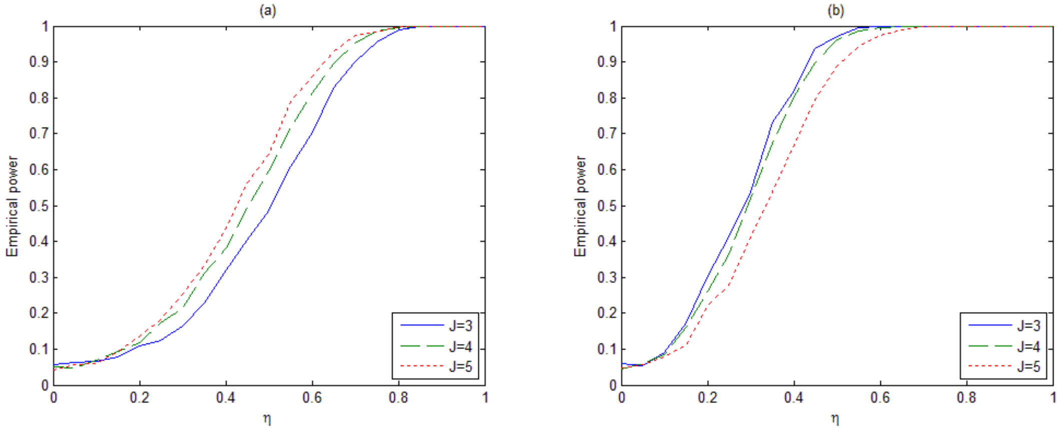

Figure 4: Empirical power for K=3 and n=400 with weak (a) and strong (b) instruments

If the instrument is weak, for fixed K, the empirical power of the test increases when additional point of support is added. Therefore, for weak instruments, the larger the support of Z, the more powerful the test is. This suggests that in practice the researcher should look for an instrument with many support points to increase the probability of detecting the endogeneity of regressor.

On the other hand, if the instrument is strong, the empirical power remains roughly the same if the difference between the support of X and Z is small, but decreases with the gap between J and K.

5.4 Size analysis $J < K$

We have experimented with different methods of computing the critical values for the proposed test. The three methods proposed in Section 4.3 produce very similar results for the empirical size and power of the test. In this section, we present the results based on the chi-square approximation, which minimizes the computational time. The empirical size of the proposed test is presented in Tables 3 and 4.

K=5		J=2			J=3			J=4		
α	sample size	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.35	50	0.85	5.80	11.10	1.25	5.30	9.65	1.45	5.65	10.80
	100	1.10	5.00	10.45	1.20	5.65	10.85	0.85	5.25	10.80
	200	0.90	4.85	10.00	0.80	4.55	9.65	0.95	4.55	9.65
	400	0.85	5.50	10.50	1.15	6.15	10.55	1.10	5.50	9.75
0.7	50	1.30	5.95	11.20	1.05	5.65	11.30	1.35	5.60	10.80
	100	1.20	6.10	11.50	1.35	5.85	11.35	1.55	5.80	11.20
	200	1.15	5.80	10.90	1.05	4.75	9.70	1.10	5.35	10.35
	400	1.20	5.65	9.80	1.05	5.30	10.10	0.95	4.95	10.50

Table 3: Proportion of rejections under the null hypothesis; K=5

K=6		J=3			J=4			J=5		
α	sample size	1%	5%	10%	1%	5%	10%	1%	5%	10%
0.35	50	1.20	5.80	11.10	1.25	6.10	12.05	1.60	5.90	11.20
	100	1.20	5.70	10.50	0.85	5.25	10.95	1.10	6.15	11.50
	200	1.05	4.80	10.15	1.00	5.60	11.10	1.50	5.55	10.40
	400	0.95	4.65	9.80	0.90	5.05	9.85	0.90	5.25	10.50
0.7	50	1.20	6.15	11.30	1.10	5.05	9.25	1.25	5.85	9.75
	100	1.05	5.35	10.55	0.95	4.75	10.30	1.60	5.70	10.15
	200	1.15	5.10	9.85	0.85	5.05	9.90	1.65	5.90	11.20
	400	0.90	5.40	10.75	1.05	5.65	11.10	0.95	5.20	10.40

Table 4: Proportion of rejections under the null hypothesis; K=6

.

The test has adequate size in all cases, even in the small samples of 50 observations. The size is not sensitive to changes in the number of points of support and the strength of the relationship between endogenous regressor and the instrument.

5.5 Power analysis $J < K$

The results of a power analysis at 5% significance level are presented in Figures 5 and 6.

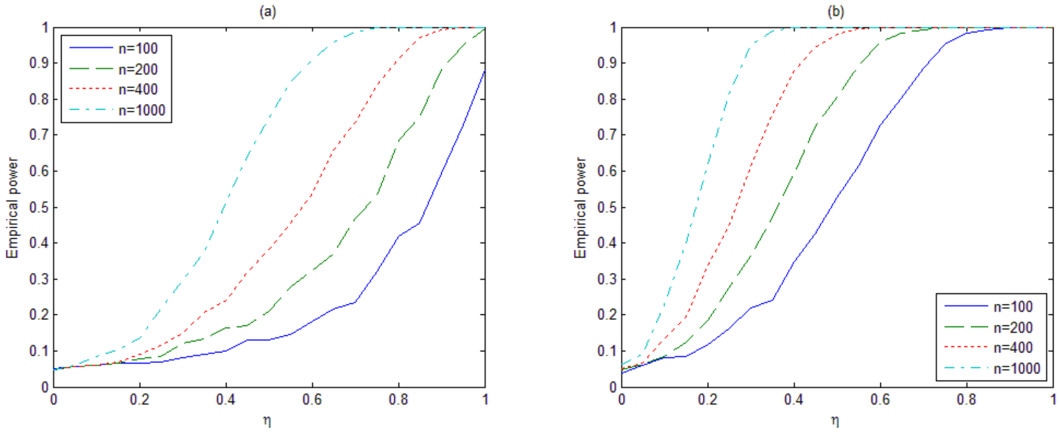

Figure 5: Empirical power for K=5 and J=2 with weak (a) and strong (b) instruments

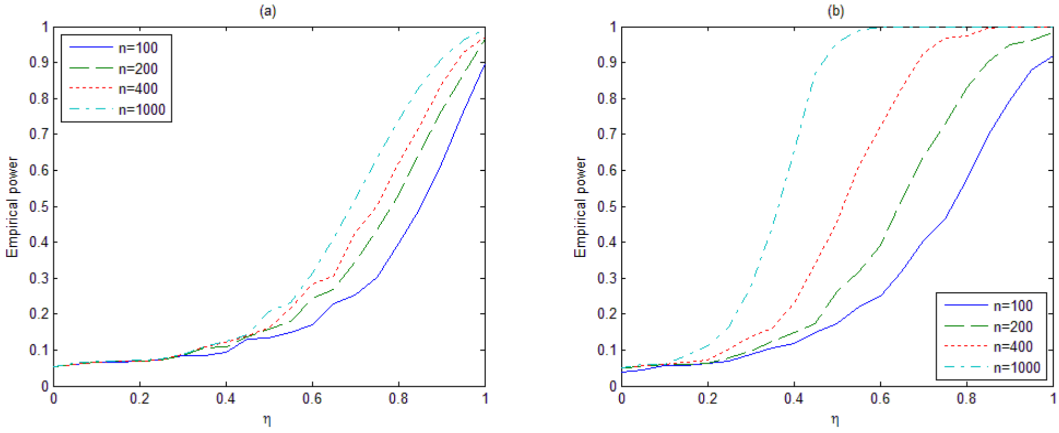

Figure 6: Empirical power for K=6 and J=3 with weak (a) and strong (b) instruments

The empirical power increases with the sample size and in some cases (strong instruments and large η) converges quickly to 1. The proposed test performs particularly well if the instruments used in experiment are strong. In general, the results are more than satisfactory given the fact that the model is only partially identified under the alternative hypothesis. A few testing procedures for partially identified models developed recently are typically complicated and allow to test a limited range of hypotheses. We provide the simple exogeneity test based on the standard results that can be applied in this conventionally untestable context.

Figures 7 and 8 show how the empirical power changes with the number of points of support of the instrument.

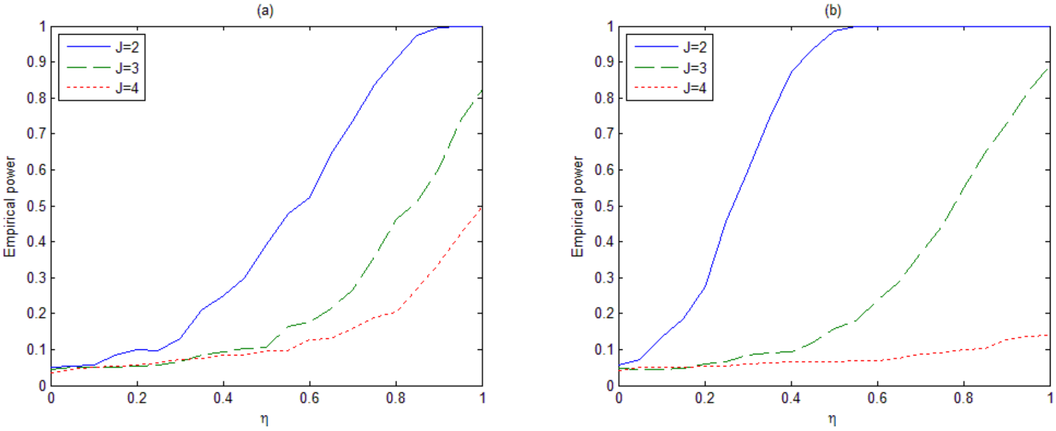

Figure 7: Empirical power for K=5 and n=400 with weak (a) and strong (b) instruments

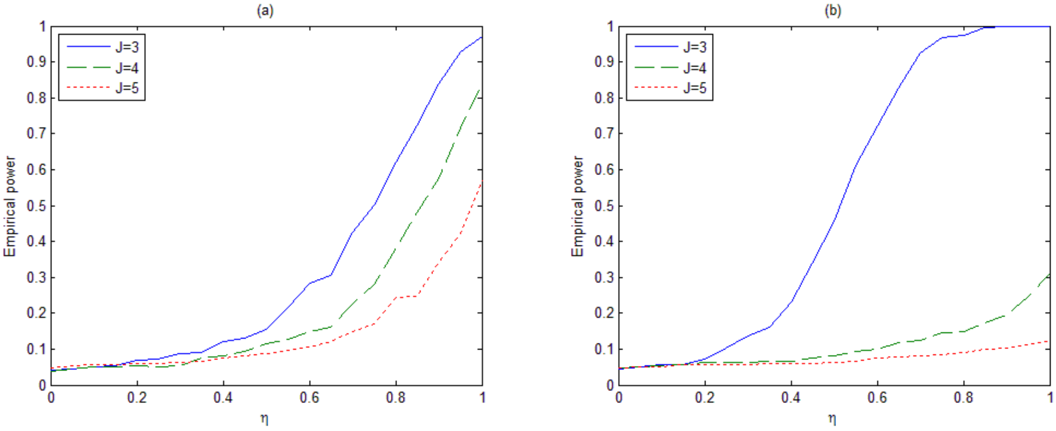

Figure 8: Empirical power for K=6 and n=400 with weak (a) and strong (b) instruments

For a fixed number of points of support of the regressor, the proposed test detects endogeneity of the regressor better when the support of the instrument is smaller. Hence, for both, weak and strong instruments, the power of the test is decreasing with the number of points of support in the Z. Therefore, in applications in order to obtain higher power in detecting endogeneity, among all the instruments available, the one with the smallest number of support points should be chosen. Note that if the gap between K and J is small, the test tends to be more powerful with weak instruments. This counter intuitive behaviour of the power function might be due to the fact that the chi-squared approximation is more accurate with smaller J. Simulations reveal that the approximation error is small up to 5 terms in the weighted sum. Therefore, in experiments with large support of instrumental variable, the critical values should be computed using another method discussed above.

6 Conclusion

The consistency of a standard nonparametric estimation procedures fails in the presence of endogeneity in the model. Therefore, in order to choose a consistent estimation technique, the applied researcher should test whether the explanatory variable(s) used in the model are exogenous. This paper has provided two consistent tests for exogeneity in nonparametric models, when the single explanatory variable is discrete. To the best of our knowledge, there exist no such tests for nonparametric models with discrete regressors. In models that point identify the unknown function of interest, the test is built on a quadratic form of a difference between two estimators, one of which is consistent only under exogeneity and the other is consistent under both scenarios. This testing framework follows closely the Wu-Hausman-type of test. It has been shown that under the null hypothesis of exogeneity, the test statistic follows chi-square distribution asymptotically and that the test is consistent against fixed alternatives.

In models that set identify the structure of interest, the test-statistic is based on a constrained minimized sum of squares. We have shown that under the null hypothesis, the proposed test-statistic converges to a weighted sum of chi-square (1) random variables. Under the alternative hypothesis the test-statistic converges to a weighted sum of noncentral chi-square (1) random variables. The proposed test is thus shown to be consistent with asymptotic power approaching 1 as the sample size increases.

The results of Monte Carlo simulations have shown satisfactory finite-sample properties of the proposed tests. Based on our experiment, we can conclude that:

- both tests have correct size even in small samples,
- empirical power increases with the sample size and converges to 1,
- using a strong instrument leads to better power properties,
- empirical power changes with the number of support points of both endogenous regressor and instrument.

Particularly interesting is the fact that the power increases with the gap between the number of points of support of the variables. Therefore, assuming that there is a choice between valid instruments for the applied researcher, when $J \geq K$, they should choose the one with the most points of support (when the instruments are weak), and when $J < K$ choose the one with the smallest number of support points in order to increase the probability of detecting endogeneity of the regressor.

Appendix: Proofs

The crucial result underlying the analysis is the asymptotic distribution of a vector $(\varepsilon' L_X \ \varepsilon' L_Z)'$. Therefore, we first prove Lemma 1 and use it to discuss other results.

Proof of Lemmas 1 and 2

Clearly $E[w_n] = 0$, since for each $j = 1, ...J$ $E[u_{nj}] = E_Z[\sum_i I(z_i^s = z_j)E[\varepsilon_i|z_i^s]] = 0$ and for each $k = 1, ...K$,

$$E[v_{nk}] = \sum_i E_X[I(x_i^s = x_k)E[\varepsilon_i|x_i^s]] = 0.$$

Therefore,

$$Var(u_{nj}) = E\left(\sum_{i=1}^n \varepsilon_i I(z_i^s = z_j)\right)^2 = n\sigma^2 q_j$$

and

$$Var(v_{nk}) = E\left(\sum_{i=1}^n \varepsilon_i I(x_i^s = x_k)\right)^2 = n\sigma^2 p_k.$$

The covariance between u_{nj} and v_{nk} is

$$\begin{aligned} cov(u_{nj}, v_{nk}) &= E\left(\sum_{i=1}^n \varepsilon_i I(z_i^s = z_j)\right)\left(\sum_{i=1}^n \varepsilon_i I(x_i^s = x_k)\right) \\ &= E\left(\sum_{i=1}^n \varepsilon_i^2 I(z_i^s = z_j) I(x_i^s = x_k)\right) \\ &= \sigma^2 E\left(\sum_{i=1}^n I(z_i^s = z_j) I(x_i^s = x_k)\right) \\ &= n\sigma^2 p_{jk}. \end{aligned}$$

The covariance between two different elements of u_n is zero, because for $j \neq l$

$$cov(u_{nj}, u_{nl}) = E\left(\sum_{i=1}^n \varepsilon_i^2 I(z_i^s = z_j) I(z_i^s = z_l)\right) = 0,$$

since $I(z_i^s = z_j)I(z_i^s = z_l) = 0$, and the indicated events cannot occur simultaneously. Similarly, for $k \neq s$

$$cov(v_{nk}, v_{ns}) = E\left(\sum_{i=1}^n \varepsilon_i^2 I(x_i^s = x_k) I(x_i^s = x_s)\right) = 0.$$

The covariance matrix of the vector $w_n = \begin{pmatrix} u_n \\ v_n \end{pmatrix}$ is therefore

$$nV = n\sigma^2 \begin{pmatrix} D_Z & P \\ P' & D_X \end{pmatrix}$$

with V finite. Because the components of w_n are correlated, we need a multivariate version of the Lindeberg-Feller central limit theorem to establish the asymptotic normality of $\frac{1}{\sqrt{n}} w_n$ (see, for example, van der Vaart (1998), Section 2.8). The stability condition (finite V) is clear, so to establish the result we need to confirm the Lindeberg condition

$$\frac{1}{n} E\left[\sum_{i=1}^{n} ||w_i||^2 I\{|w_i| > \sqrt{n}\delta\}\right] \to 0 \text{ for all } \delta > 0.$$

Firstly, observe that

$$\begin{aligned} ||w_i||^2 &= \sum_{k=1}^{K} \varepsilon_i^2 I(x_i^s = x_k) + \sum_{j=1}^{J} \varepsilon_i^2 I(z_i^s = z_j) \\ &= \varepsilon_i^2 \left(\sum_{k=1}^{K} I(x_i^s = x_k) + \sum_{j=1}^{J} I(z_i^s = z_j) \right) \\ &= 2\varepsilon_i^2 \end{aligned}$$

since $\sum_{k=1}^{K} I(x_i^s = x_k) = \sum_{j=1}^{J} I(z_i^s = z_j) = 1$. These results give

$$||w_i||^2 I\{|w_i| > \sqrt{n}\delta\} \leq ||w_i||^2 = 2\varepsilon_i^2,$$

with $E[2\varepsilon_i^2] = 2\sigma^2 < \infty$, and

$$\begin{aligned} \lim_{n\to\infty} ||w_i||^2 I\{|w_i| &> \sqrt{n}\delta\} = \lim_{n\to\infty} 2\varepsilon_i^2 I\{|\sqrt{2}\varepsilon_i| > \sqrt{n}\delta\} \\ &= \lim_{n\to\infty} 2\varepsilon_i^2 I\{2\varepsilon_i^2 > n\delta^2\} = 0. \end{aligned}$$

Therefore, by the dominated convergence theorem (see for example Severini, (2005), Theorem 1.10 (vi), p. 31), we have the Lindeberg condition:

$$\lim_{n\to\infty} E\left[||w_i||^2 I\{|w_i| > \sqrt{n}\delta\}\right] = 0.$$

Thus,

$$\frac{1}{\sqrt{n}} w_n \to^d N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{bmatrix} D_Z & P \\ P' & D_X \end{bmatrix} \right), \tag{17}$$

as claimed. Since z_n is a linear combination of u_n and v_n, by Slutsky's Theorem, Lemma 2 follows immediately.

Since Σ represents the covariance matrix, we have to show that it is positive definite. To do so, first observe that neither the support of Z, nor that of X, can affect the properties of w_n. That is to say, such properties must be *invariant to the support of* Z (or X), and hence hold for arbitrary support vectors z (or x). Now, the key matrix in Σ is $D_Z - PD_X^{-1}P' = D_Z - PD_X^{-1}D_X D_X^{-1}P'$. Let a denote a J-vector of hypothetical support points of Z, and consider the quadratic form in the matrix $D_Z - PD_X^{-1}D_X D_X^{-1}P'$:

$$a'D_Z a - \left(a'PD_X^{-1}\right) D_X \left(D_X^{-1}P'a\right). \tag{18}$$

The first term is $E_Z[Z^2] = E_X[E_{Z|X}[Z^2|X]]$- the second moment of Z when its support is a. The term $D_X^{-1}P'a$ is the vector of conditional means $E[Z|X = x_k]$, $k = 1, ..., K$, so the whole second term is $E_X[E_{Z|X}[Z|X]^2]$. Hence, the complete expression in (18) can be interpreted as

$$E_X\left[E_{Z|X}\left[Z^2 - E_{Z|X}[Z|X]^2\right]|X\right] = E_X[Var(Z|X)] > 0$$

i.e. the expectation of the conditional variance of Z given X when the support of Z is a. Since this must hold for all a, it follows that the matrix $D_Z - PD_X^{-1}P'$ is positive definite as required. The only exception would be if the conditional variance of Z given X vanished for each value of X, which we rule out.

Proof of Theorem 1

To determine the asymptotic distribution of the OLS estimator $\widehat{\beta}$, we need to study the asymptotic behaviour of $n^{-\frac{1}{2}}L'_X\varepsilon$, which could be derived by using standard Lindeberg CLT. However, in the proof of Lemma 1, we have already derived that the joint distribution of $L'_Z\varepsilon$ and $L'_X\varepsilon$. Given (17), we immediately get

$$n^{-\frac{1}{2}}L'_X\varepsilon \rightarrow^d N(0, \sigma^2 D_X).$$

It follows that, under exogeneity,

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) = \left(\frac{L'_X L_X}{n}\right)^{-1}\frac{L'_X\varepsilon}{\sqrt{n}} \rightarrow^d N\left(0, \sigma^2 D_X^{-1}\right).$$

Proof of Theorem 2

To determine the asymptotic distribution of the IV estimator $\widehat{\beta}_{IV}$, we need to study the asymptotic behaviour of $n^{-\frac{1}{2}}L'_Z\varepsilon$. Given (17), we have

$$n^{-\frac{1}{2}}L'_Z\varepsilon \rightarrow^d N(0, \sigma^2 D_Z).$$

Since

$$\widehat{\beta}_{IV} - \beta = \left(L'_X P_{L_Z} L_X\right)^{-1} L'_X P_{L_Z}\varepsilon = \left(L'_X L_Z (L'_Z L_Z)^{-1} L'_Z L_X\right)^{-1} L'_X L_Z (L'_Z L_Z)^{-1} L'_Z\varepsilon$$

it follows that

$$\begin{aligned}\sqrt{n}(\widehat{\beta}_{IV} - \beta) &= \left(\frac{L'_X L_Z}{n}\left(\frac{L'_Z L_Z}{n}\right)^{-1}\frac{L'_Z L_X}{n}\right)^{-1}\frac{L'_X L_Z}{n}\left(\frac{L'_Z L_Z}{n}\right)^{-1}\frac{L'_Z\varepsilon}{\sqrt{n}} \\ &\rightarrow {}^d N\left(0, \sigma^2\left(P'D_Z^{-1}P\right)^{-1}\right).\end{aligned}$$

Proof of Theorem 3

Note that under H_0 we have $E[y|X = L_X x] = L_X\beta$ and the standard arguments show easily that

$$n^{-1}y'M_{L_X}y \to^p \sigma^2.$$

The representation of T_n^* as a quadratic form in z_n follows from the fact that

$$\begin{aligned} W_{XZ} &= M_{L_X}L_Z[l_J, C_J][K^{-1}l_J, C_J]'\left(L_Z'L_Z\right)^{-1}L_Z'L_XC_K \\ &= [0, M_{L_X}L_ZC_J][K^{-1}l_J, C_J]'\left(L_Z'L_Z\right)^{-1}L_Z'L_XC_K \\ &= [M_{L_X}L_ZC_J][C_J'\left(L_Z'L_Z\right)^{-1}L_Z'L_XC_K]. \end{aligned}$$

Since the matrices C_J and C_K are non-random, we have

$$A_n = C_J'\left(\frac{L_Z'L_Z}{n}\right)^{-1}\frac{L_Z'L_X}{n}C_K \to^p C_J'D_Z^{-1}PC_K = A$$

and

$$\frac{1}{n}B_n \to^p A'\Sigma A.$$

By Lemma 2

$$\frac{1}{\sqrt{n}}A_n'z_n \to^d N(0, \sigma^2 A'\Sigma A)$$

It follows that

$$T_n^* \to^d \sigma^2\chi^2_{K-1}$$

and therefore

$$T_n = \frac{nT_n^*}{y'M_{L_X}y} \to^d \chi^2_{K-1}.$$

Proof of Theorem 4

Recall that we are interested in obtaining the asymptotic distribution of

$$\frac{1}{\sqrt{n}}\begin{bmatrix} L_Z'\varepsilon \\ L_X'\varepsilon \end{bmatrix} = \frac{1}{\sqrt{n}}\begin{bmatrix} u_n \\ v_n \end{bmatrix}$$

under the alternative hypothesis. Clearly $E[u_n] = 0$, since for each $j = 1, ...J$ $E[u_{nj}] = E_Z[\sum_i I(z_i^s = z_j)E[\varepsilon_i|z_i^s]] = 0$ by Assumption 2. For each $k = 1, ...K$, $E[v_{nk}] = \sum_i E_X[I(x_i^s = x_k)E[\varepsilon_i|x_i^s]] = n^{-\frac{1}{2}}p_k\sum_i m(x_k, v_i)$. Since we consider local alternatives of order $n^{-\frac{1}{2}}$, it follows that $E[v_n] = n^{-\frac{1}{2}}L_X'L_Xm$ and under the alternative

$$E\left(\frac{1}{\sqrt{n}}\begin{bmatrix} u_n \\ v_n \end{bmatrix}\right) \to \begin{bmatrix} 0 \\ D_Xm \end{bmatrix}$$

The covariance between u_{nj} and v_{nk} and the variances of u_{nj} and v_{nk} remain the same as under the null hypothesis. The stability condition in the Lindeberg-Feller CLT is satisfied and the Lindeberg condition is exactly the same as under H_0. Therefore under H_1,

$$\frac{1}{\sqrt{n}}\begin{bmatrix} u_n \\ v_n \end{bmatrix} \rightarrow^d N\left(\left(\begin{array}{c} 0 \\ D_X m \end{array}\right), \sigma^2\left(\begin{array}{cc} D_Z & P \\ P' & D_X \end{array}\right)\right).$$

Since we are interested in z_n, a linear function of u_n and v_n, it is clear that

$$\frac{1}{\sqrt{n}} z_n \rightarrow^d N(\mu, \sigma^2\Sigma)$$

with $\mu = -C_J' P m$ and $\Sigma = C_J'[D_Z - P D_X^{-1} P'] C_J$, and that

$$\frac{1}{\sqrt{n}} A_n' z_n \rightarrow^d N\left(A'\mu, A'\Sigma A\right)$$

with $A'\mu = -C_K' P' D_Z^{-1} C_J C_J' P m$. Note that the only difference in distribution between null and alternative hypotheses is a non-zero mean in the asymptotic distribution of z_n.

Therefore, the test statistic is a quadratic form in normal variables (with non-zero mean). This implies that

$$T_n = n\frac{T_n^*}{y' M_{L_X} y} \rightarrow^d \chi^2_{K-1}(\delta_L)$$

i.e. non-central chi-square distribution with non-centrality parameter

$$\delta_L = \frac{\mu' A \left(A'\Sigma A\right)^{-1} A'\mu}{\sigma^2}.$$

Proof of Theorem 5

Using z_n defined in (13), the numerator of the test-statistic can be written as

$$R_n^* = \left(n^{-\frac{1}{2}} z_n\right)' \left(\frac{C_J' L_Z' \left[P_{L_X} - P_{l_n}\right] L_Z C_J}{n}\right)^{-1} \left(n^{-\frac{1}{2}} z_n\right)$$

The matrix in the middle converges to $\Omega = C_J'\left(P D_X^{-1} P' - p_Z p_Z'\right) C_J$, where p_Z is a J-vector with elements q_j. Therefore,

$$R_n^* \rightarrow^d \sigma^2 z' \Omega^{-1} z$$

with $z \sim N(0, \Sigma)$. It follows that

$$R_n = n\frac{R_n^*}{y' M_{L_X} y} \rightarrow^d z'\Omega^{-1} z \sim \sum_{j=1}^{J-1} \omega_j \chi_j^2(1),$$

where the ω_j are the eigenvalues of $\Sigma\Omega^{-1}$, *i.e.*, those of $\Sigma^{\frac{1}{2}}\Omega^{-1}\Sigma^{\frac{1}{2}}$.

We have to show that all weights ω_j are positive. The argument is similar to that used to prove that Σ is positive definite, which we have already shown in the Proof of Lemma 1. Since the inverse of a positive definite matrix is positive definite itself, we only have to prove that Ω is positive definite. This will be so if

$$PD_X^{-1}P' - q_Z q_Z' = PD_X^{-1}D_X D_X^{-1}P' - q_Z q_Z'$$

where $q_Z = (q_1, ..., q_J)'$, is itself positive definite. Let a denote J- vector of the hypothetical support points of Z. Then the first term in

$$a'PD_X^{-1}D_X D_X^{-1}P'a - a'q_Z q_Z' a$$

is familiar (it also appears in the proof of positive definiteness of Σ) and equals

$$E_X \left[E_{Z|X}[Z|X]^2 \right].$$

The second term is simply $\left(E_Z\left[Z\right]\right)^2 = \left(E_X \left[E_{Z|X}[Z|X]\right]\right)^2$. Hence, the complete expression is

$$E_X \left(E_{Z|X}[Z|X]^2 \right) - \left(E_X \left[E_{Z|X}[Z|X] \right] \right)^2 = Var\left(E_{Z|X}[Z|X] \right) > 0,$$

the variance of the conditional expectation of Z given X. Since this must again be true for every support vector a. It follows that the matrix $PD_X^{-1}P' - q_Z q_Z'$ is positive definite as required. The only case, in which this term would be zero is when $E_{Z|X}[Z|X]$ is a constant i.e. the expectation of Z doesn't vary with X.

Proof of Theorem 6

Recall that for a quadratic form $T = Y'AY$ with A being a symmetric matrix and $Y \sim N_r(\mu, \Sigma)$ the standard results deliver $T \sim \sum_{i=1}^{r} \lambda_i \chi^2_{(1)}(\delta_i^2)$ with λ_i denoting the eigenvalues of ΣA and $(\delta_1, ..., \delta_r)' = S'L^{-1}\mu$ with L coming from the decomposition of $\Sigma = LL'$. S is an orthogonal matrix of the eigenvectors of $L'AL$.

Since $n^{-1}C_J' L_Z' (P_{L_X} - P_{l_n}) L_Z C_J \rightarrow^p \Omega$, it follows that,

$$R_n = n \frac{R_n^*}{y' M_{L_X} y} \rightarrow^d \sum_{j=1}^{J-1} \omega_j \chi_1^2(\delta_j^2)$$

with ω_j denoting the eigenvalues of $\Sigma\Omega^{-1}$ and non-centrality parameters

$$(\delta_1, ..., \delta_{J-1})' = S'\Sigma^{-\frac{1}{2}}\mu = -S'\Sigma^{-\frac{1}{2}} C_J' P m$$

References

[1] Anscombe, F.J. (1952) "Large sample-theory of sequential estimation" *Proceeding of Cambridge Philosophical Society*, 48, 600-607

[2] Bierens, H.J. (1990) "A consistent conditional moment test of functional form" *Econometrica*, 58, 1443-1458

[3] Blundell, R.W., Horowitz, J.L. (2007) "A nonparametric test of exogeneity" *Review of Economic Studies*, 74, 1035-1058

[4] Blundell, R.W., Horowitz, J.L., Parey, M. (2012) "Measuring the price responsiveness of gasoline demand: economic shape restrictions and nonparametric demand estimation" *Quantitative Economics*, 3, 29-51

[5] Buckley, M.J., Eagleson, G.K. (1988) "An approximation to the distribution of quadratic forms in normal random variables" *Australian Journal of Statistics*, 30A, 150-159

[6] Chesher, A. (2004) "Identification in additive error models with discrete endogenous variables" CeMMap working paper, CWP11/04

[7] Das, M. (2005) "Instrumental variables estimators of nonparametric models with discrete endogenous regressors" *Journal of Econometrics*, 124, 335-361

[8] Deaton, A. (2010) "Instruments, randomization and learning about development" *Journal of Economic Literature*, 48, 424-455

[9] Fan, Y., Li, Q. (1996) "Consistent model specification tests: omitted variables, parametric and semiparametric functional forms" *Econometrica*, 64, 865-890

[10] Florens, J.P., Malavolti L. (2003) "Instrumental regression with discrete endogenous variables" Working paper, GREMAQ, Universite des Sciences Sociales, Toulouse

[11] Greene, W.H. (1993) "Econometric Analysis", Second Edition, Macmillan Publishing Company, New York

[12] Hall, P. (1983) "Chi squared approximations to the distribution of a sum of independent random variables" *The Annals of Probability*, 11, 1028-1036

[13] Hall, P., Huang, L.S. (2001) "Nonparametric kernel regression subject to monotonicity constraints" *Annals of Statistics*, 29, 624-647

[14] Hausman, J.A. (1978) "Specification tests in econometrics" *Econometrica*, 46, 1251-1272

[15] Hu, Y., Lewbel, A. (2008) "Idenifying the returns to lying when the truth is unobserved" CeMMap working paper, CWP06/08

[16] Imhof, J.P. (1961) "Computing the distribution of quadratic forms in normal variables" *Biometrika*, 48, 419-426

[17] Iori, G., Kapar, B., Olmo, J. (2014) "Bank characteristics and the interbank money market: a distributional approach" accepted in *Studies in Nonlinear Dynamics and Econometrics*

[18] Johnson, N.L., Kotz, S., Balakrishnan, N. (1994) "Continuous univariate distributions. Volume 1", Second Edition, John Wiley & Sons Inc., New Jersey

[19] Lavergne, P., Patilea V. (2008) "Breaking the curse of dimensionality in nonparametric testing" *Journal of Econometrics*, 143, 103-122

[20] Lavergne, P., Vuong, Q. (2000) "Nonparametric significance testing" *Econometric Theory*, 16, 576-601

[21] Matzkin, R.L. (2007) "Nonparametric identification" in Handbook of Econometrics, Volume 6, Part B, 5307-5368

[22] Muirhead, R.J. (1982) "Aspects of multivariate statistical theory" John Wiley & Sons Inc., Hoboken, New Jersey

[23] Newey, W.K. (1985) "Maximum likelihood specification testing and conditional moment tests" *Econometrica*, 53, 1047-1070

[24] Newey, W.K., Powell, J.L. (2003) ”Instrumental variable estimation of nonparametric models” *Econometrica,* 71, 1565-1578

[25] Robbins, H. (1948) "The asymptotic distribution of the sum of a random number of random variables" *Bulletin of the American Mathematical Society,* 54, 1151-1161

[26] Severini, T.A. (2005) "Elements of distribution theory" Cambridge University Press

[27] Sheil, J., Muircheartaigh, I. (1977) "Algorithm AS106: The distribution of non-negative quadratic forms in normal variables" *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26, 92-98

[28] Tamer, E. (2010) "Partial identification in econometrics" *Annual Review of Economics*, 2, 167-195

[29] van der Vaart, A.W. (1998) "Asymptotic statistics" Cambridge University Press

[30] Zhang, J.T. (2005) "Approximate and asymptotic distributions of chi-squared-type mixtures with applications" *Journal of the Americal Statistical Association*, Vol. 100, No. 469, 273-285

[31] Zheng, J.X. (1996) "A consistent test of functional form via nonparametric estimation techniques" *Journal of Econometrics*, 75, 263-289