Statistical Papers
https://doi.org/10.1007/s00362-019-01116-9

REGULAR ARTICLE

Heterogeneous endogeneity

Pallab Ghosh[1] · Kevin Grier[2] · Jaeho Kim[1]

Received: 25 June 2018 / Revised: 21 May 2019

Abstract

We define heterogeneous endogeneity as the case where a potentially endogenous regressor is endogenous for some sub-groups of the data but exogenous for other subgroups. We derive an estimator and test procedure based on the control function approach to deal with the phenomenon. We show that accounting for heterogeneous endogeneity can greatly increase the power of endogeneity tests and increase the precision of our estimator over traditional IV. While the gains get larger as the instrument gets weaker and as the relative size of the non-endogenous subgroup gets larger, we find efficiency gains even when the underlying instrument is very strong. We illustrate our approach with an example using data from Abaide et al. (Econometrica 70:91–117).

Keywords Heterogeneous Endogeneity · Instrumental variables · Heterogeneity · Control function

Mathematics Subject Classification 62J05 · 62F12

1 Introduction

Instrumental variables (IV) estimation has been a workhorse method of obtaining causal estimates in observational data. In this paper, we study refinements for standard IV methods to achieve a considerable reduction in estimation uncertainty when for some subgroups in the sample, the endogenous regressor is not correlated with the

✉ Jaeho Kim
jaeho@ou.edu

Pallab Ghosh
pallab.ghosh@ou.edu

Kevin Grier
kevin.grier@ttu.edu

1 Department of Economics, University of Oklahoma, 428 Cate Center Drive, Norman, OK 73072, USA

2 Department of Political Science, Texas Tech University, Lubbock, TX 79409, USA

Published online: 06 June 2019

error term and thus no instrument is needed. We call this phenomenon heterogeneous endogeneity.

The main conclusion of our theoretical and empirical analyses is that when there exists heterogeneous endogeneity in the population, applying the standard IV methods result in inefficient estimation because even valid instruments lose relevant information on causal relations in such a case. The intuition behind our main result is straightforward. When the instrument is not very strong, identifying subgroups in the sample where the endogenous regressor is actually exogenous can greatly reduce the variance of our estimator relative to traditional IV, as we do not have to lose any relevant information in the variable that is not perfectly correlated with the instrument. The less strong the instrument and the larger the size of the exogenous subgroup, the more benefit there is to accounting for heterogeneous endogeneity when it exists.

In many areas of economics, the notion of heterogeneous endogeneity is quite plausible. For example, investment rates are frequently thought to be an endogenous regressor in growth equations. However, in planned as opposed to market economies, investment could be plausibly exogenous. Similarly, we generally think of labor market participation as an endogenous regressor, but for some gender or ethnic groups, labor supply could also plausibly exogenous. The heterogeneous endogeneity could occur over time as well as across groups. It is plausible that monetary policy variables were not endogenous regressors during the gold standard or even during the Bretton Woods era.

A more specific example of potential heterogeneous endogeneity is the study by Autor et al. (2013) who show that from 1990 to 2007 the rising Chinese import competition decreases manufacturing employment and wages in the US. In a more recent study, Autor et al. (2017) show that specific geographic regions were more affected by the Chinese import competition because historically those areas had a relatively larger manufacturing employment shares. Since the composition of labor force in some states was not much affected because of the Chinese import competition, we can treat those most unaffected states as an exogenous group. Another example of potential heterogenous endogeneity can be in the study by Hashimoto (1987) in which the author estimates the impact of federal minimum wage on the local crime rate of youth.[1] Since federal minimum wage is not a binding constraint for some states which had a higher state minimum wage than the new federal minimum wage, we can also treat those states who are least affected because of a change in federal minimum wage as an exogenous group.

We propose a new estimation approach for heterogeneous endogeneity based a control function approach, popularized by Smith and Blundell (1986) and Rivers and Vuong (1988). In Monte Carlo simulations we conduct, the proposed control function estimator often produces efficiency gains of 50% and above compared to a traditional IV method. We also show that accounting for heterogeneous endogeneity, when it is present, can greatly increase the power of standard endogeneity tests. The weaker is the instrument and the larger is the exogenous subgroups, the more likely a standard

[1] One can argue that the federal minimum wage is an endogenous regressor because it is most likely correlated with some states' unobserved labor market conditions especially for those states which have relatively larger shares on the national income.

test accounting for heterogeneous endogeneity can correctly reject the null of no endogeneity.[2]

Many previous studies employ the control function approach for various issues. In a linear regression framework, the control function approach is essentially an instrumental variable method.[3] Although Goldberger (1972) did not use the phrase 'Control Function', notably this is the first study which uses the concept of control function to address self-selection into program participation. Subsequently, Barnow et al. (1981) and Heckman and Robb (1985) implement this approach in the context of program evaluation for longitudinal data and since then the control function approach has been widely used in applied research in the last three decades. Imbens and Wooldridge (2009) and Wooldridge (2015) provide an overview of how the control function approach has been implemented in various linear and non-linear models to resolve estimation problems regarding selection and endogeneity bias.

Lee (2007) extends the Koenker and Bassett (1978) quantile regression method to an endogenous quantile regression model by using a control function approach. In addition to the linear and non-linear additive separable models, the control function approach can be also used to correct endogeneity issue for the nonseparable models and the topic has been an active research area since the late 1990s. By extending the Newey et al. (1999) control function approach, Newey et al. (2003) show the identification and estimation of nonseparable, nonparametric triangular simultaneous equation models. Chesher (2003) and Chesher (2005) develop conditional quantile restrictions to identify nonseparable models and Ma and Koenker (2006) use those identification results to a nonseparable parametric model and also develop a control function method for the parametric model.[4] Our estimator can be thought of as an extended control function approach for heterogeneous endogeneity.

We extend our baseline model specification into two important directions. The first case is when clear classification for exogenous and endogenous groups is not possible because there exists no relevant economic theory or no information about the group membership is available prior to estimation. To address the issue, we employ k-mean clustering following the idea of Bonhomme and Manresa (2015). This approach extracts signals about group membership from the data without explicitly modeling the group structure and uses the signals to identify exogenous and endogenous groups. The finite sample performance of the resulting estimator is provided in *Appendix B*. The first model extension will be useful for panel data with large or moderately large T. Such cases can be easily found in country-level macro or international finance panel data sets.

Secondly, we incorporate heterogeneity in the coefficients of endogenous variables by using the correlated random coefficient (CRC) model proposed by Heckman and

[2] Though the strength of the instrument is a factor we vary in our simulations, none of the cases we consider suffer from a weak instrument problem in the sense of Nelson and Startz (1990b) or Stock et al. (2002).

[3] In Sect. 2, we explain the control function method in detail.

[4] Pinkse (2000), Blundell and Powell (2003), Darolles et al. (2003), Hall and Horowitz (2005), Vytlacil and Yildiz (2007), Blundell et al. (2007), Florens et al. (2008), and Imbens and Newey (2009) among others have extended the literature.

Vytlacil (1998) and Wooldridge (2015).[5] This extension is important in that a model that only considers homogeneous effects can produce inconsistent estimates when heterogeneity in a coefficient of an endogenous variable is correlated with the same endogenous variable in the population. If we fail to properly handle the issue, the correlated heterogeneity is absorbed in the error term of a model and creates another endogeneity issue. We apply different extended models in our empirical application about a government job training program for robust checks.

In the rest of this paper we formally present our notation of heterogeneous endogeneity, derive its asymptotic properties, present a test for the presence of heterogeneous endogeneity, show the efficiency gains from modeling heterogeneous endogeneity when it is present through a series of simulation exercises and apply it to data on the effects of job training on earnings in an example.

2 Heterogeneous endogeneity

2.1 Model specification

In many empirical studies, individual observations are often categorized into several groups according to their qualitative or quantitative characteristics. We therefore introduce a linear regression model which allows for 'heterogenous endogeneity' across different groups:

$$y_{ij} = x_{ij}'\beta + u_{ij}, \tag{1}$$

$$x_{ij} = Z_{ij}'\pi + \nu_{ij} \tag{2}$$

$$\begin{bmatrix} \nu_{ij}^* \\ u_{ij} \end{bmatrix} \sim i.i.d \left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} I_K & \rho_j \sigma_u \\ \rho_j' \sigma_u & \sigma_u^2 \end{bmatrix} \right) \tag{3}$$

$$for\ i = 1, 2, \ldots, N_j,\ and\ \ j = 1, 2, \ldots, J$$

where y_{ij} is a 1×1 dependent variable; x_{ij} is a $K \times 1$ vector of endogenous variables; z_{ij} is an $L \times 1$ vector of instrumental variables,[6] $\nu_{ij}^* = \Sigma_\nu^{-1} \nu_{ij}$, $\Sigma_\nu = Var(\nu_{ij})$,

[5] There is a large literature on models where the heterogeneous effects of endogenous variables on an outcome variable exist in the population. In the literature, many important works such as Abadie (1991) and Angrist et al. (1996) have proposed methods for estimating the average effect of receiving or not receiving a binary treatment. Recently, Florens et al. (2008) provided a non-parametric way to analyze heterogeneous effects of a continuous treatment. For heterogeneous response models, see Heckman et al. (1997), Card (1999) , Card (2001), Heckman and Vytlacil (2005), Heckman and Vytlacil (2007a), Heckman and Vytlacil (2007b) and references therein.

[6] For notational simplicity, we do not include a vector of exogenous covariates in Eq. (1). We assume that we have a set of valid instruments z_{ij} in Eq. (2) and thus the parameter β in Eq. (1) can be consistently estimated by 2SLS. This assumption also implies that our instrument is exogenous to all subgroups of population. Some specific examples of such instruments can be found in Graddy (1995) and Acemoglu et al. (2013).

and $Z_{ij} = (I_K \otimes z_{ij})$; $\rho_j = [\rho_{1j}, \ \rho_{2j}, \ \dots \ , \rho_{Kj}]'$ is a $K \times 1$ vector of correlation coefficients. The subscript ij denotes the i-th individual observation in the j-th group. We take the standard assumptions that all the elements of z_{ij} are not perfectly collinear and the rank condition hold for identification. The above model incorporates the heterogeneous degrees of association between x_{ij} and u_{ij} for J groups:

$$Corr(x_{ij}, u_{ij}|z_{ij}) = Corr(\nu_{ij}, u_{ij}|z_{ij}) = \rho_j \neq \rho. \text{ for } j = 1, 2, \dots, J.$$

We define this important feature of the population as 'heterogeneous endogeneity'.

In this section, we consider a simple linear model without any included exogenous variables and with predetermined group assignments[7] to better understand how heterogeneous endogeneity affects conventional methods and how our proposed method works in familiar settings. However, many important implications suggested by this paper are easily generalized to fairly complicated linear and non-linear models.

2.2 Asymptotic properties of a proposed estimator under heterogeneous endogeneity

If ρ_j is significantly different from zero for one group while ρ_j is zero for another group, the OLS estimator is inconsistent due to the former group's degree of endogeneity. Conversely, the 2SLS estimator with a set of valid instruments would be consistent but the asymptotic variance would be higher than necessary as we lose a significant amount of information in the latter group. This type of potential heterogeneous endogeneity and associated econometric issues have been largely ignored in literature. As shown in Sect. 2.4, the power of a test for endogeneity and the sampling distributions of applied estimators are significantly affected by the presence of heterogeneous endogeneity.

In this section, we introduce an estimator based on a Control Function (CF) approach and study its asymptotic properties under the existence of heterogeneous endogeneity. In particular, we show that the proposed CF method provides an asymptotically efficient estimator within the class of IV estimators under heterogeneous endogeneity.

To implement a CF approach, we can write the linear projection of u_{ij} on ν_{ij}:

$$u_{ij} = \nu_{ij}'\gamma_j + \eta_{ij}, \quad \eta_{ij} \sim (0, \sigma_{\eta_j}^2) \tag{4}$$

where $\gamma_j = \left[E(\nu_{ij}'\nu_{ij})\right]^{-1} E[\nu_{ij} u_{ij}]$ is the population regression coefficients. In a similar fashion, a group-wise heteroskedasticity in η_{ij} caused by heterogeneous endogeneity should be considered without the normality assumption. Substituting Eqs. (4) into (1) for group j gives

$$y_{ij} = x_{ij}'\beta + \nu_{ij}'\gamma_j + \eta_{ij}, \quad \eta_{ij} \sim (0, \sigma_{\eta_j}^2). \tag{5}$$

[7] The group assignments in terms of heterogeneous endogeneity would be viable if economic theories suggest that observed individual characteristics decide the group structure.

The conditional expectation of y_{ij} given x_{ij} and z_{ij} is $E[y_{ij}|x_{ij}, z_{ij}] = E[y_{ij}|x_{ij}, \nu_{ij}] = x'_{ij}\beta + \nu'_{ij}\gamma_j$ where $\nu_{ij} = (x_{ij} - Z'_{ij}\pi)$. The error term η_{ij} is not correlated with both of x_{ij} and ν_{ij} and therefore, we can consistently estimate β by including a control function ν_{ij} in the regression model. While ν_{ij} is not directly observed, the error term can be consistently estimated by running OLS. By replacing ν_{ij} in Eq. (5) with OLS residuals $\hat{\nu}_{ij}$, we obtain the following reduced form equation:

$$y_{ij} = x'_{ij}\beta + \hat{\nu}'_{ij}\gamma_j + \epsilon_{ij} \tag{6}$$

where $\epsilon_{ij} = (\nu_{ij} - \hat{\nu}_{ij})'\gamma_j + \eta_{ij}$. The OLS estimation of Eq. (6) is known as a control function approach. It is worth mentioning that the CF approach produces numerically identical estimates to 2SLS estimates when both methods ignore heterogeneous endogeneity.

Using matrix notation, we rewrite Eq. (6) as:

$$Y = X\beta + \widehat{V}\Gamma + E \tag{7}$$

where the $N \times K$ matrix X contains the regressors x_{ij} and $N = \sum_{j=1}^{J} N_j$; N_j is the total number of observations in the jth group; $\hat{V} = diag[\hat{\nu}_1, \hat{\nu}_2, \hat{\nu}_3, \cdots \hat{\nu}_J]$ is the $N \times KJ$ control function matrix; $\hat{\nu}_j = [\hat{\nu}_{1j}\ \hat{\nu}_{2j}\ \ldots\ \hat{\nu}_{N_jj}]'$ is a $N_j \times K$ error matrix of j-th group; the $KJ \times 1$ vector Γ contains the coefficients of the control function; $E = [\epsilon'_1\, \epsilon'_2\, \epsilon'_3\, \ldots\, \epsilon'_J]'$ where $\epsilon_j = [\epsilon_{1j}\, \epsilon_{2j}\, \ldots\, \epsilon_{N_jj}]'$. Based on Frisch-Waugh-Lovell (FWL) theorem, we get the CF estimator for β:

$$\widehat{\beta}_{CF} = \left(X'M_{\hat{V}}X\right)^{-1}\left(X'M_{\hat{V}}Y\right) \tag{8}$$

where $M_{\hat{V}} = I_N - \hat{V}(\hat{V}'\hat{V})^{-1}\hat{V}'$. The asymptotic distribution of the CF estimator is given in the following Lemma 1.

Lemma 1 *Suppose that Assumptions 1 to 3 as shown in Appendix hold. Then,*

$$\sqrt{N}\left(\widehat{\beta}_{CF} - \beta\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma_u^2\left(\sum_{j=1}^{J} a_j Q'_{Z_jX_j} Q^{-1}_{Z_jZ_j} Q_{Z_jX_j}\right)^{-1}\right)$$

where $a_j = \lim_{N\to\infty}\left(\frac{N_j}{N}\right)$ *and* $\sum_{j=1}^{J} a_j = 1.$

Note that the asymptotic variance of the estimator is a function of a weighted average of the population moments of different groups. If the population moments of different groups are same, the asymptotic distribution of the estimator $\hat{\beta}_{CF}$ becomes equivalent to that of the conventional 2SLS estimator. That is,

$$\sqrt{N}\left(\hat{\beta}_{CF} - \beta\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma_u^2\left(Q'_{ZX}Q^{-1}_{ZZ}Q_{ZX}\right)^{-1}\right)$$

if $Q'_{Z_jX_j}Q^{-1}_{Z_jZ_j}Q_{Z_jX_j} = Q'_{ZX}Q^{-1}_{ZZ}Q_{ZX}$ for $j = 1, 2, \ldots, J$. However, even when a 2SLS estimation is asymptotically equivalent to the CF approach, the CF method has a very attractive feature for endogeneity tests as we will describe in Sect. 2.4.

Suppose that m out of J groups are not subject to endogeneity with $\rho_j = 0$ for $j = j_1, \ldots, j_m$. Then, we have:

$$Corr(x_{ij}, u_{ij}|z_{ij}) = Corr(\nu_{ij}u_{ij}|z_{ij}) = 0, \;\; for \;\; j = j_1, \ldots, j_m.$$

Given the m exogenous groups, we rewrite Eq. (7) in the following block matrix

$$\begin{bmatrix} y_e \\ y_{ne} \end{bmatrix} = \begin{bmatrix} x_e \\ x_{ne} \end{bmatrix}\beta + \begin{bmatrix} \hat{\nu}_e & 0 \\ 0 & \hat{\nu}_{ne} \end{bmatrix}\begin{bmatrix} \gamma_e \\ \gamma_{ne} \end{bmatrix} + \begin{bmatrix} \epsilon_e \\ \epsilon_{ne} \end{bmatrix}$$

where the subscript 'e' represents the $J - m$ endogenous groups with $\rho_j \neq 0$ and the subscript 'ne' represents the remaining m exogenous groups with $\rho_j = 0$; the matrix of the control functions of the m groups is denoted by $\hat{\nu}_{ne} = diag[\hat{\nu}_{j_1} \; \hat{\nu}_{j_2} \; \ldots \; \hat{\nu}_{j_m}]$ and $\hat{\nu}_e$ is the matrix of the control functions for the remaining $J - m$ groups. Since $\gamma_{ne} = 0$, we get the following equivalent linear equation:

$$\begin{bmatrix} y_e \\ y_{ne} \end{bmatrix} = \begin{bmatrix} x_e \\ x_{ne} \end{bmatrix}\beta + \begin{bmatrix} \hat{\nu}_e \\ 0 \end{bmatrix}\gamma_e + \begin{bmatrix} \epsilon_e \\ \epsilon_{ne} \end{bmatrix}$$
$$(Y = X\beta + \tilde{V}\Gamma_e + E) \tag{9}$$

where $\tilde{V} = \begin{bmatrix} \hat{\nu}_e \\ 0 \end{bmatrix}$, $\Gamma_e = \gamma_e$, and $E = \begin{bmatrix} \epsilon_e \\ \epsilon_{ne} \end{bmatrix}$. By using the FWL theorem, we get

$$\widehat{\beta}^*_{CF} = \left(X'M_{\tilde{V}}X\right)^{-1}\left(X'M_{\tilde{V}}Y\right) \tag{10}$$

where $M_{\tilde{V}} = I_N - \tilde{V}(\tilde{V}'\tilde{V})^{-1}\tilde{V}'$. Note that $\widehat{\beta}^*_{CF}$ is a special case of $\widehat{\beta}_{CF}$ because $\tilde{V}$ is a special form of the matrix $\widehat{V}$. Hence, we get the following result.

Proposition 1 *Suppose that $\rho_j = 0$ for $j = j_1, j_2, \ldots, j_m$. Then, the control function estimator $\widehat{\beta}^*_{CF}$ can be written as*

$$\widehat{\beta}^*_{CF} = W\widehat{\beta}_{e,2SLS} + (I - W)\widehat{\beta}_{ne,OLS}$$

where $\widehat{\beta}_{e,2SLS} = \left(x'_eM_{\hat{\nu}_e}x_e\right)^{-1}\left(x'_eM_{\hat{\nu}_e}y_e\right)$; $M_{\hat{\nu}_e} = I_{N_e} - \hat{\nu}_e(\hat{\nu}'_e\hat{\nu}_e)^{-1}\hat{\nu}'_e$ and N_e is the total number of observations in endogenous groups. $\widehat{\beta}_{e,2SLS}$ is the 2SLS estimator for $J - m$ endogenous groups and $\widehat{\beta}_{ne,OLS} = \left(x'_{ne}x_{ne}\right)^{-1}x'_{ne}y_{ne}$ is the OLS estimator for m exogenous groups, respectively. The weight matrix W is given by

$$\begin{aligned} W &= \left(x'_{ne}x_{ne} + x'_eM_{\hat{\nu}_e}x_e\right)^{-1}(x'_eM_{\hat{\nu}_e}x_e) \\ &= \left(\widehat{Avar}\left[\widehat{\beta}_{e,2SLS}\right]^{-1} + \widehat{Avar}\left[\widehat{\beta}_{ne,OLS}\right]^{-1}\right)^{-1}\left(\widehat{Avar}\left[\widehat{\beta}_{e,2SLS}\right]^{-1}\right). \end{aligned}$$

The proof of Proposition 1 is given in the Appendix. It is important to recognize that the weighting matrix W is a function of the asymptotic covariance matrices of $\beta_{e,2SLS}$ and $\beta_{ne,OLS}$.[8] Proposition 1 is intuitively very appealing in that the CF estimator imposes relatively a larger (smaller) weight matrix to an estimator with a smaller (larger) covariance matrix. The following Lemma 2 provides the asymptotic distribution of the CF estimator when some groups do not suffer from endogeneity.

Lemma 2 *Suppose that $\rho_j = 0$ for $j = j_1, j_2, \ldots, j_m$ and the standard Assumptions 1 and 2 hold. Then, the control function estimator $\widehat{\beta}^*_{CF}$ has the following asymptotic distribution given instrumental variables Z:*

$$\sqrt{N}\left(\widehat{\beta}^*_{CF} - \beta\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma_u^2 \left(w\ Q'_{Z_e X_e} Q^{-1}_{Z_e Z_e} Q_{Z_e X_e} + (1-w)\ Q^{-1}_{X_{ne} X_{ne}}\right)^{-1}\right)$$

where X_e and Z_e present the endogenous variable and the instrumental variable matrices for endogenous groups; X_{ne} presents the exogenous variable matrix for exogenous groups and $w = \lim_{N\to\infty}\left(\frac{N_e}{N}\right)$.

We provide the Proof of Lemma 2 in the Appendix. From Lemma 2, we can show that

$$Avar\left(\hat{\beta}^*_{CF}\right) = \left(Avar\left(\beta_{e,2sls}\right)^{-1} + Avar\left(\beta_{ne,ols}\right)^{-1}\right)^{-1}$$

If all the groups are subject to endogeneity, $\hat{\beta}^*_{CF}$ becomes $\hat{\beta}_{2SLS}$ as $w = 1$. On the other hand, if all the groups are free from endogeneity, $\hat{\beta}^*_{CF}$ becomes a simple $\hat{\beta}_{OLS}$ with $w = 0$. Therefore, our CF estimator encompasses the 2SLS and OLS estimators in those special cases. For a general case where we have both endogenous and exogenous groups together, Proposition 2 shows that the CF estimator is asymptotically more efficient than the other two conventional estimators.

Proposition 2 *Suppose that $\rho_j = 0$ for $j = j_1, j_2, \ldots, j_m$ and the Assumption 1 and 2 hold in Appendix. Then, the control function estimator $\widehat{\beta}^*_{CF}$ is asymptotically efficient within the class of instrumental variables estimators given instrumental variables Z.*

We prove Proposition 2 by comparing the inverse of the asymptotic variances of the CF estimator and the conventional 2SLS estimator. More details of the proof are provided in the Appendix.

As for the estimator of σ_u^2, we calculate the residual $\hat{\epsilon}_{ij}$ by running the OLS regression of $y_{i,j}$ on x_{ij} and $\hat{\nu}_{ij}$ in Eq. (9). Since the variances of ϵ_{ij} are different across J groups due to heterogeneous endogeneity, we have a consistent estimator of σ_u^2:

$$\hat{\sigma}_u^2 = \frac{1}{N}\hat{U}'\hat{U} \tag{11}$$

[8] We note that by definition W is a $K \times K$ matrix. In a special case, when we have only one endogenous variable, W is a scalar and $0 \leq W \leq 1$.

where $\hat{U}$ is

$$\hat{U} = \begin{bmatrix} \hat{\epsilon}_e + \hat{\nu}_e \hat{\gamma}_e \\ \hat{\epsilon}_{ne} \end{bmatrix}$$

and $\hat{\epsilon}_e$ is the vector of residual $\hat{\epsilon}_{ij}$ for endogenous groups and $\hat{\gamma}_e$ is the estimated γ_e by OLS. Because $\hat{\beta}^*_{CF}$ is a consistent estimator, the estimator in Eq. (11) also provides a consistent estimator of σ_u^2. Thus, the covariance matrix of $\hat{\beta}^*_{CF}$ is estimated by

$$\widehat{Avar}\left(\widehat{\beta}^*_{CF}\right) = \hat{\sigma}_u^2 \left(N_e \left(x_e' P_{Z_e} x_e\right)^{-1} + N_{ne}\left(x_{ne}' x_{ne}\right)^{-1}\right)^{-1} \tag{12}$$

where N_e and N_{ne} are the numbers of observations in endogenous and exogenous groups, respectively.

Under the homoskedasticity setup specified in Eq. (3) and Assumption 1, we have derived the asymptotic covariance matrix of $\hat{\beta}^*_{CF}$. The results provided in this section can be carried over to a case with a more general form of heteroskedasticity by using the residuals $\hat{u}_{ij}$ obtained by the consistent estimators. For instance, we can easily handle the group-wise heteroskedasticity in u_{ij} by estimating $\hat{\sigma}_j^2$ and pre-multiplying its inverse to the corresponding dependent and explanatory variables in Eq. (3). Then, the asymptotic results presented here can be applied to the observations normalized by $\hat{\sigma}_j^2$.

2.3 Inference under heterogeneous endogeneity

In Sect. 2.2, we have investigated the asymptotic properties of the proposed control function estimator for β when there exists heterogeneous endogeneity. In addition to the asymptotic properties, we also need to study the asymptotic distributions of $\hat{\gamma}_j$ for $j = 1, 2, \ldots, J$. Therefore, in this section we provide further asymptotic results for the estimators of γ_j and also propose a proper test for heterogeneous endogeneity. For this purpose, we first rewrite the model in Eqs. (1) and (2) as:

$$y_n = x_n^{*'} \delta + \eta_n, \tag{13}$$
$$x_n = Z_n' \pi + \nu_n \tag{14}$$

where $x_n^* = [x_n' \quad V_n']'$; $\nu_n = \Sigma_\nu \nu_n^*$; $V_n = (D_n \otimes \nu_n)$; $\delta = [\beta' \quad \Gamma']'$; $\Gamma = [\gamma_1' \quad \gamma_2' \quad \ldots \quad \gamma_J']'$; $n = \sum_{g=0}^{j-1} N_g + i$ and $N_0 = 0$ for $i = 1, 2, \ldots, N_j$ and $j = 1, 2, \ldots, J$. D_n is a $(J \times 1)$ row vector, whose j-th element takes one if the n-th individual is in group j and all remaining elements are zeros. From Eq. (14), we obtain the control function, $\hat{V}_n = x_n - Z_n'\hat{\pi}$ where $\hat{\pi}$ is the OLS estimator of π. Using $\hat{\pi}$, we define $\hat{x}_n^* = [x_n' \ \hat{V}_n']'$ and obtain the following OLS estimator of δ:

$$\widehat{\delta} = \left(\sum_{n=1}^{N} \widehat{x}_n^* \widehat{x}_n^{*'}\right)^{-1} \left(\sum_{n=1}^{N} \widehat{x}_n^* y_n\right). \tag{15}$$

Before investigating the asymptotic distribution of $\hat{\delta}$, in the following Lemma 3 we derive the consistency property of $\hat{\delta}$.

Lemma 3 *Suppose that the standard Assumptions 1 and 2 (stated in the appendix) hold. Then the OLS estimator $\hat{\delta}$ is a consistent estimator for δ.*

Because $\hat{\delta} = [\hat{\beta}' \ \hat{\Gamma}']'$, Lemma 3 also implies that $\hat{\Gamma}$ is a consistent estimator of Γ. Therefore, the above result suggests that even if we do not observe V_n, we can still consistently estimate the coefficient of V_n in Eq. (13) as long as we have a set of valid instruments. We use this consistency property of $\hat{\delta}$ to derive its asymptotic distribution in the following Proposition 3.

Proposition 3 *Suppose that the standard Assumptions 1 and 2 (stated in the appendix) hold. Then the asymptotic distribution of the OLS estimator $\hat{\delta}$ is given by:*

$$\sqrt{N}\left(\hat{\delta} - \delta\right) \xrightarrow{d} N\left(\mathbf{0}, C^{-1}(H\ M\ H' + Var[x_n^* \eta_n])(C^{-1})'\right),$$

where $C = E[x_n^* x_n^{*'}]$, $H = E\left[x_n^*\left(\Gamma' \times \mathbf{Z_n}\right)\right]$, $\mathbf{Z_n} = (D_n \otimes Z_n')$, $\sqrt{N}(\hat{\pi} - \pi) \xrightarrow{d} N(0, M)$.

From Proposition 3, we obtain the following estimate for the asymptotic variance of $\hat{\delta}$,

$$\widehat{Avar}(\hat{\delta}) = \frac{1}{N}\hat{C}^{-1}\left(\hat{H}\ \hat{M}\ \hat{H}' + \widehat{Var}[x_n^* \eta_n]\right)(\hat{C}^{-1})'$$

where $\hat{C} = \frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^* \hat{x}_n^{*'}$, $\hat{H} = \frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^*(\hat{\Gamma}'\mathbf{Z_n})$ and $\widehat{Var}[x_n^* \eta_n] = \frac{1}{N}\sum_{n=1}^{N}$ $(\hat{x}_n^* \hat{\eta}_n)(\hat{x}_n^* \hat{\eta}_n)'$. Note that $\hat{\eta}_n = y_n - \hat{x}_n^{*'}\hat{\delta}$ and $\hat{M}$ which is an estimate of the asymptotic variance of $\sqrt{N}(\hat{\pi} - \pi)$ can be obtained from the first stage regression. By calculating the Wald statistic from the estimated asymptotic variance of $\hat{\delta}$, we perform various tests such as $H_0 : \beta_k = 0$ for $k = 1, 2, \ldots, K$ or $H_0 : \rho_j = 0$ (equivalently $H_0 : \gamma_j = 0$) for $j = 1, 2, \ldots, J$. Details of proofs for Lemma 3 and Proposition 3 are given in the appendix.

Detection of heterogeneous endogeneity is based on the null hypothesis, $H_0 : \rho_1 = \rho_2 = \cdots = \rho_J$. To implement the heterogeneous endogeneity test, we use the normalized residual vector $\hat{\nu}_{ij}^* = \hat{\Sigma}_\nu^{-1}(x_n - Z_n'\hat{\pi})$ from Eq. (2). In the second-stage, we use the control function, $\hat{\nu}_{ij}^*$ as an additional regressor in Eq. (1) and obtain $\hat{\gamma}_j$, which is the OLS estimate of $\gamma_j = \left[E(\nu_{ij}^{*\prime}\nu_{ij}^*)\right]^{-1}E\left(\nu_{ij}^* u_{ij}\right) = E[\nu_{ij}^* u_{ij}] = \sigma_u \rho_j$ for $j = 1, 2, \ldots, J$. Using the estimated coefficients, we perform a test for heterogeneous endogeneity under the null hypothesis, $H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_J$.[9] Since the asymptotic distribution and the standard error for $\hat{\gamma}_j$ are not standard due to the generated regressors $\hat{\nu}_{ij}^*$, we need to derive a proper asymptotic distribution taking into account the generated regressor problem. The asymptotic distribution of

[9] Note that when there exists heteroskedasticity in u_{ij}, we should consistently estimate the heteroskedasticity first and then employ normalized data for this test.

$\hat{\gamma}_j$, when $\hat{\nu}_{ij}^*$ is employed, is similar to one given in Proposition 3 except that $\mathbf{Z_n}$ should be normalized as $\mathbf{Z_n} = \left(D_n \times \hat{\Sigma}_\nu^{-1} Z_n'\right)$. Influence of the use of $\hat{\Sigma}_\nu$ on the asymptomatic distribution of $\hat{\gamma}_j$ is only marginal as long as $\hat{\Sigma}_\nu$ is consistent. Therefore, we can obtain satisfactory test results by simply computing the asymptotic variance in Proposition 3 with the normalized instruments, $\mathbf{Z_n} = \left(D_n \times \hat{\Sigma}_\nu^{-1} Z_n'\right)$.[10]

2.4 Finite sample properties: Monte-Carlo study

In addition to the heterogeneous endogeneity test, we also investigate how incorrectly assuming homogeneous endogeneity affects hypothesis testing for endogeneity. More specifically, we calculate and compare the power functions of tests with homogeneous and heterogeneous endogeneity assumptions using 10,000 sets of data generated by the following linear model:

$$
\begin{aligned}
y_{ij} &= \beta_0 + \beta_1 x_{ij} + u_{ij}, \qquad (16)\\
x_{ij} &= \pi_0 + \pi_1 z_{ij} + \nu_{ij}\\
\begin{bmatrix} \nu_{ij} \\ u_{ij} \end{bmatrix} &\overset{i.i.d.}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\nu^2 & \rho_j \sigma_\nu \sigma_u \\ \rho_j \sigma_u \sigma_\nu & \sigma_u^2 \end{bmatrix}\right)\\
i &= 1, 2, \ldots, N_j, \quad j = 1, 2 \qquad (17)
\end{aligned}
$$

where ρ_1 is not necessarily same as ρ_2. The assumption of homogeneous endogeneity often used in empirical applications implicitly imposes a restriction to the 2nd stage regression:

$$y_i = \beta_0 + \beta_1 x_i + \gamma \hat{\nu}_i + \epsilon_i$$

for $i = 1, 2, \ldots, N$. This leads to a test for endogeneity of the following null and alternative hypotheses[11]:

$$
\begin{aligned}
H_0 &: \gamma = 0 \,(\text{No Endogeneity})\\
H_a &: \gamma \neq 0 \,(\text{Endogeneity})
\end{aligned}
$$

Conversely, by recognizing that heterogeneous groups may suffer from different degrees of endogeneity, we consider the following regression model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_j \hat{\nu}_{ij} + \epsilon_{ij}$$

[10] Through Monte-Carlo simulations, we confirm that a power of the test for heterogeneous endogeneity is well controlled when a concentration parameter is higher than 10.

[11] To perform this endogeneity test, we implicitly assume that the instrumental variables z_{ij} are completely exogenous. Therefore, $\hat{\nu}_{ij}$ contains all the contaminated part of x_{ij} because in the first stage we project x_{ij} on the instrumental variable space.

where $i = 1, 2, \ldots, N_j$ and $j = 1, 2$. This model leads to a test of the null and alternative hypotheses:

$$H_0: \gamma_1 = 0 \text{ and } \gamma_2 = 0 \text{ (No Endogeneity)}$$
$$H_a: \gamma_1 \neq 0 \text{ or } \gamma_2 \neq 0. \text{ (Endogeneity)}$$

We perform the aforementioned hypothesis tests based on a heteroskedasticity-robust Wald test statistic and compare the sizes and powers of the two tests for endogeneity. Note that under H_0, the generated regressor problem disappears because $H = E\big[(\Gamma' \otimes x_n)Z_n\big]$ is a null matrix. Therefore, a conventional Wald-type test can be used for the endogeneity test.

In generating data, we use the parameter values, $\{\beta_0 = 1, \beta_1 = 0.5, \sigma_u^2 = 1, \delta_0 = 1, \delta_1 = 1\}$. The correlation parameter ρ_1 is fixed at 0 in all simulations, which means that the first group does not suffer from endogeneity. The different degrees of endogeneity for group 2 are assumed with ρ_2 ranging from -0.9 to 0.9, $\{\rho_2 = -0.9, -0.8, \ldots, 0, 0.1, \ldots, 0.8, 0.9\}$. For instance, when $\rho_1 = 0$ and $\rho_2 = 0$, we can obtain the sizes of the two tests. On the other hand, when $\rho_1 = 0$ and $\rho_2 = 0.5$, the powers of the two tests can be evaluated by calculating rejection rates. We make use of the concentration parameter to analyze the effects of strength of instruments on the power functions of these endogeneity tests:

$$\mu^2 = N\left(\frac{Var(\pi_1 z_{i,j})}{Var(\nu_{i,j})}\right).$$

We consider different values of $\mu^2 = \{10, 40, 70, 100\}$ and different numbers of observations in the two groups in our simulation. We note that though a smaller value of μ^2 means that the instrument is weaker, none of the instruments we study are weak in the sense of Nelson and Startz (1990a) and Nelson and Startz (1990b).

The upper left quadrant of Table 1 shows our simulation results when the concentration parameter is small ($\mu^2 = 10$). We consider three cases. The first is a sample size of 100 equally split between endogenous and exogenous groups ($N_1 = 50$, $N_2 = 50$). Here the power of a standard endogeneity test is quite low compared to that of our heterogeneous endogeneity approach. For example, at a correlation parameter of 0.5, the standard test rejects the (false) null only 12% of the time compared to the 69% rejection rate achieved by the heterogeneous approach. The upper right quadrant of Fig. 1 displays these two power curves.

Our second case is a sample size of 500 with 400 observations exogenous and 100 endogenous ($N_1 = 400$, $N_2 = 100$). Here the performance gap between the two approaches is even wider. Again, taking the case where the correlation parameter is 0.5, the standard approach rejects 6% of the time compared to 99% from the procedure that allows for heterogeneous endogeneity. The third case is a sample size of 500 with 100 observations exogenous and 400 endogenous ($N_1 = 100$, $N_2 = 400$). While the standard testing approach has a better rejection rate here, it is still far below that of our proposed test procedure at all levels of the correlation parameter.

Table 1 Rejection rates of endogeneity tests based on control function approach

ρ_2	Rejection rates for concentration parameter $\mu^2 = 10$						ρ_2	Rejection rates for concentration parameter $\mu^2 = 40$					
	N = 100		N = 500		N = 500			N = 100		N = 500		N = 500	
	$N_1 = 50, N_2 = 50$		$N_1 = 400, N_2 = 100$		$N_1 = 100, N_2 = 400$			$N_1 = 50, N_2 = 50$		$N_1 = 400, N_2 = 100$		$N_1 = 100, N_2 = 400$	
	Homogenous	Heterogenous	Homogenous	Heterogenous	Homogenous	Heterogenous		Homogenous	Heterogenous	Homogenous	Heterogenous	Homogenous	Heterogenous
0	0.046	0.053	0.060	0.053	0.050	0.053	0	0.048	0.052	0.049	0.048	0.049	0.049
0.1	0.054	0.072	0.045	0.109	0.054	0.117	0.1	0.056	0.077	0.061	0.120	0.078	0.134
0.2	0.066	0.141	0.053	0.338	0.083	0.361	0.2	0.082	0.159	0.055	0.349	0.170	0.425
0.3	0.075	0.268	0.052	0.692	0.124	0.707	0.3	0.125	0.313	0.067	0.707	0.318	0.797
0.4	0.094	0.465	0.054	0.934	0.176	0.929	0.4	0.193	0.538	0.084	0.929	0.550	0.974
0.5	0.121	0.685	0.062	0.996	0.274	0.991	0.5	0.283	0.778	0.101	0.994	0.753	0.998
0.6	0.159	0.865	0.066	1	0.401	1	0.6	0.383	0.932	0.111	1	0.906	1
0.7	0.195	0.965	0.073	1	0.567	1	0.7	0.498	0.990	0.143	1	0.981	1
0.8	0.264	0.997	0.080	1	0.738	1	0.8	0.626	1	0.170	1	0.999	1
0.9	0.340	1	0.089	1	0.894	1	0.9	0.734	1	0.200	1	1	1
ρ_2	Rejection rates for concentration parameter $\mu^2 = 70$						ρ_2	Rejection rates for concentration parameter $\mu^2 = 100$					
	N = 100		N = 500		N = 500			N = 100		N = 500		N = 500	
	$N_1 = 50, N_2 = 50$		$N_1 = 400, N_2 = 100$		$N_1 = 100, N_2 = 400$			$N_1 = 50, N_2 = 50$		$N_1 = 400, N_2 = 100$		$N_1 = 100, N_2 = 400$	
	Homogenous	Heterogenous	Homogenous	Heterogenous	Homogenous	Heterogenous		Homogenous	Heterogenous	Homogenous	Heterogenous	Homogenous	Heterogenous
0	0.047	0.053	0.050	0.052	0.053	0.054	0	0.050	0.057	0.053	0.053	0.047	0.047
0.1	0.060	0.080	0.054	0.116	0.091	0.142	0.1	0.065	0.078	0.109	0.116	0.107	0.155
0.2	0.100	0.166	0.067	0.354	0.232	0.480	0.2	0.112	0.181	0.338	0.353	0.312	0.531
0.3	0.160	0.351	0.083	0.708	0.479	0.844	0.3	0.183	0.370	0.692	0.711	0.621	0.899
0.4	0.253	0.591	0.108	0.939	0.762	0.983	0.4	0.296	0.624	0.934	0.937	0.861	0.993

Table 1 continued

ρ_2	Rejection rates for concentration parameter $\mu^2 = 70$						ρ_2	Rejection rates for concentration parameter $\mu^2 = 100$					
	N = 100		N = 500		N = 500			N = 100		N = 500		N = 500	
	$N_1 = 50$, $N_2 = 50$		$N_1 = 400$, $N_2 = 100$		$N_1 = 100$, $N_2 = 400$			$N_1 = 50$, $N_2 = 50$		$N_1 = 400$, $N_2 = 100$		$N_1 = 100$, $N_2 = 400$	
	Homogenous	Heterogenous	Homogenous	Heterogenous	Homogenous	Heterogenous		Homogenous	Heterogenous	Homogenous	Heterogenous	Homogenous	Heterogenous
0.5	0.375	0.829	0.156	0.995	0.921	1	0.5	0.434	0.841	0.996	0.996	0.975	1
0.6	0.514	0.953	0.193	1	0.990	1	0.6	0.583	0.967	1	1	0.997	1
0.7	0.638	0.996	0.262	1	0.999	1	0.7	0.720	0.997	1	1	1	1
0.8	0.758	1	0.315	1	1	1	0.8	0.830	1	1	1	1	1
0.9	0.856	1	0.378	1	1	1	0.9	0.911	1	1	1	1	1

To estimate the rejection rates we consider the following two models: Model 1 (homogenous endogeneity): $y_i = \beta_0 + \beta_1 x_i + \gamma v_i + \epsilon_i, \quad i = 1, 2, \ldots, N$ and test hypotheses are $H_0 : \gamma = 0$, $H_a : \gamma \neq 0$. Model 2 (heterogenous endogeneity): $y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_j v_{ij} + \epsilon_{ij}, \quad i = 1, 2, \ldots, N_j, \quad j = 1, 2$ and test hypotheses are $H_0 : \gamma_1 = \gamma_2 = 0, \quad H_a : \gamma \neq 0$. We generate 10,000 artificial samples for simulation. Endogeneity tests are conducted using Heteroskedasticity-robust Wald statistics. The concentration parameter (μ^2) measures the strength of the instrument and is defined as $\mu^2 = N\left(\frac{Var(\pi_1 z_{ij})}{Var(v_{ij})}\right)$

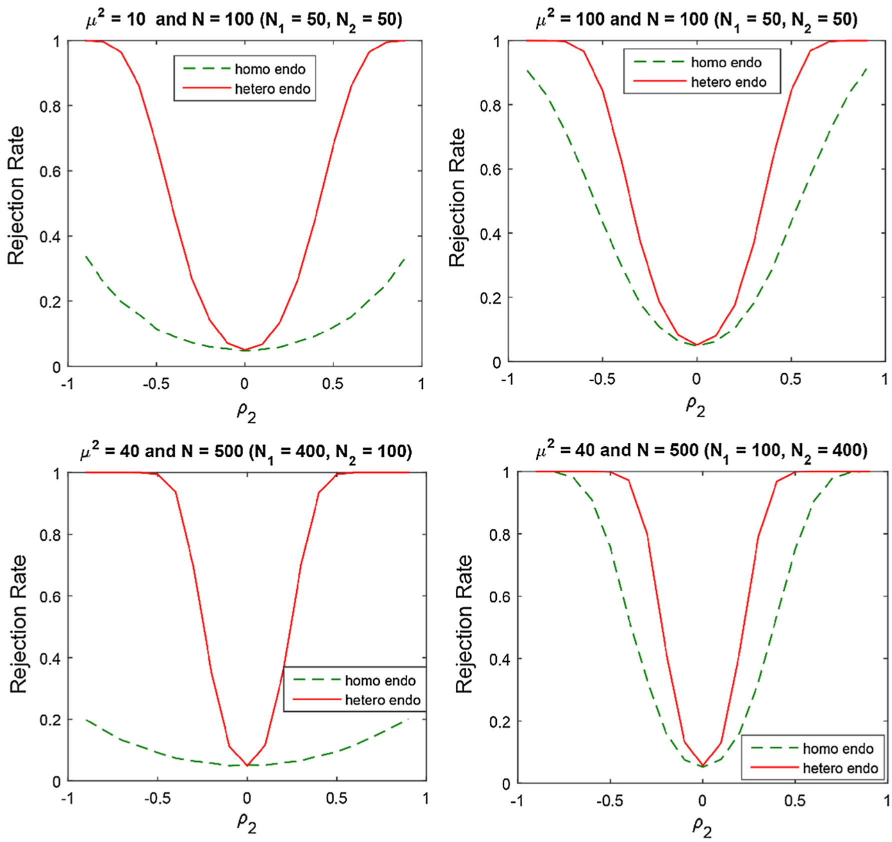

Fig. 1 Power curves of endogeneity tests based on control function approach: power comparison with different values of concentration parameter under heterogeneous endogeneity. To generate the power curves of endogeneity test based on control function approach, we consider the following two models: Model 1 (homogenous endogeneity): $y_i = \beta_0 + \beta_1 x_i + \gamma v_i + \epsilon_i$, $i = 1, 2, \ldots, N$ and test hypotheses are $H_0 : \gamma = 0$, $H_a : \gamma \neq 0$. Model 2 (heterogenous endogeneity): $y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_j v_{ij} + \epsilon_{ij}$, $i = 1, 2, \ldots, N_j$, $j = 1, 2$ and test hypotheses are $H_0 : \gamma_1 = \gamma_2 = 0$, $H_a : \gamma \neq 0$. We generate 10,000 artificial samples for simulation. The red bold line is the power curve under the heterogeneous endogeneity assumption. The dotted green line is the power curve under the homogeneous endogeneity assumption. Endogeneity tests are conducted using Heteroskedasticity-robust Wald statistics. The concentration parameter (μ^2) measures the strength of the instrument and is defined as $\mu^2 = N\left(\frac{Var(\pi_1 z_{ij})}{Var(v_{ij})}\right)$

The other three quadrants of Table 1 and Fig. 1 contain simulation results for these same three cases at larger values of the concentration parameter. While a greater proportion of the sample being endogenous clearly helps the rejection rate of the standard approach, there are considerable gains in power from using our proposed test procedure in almost all cases.

Under heterogeneous endogeneity, the conventional 2SLS no longer provide an asymptotically efficient estimator as shown in Proposition 2. In particular, we waste considerable amount of information in the regressors x_{ij} by using 2SLS when some subgroups are not subject to endogeneity. The CF procedure introduced in Sect. 3.2

produces a straightforward tool to resolve this issue by fully utilizing relevant information in estimating model parameters. The following is the summary of the control function approach for heterogeneous endogeneity.

Control Function Procedure under Heterogeneous Endogeneity

Step 1 Run the OLS regression of x_{ij} on all exogenous variables Z_{ij}:

$$X = Z\pi + V$$

and obtain OLS residuals $\hat{v}_{ij}$.

Step 2 Construct $\widehat{V} = diag[\hat{v}_1, \hat{v}_2, \hat{v}_3, \ldots \hat{v}_J]$ and run the OLS regression of y_{ij} on x_{ij} and $\hat{v}_{ij}$:

$$Y = X\beta + \widehat{V}\Gamma + E$$

and obtain $\hat{\beta}$, and $\widehat{\Gamma}$. Conduct a hypothesis testing of the null and alternative $H_0 : \gamma_j = 0, \quad H_a : \gamma_j \neq 0$ using a robust test statistic for $j = 1, 2, \ldots, J$ based on Proposition 3.

Step 3 Construct $\widetilde{V} = \begin{bmatrix} \hat{v}_e \\ 0 \end{bmatrix}$ and run the OLS regression of y_{ij} on x_{ij}, and $\hat{v}_{ij}$ for endogenous groups and only on x_{ij} for exogenous groups:

$$Y = X\beta + \widetilde{V}\Gamma_e + E$$

In this section, we perform Monte-Carlo simulations to show the utility of the proposed CF procedure in finite samples.[12] In studying the finite sample properties of alternative estimators including the CF estimator, 10,000 sets of samples are generated from the model in Eqs. (16) and (17). As before, we assume that one group is subject to endogeneity and another is not by using $\rho_1 = 0$ and $\rho_2 = -0.7$ in the data generating process. The values of all other parameters, the concentration parameter, and the number of observations in each group are assumed to be all the same as those used in the previous simulation.

Figure 2 reports the sampling distributions of OLS, 2SLS, and CF estimators of the parameter β_1 under heterogeneous endogeneity. Regardless of different values of the concentration parameter ($\mu^2 = \{10, 100\}$), the CF procedure gives the best result. Notice that while the OLS estimator is severely biased, the sampling distributions of the 2SLS and CF estimators are centered around the true value 0.5 only with $N = 100$. It is noticeable that the efficiency gain from the proposed CF procedure is significant. Especially, the variance of the sampling distribution of the CF estimator is much smaller than that of the 2SLS estimator when $\mu^2 = 10$. They are comparable when $\mu^2 = 100$. It is worth mentioning that there still exist considerable benefits of using the CF approach even when the sample size is relatively large as shown in Fig. 2. When $N = 500$, the sampling distribution of the estimator based on the CF approach has the most desirable finite sample properties comparing other alternative methods.

[12] In the simulation study, we assume that the exogenous group is priorly known. In the case where the exogenous group is not known, a straightforward test can be performed to identify the group.

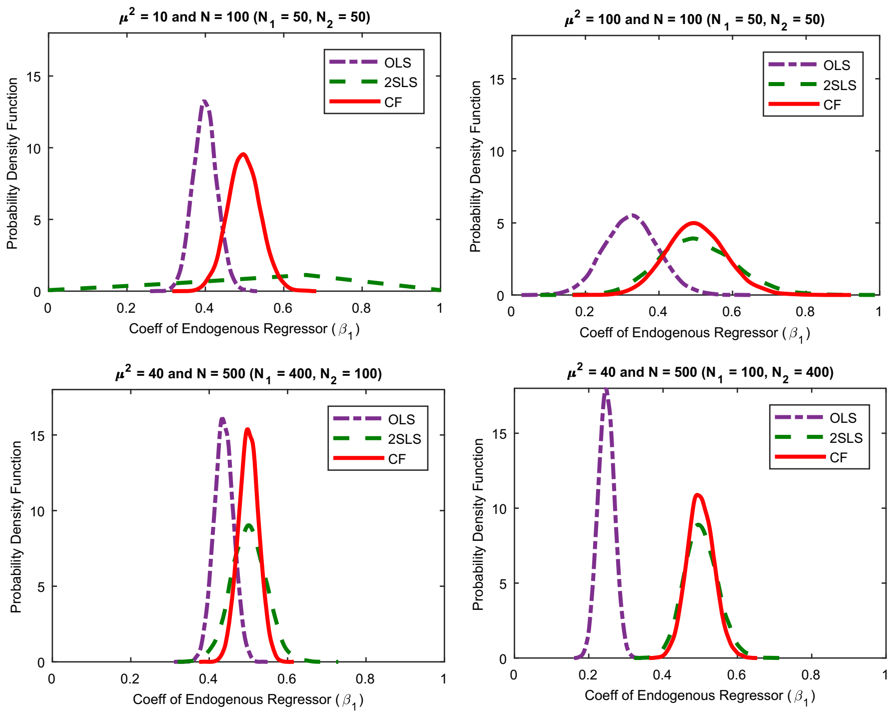

Fig. 2 Sampling distributions of β_1 from OLS, 2SLS, and control function (CF) methods: a comparison with different values of concentration parameter (μ^2) under heterogeneous endogeneity. The data generating process is given by $y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{ij}$ and $x_{ij} = \pi_0 + \pi_1 z_{ij} + v_{ij}$ where $\beta_0 = 1$, $\beta_1 = 0.5$, $\pi_0 = 1$, $\pi = 1$ and $\begin{bmatrix} v_{ij} \\ u_{ij} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \rho_j \sigma_v \sigma_u \\ \rho_j \sigma_v \sigma_u & \sigma_u^2 \end{bmatrix}\right)$. We set $\sigma_u^2 = 1$, $\rho_1 = 0$, $\rho_2 = -0.7$ and generate 10,000 artificial samples to estimate the probability density functions

Table 2 provides more specific evidence about the relative performance of 2SLS and the proposed CF estimator in simulations. Efficiency gains (as measured by how much bigger the coefficient's standard error is in 2SLS compared to our CF estimator) from adopting our approach can be quite substantial as the median gain reported is around 42%. Regardless of sample size and fraction of the observations that are endogenous, efficiency gains are large for cases where the instrument is not overwhelmingly strong. For example with 500 observations, 80% of which are exogenous, our CF estimator has a standard deviation half the size of the 2SLS estimator (97% efficiency gain) when the concentration parameter is 30. With the same sample but only 20% of the observations exogenous, the efficiency gain from our CF procedure versus 2SLS is 34%. Note that given our model has one regressor, there is a direct correspondence between the concentration parameter and the first stage F-statistic in a 2SLS procedure, so this case (concentration parameter of 30) is the case where the instrument is quite strong by conventional measures. Even in such an extreme case where the concentration parameter is 100 and only 20% of the observations are exogenous, we still achieve about a 15% efficiency gain. With the same concentration parameter but 80% of the observations exogenous, the efficiency gain is around 40%.

3 Model extensions

This section extends the model given in Eqs. (1), (2), and (3) in two important directions. The first extension is to allow for unknown group membership due to a lack of proper economic theory. We introduce two statistical methods to directly identify endogenous and exogenous groups based on data. The second extension is to allow for unobserved heterogeneity in the coefficients of the endogenous variables by employing the correlated random coefficient (CRC) model proposed by Heckman and Vytlacil (1998) and applied by Wooldridge (2015).

3.1 Identifying exogenous and endogenous groups

Suppose that there is a set of categorical variables[13] that potentially define the exogenous and endogenous groups. One can easily create non-overlapping groups and a group index variable for all individual units based on such variables. Let $g_i \in \{1, 2, \ldots, G\}$ denote the group index variable where G indicates the count of all groups. Using g_i, we rewrite the model as follows:

$$y_i = x_i'\beta + v_i'\gamma_{g_i} + \eta_i \tag{18}$$

where $i = 1, 2, \ldots, N$. Because all individual units that belong to the same group (i.e. all i such that $g_i = g$) share the same value of γ_{g_i}, the model can be easily estimated by employing group dummies as additional regressors. In order to test for which groups are subject to endogeneity and which groups are not (i.e. $\gamma_g = 0$ v.s. $\gamma_g \neq 0$ for $g = 1, 2, \ldots, G$), we can apply the asymptotic distribution in Proposition 3 or bootstrap methods. In our empirical application, endogenous and exogenous groups are selected based on this approach.

We provide another statistical procedure for estimating the group membership when no categorical variable is available but individual units have time series observations. The proposed method is built upon the well-known *kmeans clustering* algorithm by Forgy (1965). Consider the following panel data model which is featured by heterogeneous endogeneity:

$$y_{i,t} = x_{i,t}'\beta + v_{i,t}'\gamma_{g_i} + \eta_{i,t} \tag{19}$$

where $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$. Let $\phi = \{g_1, g_2, \ldots, g_N\}$ denote a particular grouping of all individual units, $i \in \{1, 2, \ldots, N\}$. Our proposed estimator that simultaneously takes into account unknown group membership and model parameters is defined by the minimization problem:

[13] A continuous variable can be used to create a categorical variable with proper threshold values.

Table 2 Mean and standard deviation of the sampling distributions of OLS, 2SLS and proposed control function (CF) estimators for the endogenous regressor

	N = 100				N = 500				N = 500			
	$N_1 = 50$, $N_2 = 50$				$N_1 = 400$, $N_2 = 100$				$N_1 = 100$, $N_2 = 400$			
μ^2	OLS	2SLS	Proposed CF	Efficiency gain	OLS	2SLS	Proposed CF	Efficiency gain	OLS	2SLS	Proposed CF	Efficiency gain
10	0.399	0.517	0.499	16.096	0.460	0.502	0.500	3.121	0.339	0.503	0.500	1.829
	(0.031)	(0.666)	(0.041)		(0.014)	(0.047)	(0.015)		(0.012)	(0.047)	(0.026)	
20	0.370	0.510	0.499	2.196	0.448	0.501	0.500	2.265	0.291	0.502	0.500	1.470
	(0.041)	(0.118)	(0.054)		(0.019)	(0.046)	(0.020)		(0.017)	(0.046)	(0.031)	
30	0.353	0.509	0.500	1.807	0.441	0.500	0.500	1.972	0.264	0.502	0.500	1.344
	(0.048)	(0.110)	(0.061)		(0.022)	(0.046)	(0.023)		(0.019)	(0.046)	(0.034)	
40	0.342	0.507	0.501	1.641	0.437	0.500	0.500	1.740	0.247	0.502	0.501	1.267
	(0.054)	(0.108)	(0.066)		(0.025)	(0.045)	(0.026)		(0.022)	(0.045)	(0.036)	
50	0.334	0.504	0.500	1.488	0.434	0.500	0.500	1.647	0.236	0.502	0.501	1.224
	(0.059)	(0.105)	(0.071)		(0.026)	(0.045)	(0.027)		(0.023)	(0.045)	(0.037)	
60	0.330	0.505	0.500	1.445	0.432	0.500	0.500	1.572	0.229	0.501	0.500	1.196
	(0.062)	(0.105)	(0.073)		(0.028)	(0.045)	(0.029)		(0.025)	(0.045)	(0.038)	
70	0.328	0.506	0.501	1.397	0.431	0.500	0.500	1.511	0.225	0.501	0.500	1.181
	(0.065)	(0.106)	(0.076)		(0.029)	(0.046)	(0.030)		(0.027)	(0.046)	(0.039)	
80	0.328	0.507	0.502	1.351	0.430	0.499	0.500	1.440	0.221	0.501	0.501	1.146
	(0.067)	(0.104)	(0.077)		(0.031)	(0.045)	(0.031)		(0.028)	(0.044)	(0.039)	

Table 2 continued

	N = 100				N = 500				N = 500			
	$N_1 = 50$, $N_2 = 50$				$N_1 = 400$, $N_2 = 100$				$N_1 = 100$, $N_2 = 400$			
μ^2	OLS	2SLS	Proposed CF	Efficiency gain	OLS	2SLS	Proposed CF	Efficiency gain	OLS	2SLS	Proposed CF	Efficiency gain
90	0.326	0.505	0.502	1.318	0.430	0.501	0.500	1.421	0.220	0.501	0.500	1.142
	(0.070)	(0.104)	(0.079)		(0.032)	(0.046)	(0.032)		(0.029)	(0.045)	(0.039)	
100	0.325	0.504	0.500	1.295	0.430	0.501	0.500	1.379	0.219	0.500	0.500	1.147
	(0.071)	(0.103)	(0.080)		(0.032)	(0.045)	(0.033)		(0.030)	(0.045)	(0.039)	

The data generating process is given by $y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{ij}$ and $x_{ij} = \pi_0 + \pi_1 z_{ij} + v_{ij}$ where $\begin{bmatrix} v_{ij} \\ u_{ij} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \rho_j \sigma_v \sigma_u \\ \rho_j \sigma_v \sigma_u & \sigma_u^2 \end{bmatrix}\right)$. We generate 10,000 artificial samples for simulation. The standard deviations are shown in the parentheses. The 'Efficiency Gain' is defined as the ratio of the standard deviations of 2SLS and CF estimators, that is $\sigma_{2SLS}/\sigma_{CF}$. The concentration parameter (μ^2) measures the strength of the instrument and is defined as $\mu^2 = N\left(\frac{Var(\pi_1 z_{ij})}{Var(v_{ij})}\right)$

$$\left(\hat{\beta}_{CF}, \hat{\gamma}, \hat{\phi}\right) = \underset{\beta,\gamma,\phi}{argmin} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{i,t} - x_{i,t}' \beta - \hat{v}_{i,t}' \gamma_{g_i} \right)^2$$

where $g_i \in \{1, 2, \ldots, G\}$; $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_G\}$. Suppose that the values of β and γ are given. Then the group assignment is optimally determined by minimizing the following objective function:

$$\hat{\phi}(\beta, \gamma) = \underset{\phi}{argmin} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{i,t} - x_{i,t}' \beta - \hat{v}_{i,t}' \gamma_{g_i} \right)^2$$

where $\hat{\phi}(\beta, \gamma) = \{\hat{g}_1(\beta, \gamma), \hat{g}_2(\beta, \gamma), \ldots, \hat{g}_N(\beta, \gamma)\}$. Since the optimal group assignment depends on $\{\beta, \gamma\}$, we can define the estimator of $\{\beta, \gamma\}$ as:

$$\left(\hat{\beta}_{CF}, \hat{\gamma}\right) = \underset{\beta,\gamma}{argmin} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{i,t} - x_{i,t}' \beta - \hat{v}_{i,t}' \gamma_{\hat{g}_i(\beta,\gamma)} \right)^2.$$

The estimator of ϕ is $\hat{\phi}(\hat{\beta}_{CF}, \hat{\gamma})$ given $\hat{\beta}_{CF}, \hat{\gamma}$.

The theoretical underpinning of this approach is based on the study of Bonhomme and Manresa (2015) who develop a group fixed-effects (GFE) estimator. The GFE estimator is designed to estimate group-specific parameters and group membership at the same time while allowing for arbitrary correlations between the group fixed effects and covariates. One of the main theoretical results in Bonhomme and Manresa (2015) is that all model parameters including the latent group membership are consistently estimated without explicitly modeling the link between the group fixed effects and the covariates, when both N and T go to infinity. In addition to the important theoretical result, they successfully show via Monte-Carlo simulation that the GFE estimator performs well even when T is moderately small. The fact that the generated regressor $\hat{v}_{i,t}$ converges to its true value asymptotically and the theoretical conclusion of Bonhomme and Manresa (2015) directly suggest that the model parameters and the group membership of our model can be consistently estimated without any variables that define the endogenous and exogenous groups in the population. Next we summarize each step for the new method that combines the CF approach and the *kmeans clustering* algorithm.

Parameter estimation and group assignment

Step 1 Obtain the residual vectors $\hat{v}_{i,t}$ for $t = 1, 2, ..., T$ and $i = 1, 2, \ldots, N$ using the control function approach described in Sect. 2.3.

Step 2 Determine the initial values of β and γ_g for $g = 1, 2, \ldots, G$. We assume that the first group ($g = 1$) is the exogenous group for identification by restricting $\gamma_1 = 0$.

Step 3 Assign i-th individual unit to one of the groups in $\{1, 2, ..., G\}$ given the values of β and γ_g:

$$\hat{g}_i(\beta, \gamma_2) = \underset{g \in \{1,2,...,G\}}{argmin} \sum_{t=1}^{T} \left(y_{i,t} - x'_{i,t}\beta - \hat{\nu}'_{i,t}\gamma_g \right)^2.$$

Step 4 Obtain $\hat{\beta}_{CF}$ and $\hat{\gamma}_g$ for $g = 2, 3, \ldots, G$ given $\hat{\phi} = \{\hat{g}_1, \hat{g}_2, \ldots \hat{g}_N\}$:

$$\left(\widehat{\beta}_{CF}, \hat{\gamma}_2\right) = \underset{\beta,\gamma_2}{argmin} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{i,t} - x'_{i,t}\beta - \hat{\nu}'_{i,t}\gamma_{\hat{g}_i} \right)^2.$$

The analytic solution to the above minimization problem is given in Eqs. (9) and (10). Thus, this step does not require any optimization algorithm. Set $\beta = \hat{\beta}_{CF}$ and $\gamma_g = \hat{\gamma}_g$ and compute the sum-of-squared residuals, $\sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_i - x'_i\beta - \hat{\nu}'_i\gamma_{\hat{g}_i} \right)^2$ for the next iteration.

Step 5 Repeat Step 3 and Step 4 until the difference in the sum-of-squared residuals between two consecutive iterations becomes less than a user-specified tolerance level, κ.

The iterative algorithm usually converges very fast even when N is moderately large. If there exist multiple local minima, one can find the global minimum that provides the lowest sum-of-squared residuals by employing different initial values of the model parameters in Step 2. Estimation uncertainty in the control function, $\hat{\nu}_i$ and the group membership, $\hat{\phi}$ should be considered in inference for $\hat{\beta}_{CF}$. Because the analytic expression of the asymptotic variance for $\hat{\beta}_{CF}$ is not tractable, we can use bootstrap methods for inference. In *Appendix B*, we provide simulation results for the proposed algorithm and evaluate its performance for different T. The sampling distributions of $\hat{\beta}_{CF}$ in the two different cases that use the true group membership or estimate the unknown group membership become almost identical as T increases to 20.

3.2 Heterogeneity in β

So far, we have assumed that the coefficients of the endogenous variables are homogeneous. This section introduces a regression model that incorporates heterogeneity in β:

$$y_i = x'_i\beta_i + \nu'_i\gamma_{g_i} + \eta_i \tag{20}$$

where $\beta_i = \gamma_0 + \epsilon_i$ and $E[\epsilon_i] = 0$. The equation for β_i implies that the coefficient γ_0 represents the average marginal effect of x_i as $E[\beta_i] = \gamma_0$. By assuming that the instrumental variables are exogenous of ϵ_i (i.e. $E[\epsilon_i|z_i] = 0$) and the source of heterogeneity is linearly associated with ν_i as in Wooldridge (2015), we can obtain the conditional expectation equation of y_i given x_i and z_i:

$$
\begin{aligned}
E[y_i|x_i,z_i] &= E[y_i|x_i,\nu_i] \\
&= x_i'\gamma_0 + x_i'E[\epsilon_i|\nu_i] + \nu_i'\gamma_{g_i} + \eta_i \quad = x_i'\gamma_0 + (x_i \circ \nu_i)'\psi + \nu_i'\gamma_{g_i} + \eta_i
\end{aligned} \tag{21}
$$

where $E[\epsilon_i|\nu_i] = (\nu_i \circ \psi)$ and $\circ$ is the Hadamard product. Equation (21) suggests a simple control function approach given below:

$$
y_i = x_i'\gamma_0 + (x_i \circ \hat{\nu}_i)'\psi + \hat{\nu}_i'\gamma_{g_i} + \eta_i^* \tag{22}
$$

where $E[\eta_i^*|x_i,z_i] = 0$. The effect of ignoring heterogeneity in β seems worth discussing. If ψ is not zero in the population and heterogeneity in β is not taken into account, the error term η_i^* absorb the heterogeneity which is related with ν_i. In the case, since the correlation between the control function and the error term η_i^* is no longer 0, the CF estimator for γ_0 becomes biased. For this reason, we recommend conducting a robust check by including the additional interaction terms $(x_i \circ \hat{\nu}_i)$ in the regression model.

We can take a more flexible functional form for β_i by including some exogenous variables $\tilde{z}_i$ as $\beta_i = \gamma_0 + \tilde{z}_i'\gamma_1 + \epsilon_i$ where $E[\epsilon_i] = 0$ and $E[\epsilon_i|\nu_i] = (\nu_i \circ \psi)$. The variables $\tilde{z}_i$ could be a subset of z_i. This extension leads to the following regression model:

$$
y_i = x_i'\gamma_0 + (x_i \circ \tilde{z}_i)'\gamma_1 + (x_i \circ \hat{\nu}_i)'\psi + \hat{\nu}_i'\gamma_{g_i} + \eta_i^*. \tag{23}
$$

where $(x_i \circ \tilde{z}_i)$ is the vector of interaction terms between x_i and $\tilde{z}_i$. One can perform inference for the model parameters using the asymptotic distribution in Proposition 3 or bootstrap methods.

4 Empirical example

In this section, we use the data and basic model from Abadie et al. (2002) to illustrate our approach. Abadie et al. (2002) study the effect of a job-training program on earnings. The offer of training is randomized, but the take-up is not, making training an endogenous regressor in the earnings equation. Abadie et al. (2002) use the offer as their instrument to estimate the causal effect of the training on earnings. We choose this as our example because the validity of the instrument is not in question and we can focus on whether the training variable exhibits heterogeneous endogeneity knowing that we have a good instrument.

We illustrate our approach by testing for and modeling heterogeneous endogeneity by race in the data. We consider 9979 observations from the Abadie et al. (2002) data, of which 7070 are white and 2909 are black. Table 3 provides summary statistics on our sample. A large literature including Holzer (1987), Smith (1993), Bound and Holzer (1993), Altonji and Blank (1999), and others shows that white and black workers have very different labor force attachment and elasticities with respect to wage. Therefore, it is natural to test whether white and black workers' decisions to accept the training offer are differentially correlated with their earnings outcomes.

When we implement our test for heterogeneous endogeneity, we find that the null hypothesis of homogeneous heterogeneity is decisively rejected at the 0.001 level as

Table 3 Means and standard deviations of the national job training program earnings sample

		Assignment			Treatment		
	Entire sample	Treatment	Control	Difference (t-stat)	Trainees	Non-trainees	Difference (t-stat)
No. of observation	9979	6672	3307		4292	5687	
Treatment	0.430	0.635	0.016	0.620			
(Job training)	(0.495)	(0.481)	(0.124)	72.848			
Outcome variable	15,967	16,357	15,179	1,178	17,544	14,776	2,768
(30 Month earnings)	(16,848)	(16,218)	(17,139)	(3.292)	(17,141)	(16,525)	(8.153)
Baseline characteristics							
High school or GED	0.719	0.723	0.709	0.014	0.746	0.698	0.048
	(0.435)	(0.433)	(0.440)	(1.555)	(0.421)	(0.445)	(5.427)
Married	0.276	0.283	0.263	0.019	0.287	0.268	0.020
	(0.434)	(0.437)	(0.428)	(2.101)	(0.440)	(0.430)	(2.234)
Men	0.461	0.459	0.467	−0.008	0.445	0.474	− 0.028
	(0.499)	(0.498)	(0.499)	(− 0.765)	(0.497)	(0.499)	(−2.825)
Women	0.539	0.541	0.533	0.013	0.555	0.526	0.008
	(0.499)	(0.498)	(0.499)	(0.765)	(0.497)	(0.499)	(2.825)
Black	0.292	0.294	0.287	0.007	0.292	0.291	0.001
	(0.454)	(0.456)	(0.452)	(0.703)	(0.455)	(0.454)	(0.081)
Worked less than 13	0.453	0.453	0.451	0.002	0.448	0.456	− 0.008
weeks in past year	(0.471)	(0.471)	(0.473)	(0.166)	(0.473)	(0.470)	(− 0.808)

The summary statistics table reports mean and standard deviations (in brackets) for the National JTPA study 30-month earnings sample. The two panels ‘Assignment’ and ‘Treatment’ show differences in means by random assignment and self selected treatment status. The column ‘Difference’ reports the t-statistic (in parentheses) for the null hypothesis of equality in means reported in the right panel of Table 4. When we examine the component parts of the test, we can see that training is significantly endogenous for white (significant at the 0.000 level) but not for black (insignificant even at the 0.878 level) as shown in the left panel of Table 4. We conclude that job training is an endogenous regressor for the whites in our sample but not for the blacks.

We then estimate the treatment effect in our data by OLS, by conventional 2SLS (homogeneous endogeneity) and by our proposed control function approach that allows for heterogeneous endogeneity. The estimation results of all three methods are reported in Table 5. As expected, the OLS result produces the largest estimated treatment effect, while the 2SLS and CF models produce generally similar estimates. The estimated impacts of the job training on annual average earnings are \$1,575 and \$1,620 respectively based on 2SLS and the CF estimation. The difference between the 2SLS and the CF estimates is only \$45, an economically insignificant amount.

Although the 2SLS and CF estimates are similar, comparing the standard errors in Table 5 we see that the CF approach is able to provide a more precise estimate of the treatment effect. Its standard error is about 16% smaller than that of the 2SLS estimate, illustrating the efficiency gain from modeling heterogeneous endogeneity when it is present. Since the instrument here is very strong and the sample for the endogenous group is large, the efficiency gain in this example is relatively modest.[14] This indicates that even in the "worse case" scenario for our approach, it is still possible to reduce estimation uncertainty in practice.

We note that the CF estimate of the treatment effect does not differ much from the full-sample 2SLS. Proposition 1 shows that the proposed CF is a weighted average of the two consistent estimators obtained using the sub-sample of the exogenous group for the OLS estimator and the sub-sample of the endogenous group for the 2SLS estimator. Since the full-sample 2SLS estimator is also consistent, the CF estimate should be fairly close to it, as long as the variance of the 2SLS estimator is not too large.

Following the correlated random coefficient model proposed by Heckman and Vytlacil (1998) and Wooldridge (2015), we also consider the heterogeneous impacts of job training on earnings. We define 4 groups based on gender and race to more accurately capture heterogeneous endogeneity. In the first model (*Model 1)* that takes into account the heterogeneous treatment effects, we assume that the coefficient β_i of the job training program is determined by the sum of γ_0 and ϵ_i which represents the average effect and the heterogeneity shock respectively as in Eq. (20). We further assume that the source of heterogeneity ϵ_i is independent of whether one receives the job training or not. Under these assumptions, we can consistently estimate the average effect γ_0 using Eq. (22) that excludes the interaction term between the job training variable and the control function. The left panel of Table 6 reports the result of this model. The results indicate that the estimated treatment effects are very similar in the baseline approach (2SLS) that assumes all groups are endogenous groups and our proposed approach that assumes only white groups are endogenous groups. However,

[14] Given that we have a dummy potentially endogenous variable, we avoid the forbidden regression problem in the first stage by following the procedure outlined on pages 189–191 of Angrist and Pischke (2009) Results are very similar to those reported in the text if we simply use a linear probability model in the first stage.

Table 4 Wald test for detecting endogeneity and testing heterogeneous endogeneity for groups based on race

	Test for detecting endogeneity		Test for heterogeneous endogeneity
	$H_0 : \gamma_1$ (White) $= 0$	$H_0 : \gamma_2$ (Black) $= 0$	$H_0 : \gamma_1$ (White) $= \gamma_2$ (Black)
Test statistics value	14.112	0.023	10.409
5% level critical value	[3.841]	[3.841]	[3.841]
p-value	(0.000)	(0.878)	(0.001)
Conclusion	Reject H_0	Can't reject H_0	Reject H_0

The left panel reports the proposed Wald test statistics of γ_1 and γ_2 obtain from the following linear regression, $income_{ij} = x_{ij}^{*\prime}\delta + \epsilon_{ij} = \beta_0^* + \beta_1^* treatment_{ij} + \gamma_1 \widehat{\nu}_{i1} + \gamma_2 \widehat{\nu}_{i2} + x_i^{\prime}\beta^* + \epsilon_{ij}$ where $\widehat{\nu}_1$ (for white) and $\widehat{\nu}_2$ (for black) are the control variables obtain from the first stage regression. The set of control variables x includes the baseline characteristics along with a set of dummies for age categories. The Wald-statistics is given by $W_{\gamma_i} = \hat{\gamma}_i^2 / \widehat{AVar}(\hat{\gamma}_i) \sim {}^{2}_{1}$ where $i = 1$ and 2. The Wald-statistics are estimated using the asymptotic variance-covariance matrix of $\hat{\delta}$ as shown in Proposition 3. $\widehat{AVar}(\hat{\gamma}_1)$ and $\widehat{AVar}(\hat{\gamma}_2)$ are corresponding element of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ in the asymptotic variance-covariance matrix of $\hat{\delta}$. Therefore, this proposed Wald test incorporates the generating regressor problem by using Proposition 3. The right panel shows the proposed heterogeneity test statistic using the Wald test where the test statistics is defined as $W = \left(R\hat{\delta}\right)^{\prime}\left[\widehat{AVar}(\hat{\delta})\right]^{-1}\left(R\hat{\delta}\right) \sim {}^{2}_{1}$ and R is a row vector includes the restrictions on the estimated parameters $\hat{\delta}$

Table 5 National job training program earnings sample treatment effects from OLS, 2SLS and proposed control function (CF) estimators

Standard and proposed estimators	Group name	No. of observations		Treatment effect			
	Endogenous group	All	Endogenous group	Estimate (se)	t-stat	R^2	MSE ($\times 10^6$)
Ordinary least squares (OLS)	None	9979	0	2825 (328)	8.620	10.093	257.69
Two stage least squares (2SLS)	All	9979	9979	1575 (328)	3.112	0.087	259.44
Proposed CF estimator	White	9979	7070	1620 (438)	3.699	0.102	255.24

The table reports the OLS and two stage least square coefficients of the following regression: $income = \beta_0 + \beta_1 treatment + X\beta + u$ where for 2SLS estimation, the random assignment is used as an instrument for self-selected treatment status. The set of control variables includes dummies for age categories, high-school or GED, married, women and black. The proposed control function estimators are obtained from the regression, $income = \beta_0^* + \beta_1^* treatment + \gamma\widehat{\nu} + X\beta^* + \epsilon$ where the first stage estimated residuals $\widehat{\nu}$ are replaced by a null vector for the exogenous subgroup(s). The robust standard errors are shown in the parenthesis. MSE is referred as mean squared error

Table 6 National job training program earnings sample heterogeneous treatment effects by race and gender

	Model 1		Model 2	
	Baseline	Proposed	Baseline	Proposed
Treatment	1641	1612	1609	1586
	(543)	(442)	(986)	(979)
Treatment × control function			98	74
			(2562)	(2495)
Control function × White male	3870	3899	3861	3894
	(886)	(828)	(912)	(839)
Control function × White female	1696	1726	1690	1723
	(877)	(816)	(891)	(820)
Control function × Black male	937		929	
	(1247)	(1268)		
Control function × Black female	−893		−901	
	(1109)		(1127)	
R^2	0.108	0.108	0.108	0.108
No. of observations	9979	9979	9979	9979

The table reports the OLS coefficients of two models specified in Eqs. (20) and (22) where the outcome variable is income and the endogenous regressor treatment indicates job training. The control function $(\widehat{\nu})$ is obtained from the first stage estimated residuals. The robust standard errors are shown in the parenthesis

compared to the baseline model, the standard error is much smaller when we apply our proposed approach by treating black racial groups as exogenous groups.

As explained in Sect. 3.2, the estimates in the previous model are likely to be biased when the heterogeneity in the treatment effect is correlated with the job training variable. To address this potential problem, we include the interaction term between the job training variable and the control function in the second model. The result of the second model (*Model 2*) is reported in the right panel of Table 6. As before, the estimated average treatment effects are very similar in magnitude in both the baseline and our proposed control function approach. However, the efficiency gain seems much smaller compare to *Model 1*. This can be explained by the substantial increase in estimation uncertainty as a result of including the new interaction term. Note that the estimated coefficient of the interaction term is close to zero while the standard error is very large. Also when we consider heterogeneous treatment effects in *Model 1* and *Model 2*, our efficiency gains are relatively modest because our instrument is very strong. The first stage F-statistics is 235.

Differ from *Model 2* in which the heterogeneous treatment effects are controlled for only by the interaction term between the control function and the job training variable, *Model 3* assumes that the treatment effects are heterogeneous across some observed demographic characteristics such as race and gender. Thus *Model 3* is similar to Eq. (23) except it contains more exogenous control variables. In *Model 3* we allow the treatment effects to vary across all four groups: white male, white female, black male, and black female. The corresponding results are shown in the left panel of Table 7. Both

Table 7 Stratifying groups for heterogeneous treatment effects and heterogeneous endogeneity based on race and gender

	Model 3		Model 4	
	Baseline	Proposed	Baseline	Proposed
Treatment × White male	1951	2057		
	(1322)	(1308)		
Treatment × White female	1037	1130		
	(1178)	(1166)		
Treatment × Black male	2589	2705		
	(1706)	(1312)		
Treatment × Black female	1772	1233		
	(1479)	(1200)		
Treatment × Male			2176	2373
			(1203)	(1132)
Treatment × Female			1293	1161
			(1059)	(1026)
Treatment × control function	− 19	− 331	− 133	− 255
	(2585)	(2519)	(2578)	(2507)
Control function × White male	3567	3588	3389	3242
	(1189)	(1188)	(1077)	(963)
Control function × White female	2320	2347	2104	2287
	(1124)	(1123)	(1024)	(915)
Control Function × Black male	− 0.547	458		
	(1868)		(1391)	
Control function × Black female	− 1012		− 508	
	(1615)		(1224)	
R^2	0.108	0.108	0.108	0.108
No. of observations	9979	9979	9979	9979

The table reports the OLS coefficients of the model specified in Eq. (23) where the outcome variable is income and the endogenous regressor treatment indicates job training. The control function ($\widehat{\nu}$) is obtained from the first stage estimated residuals. The robust standard errors are shown in the parenthesis

the baseline and proposed method estimates suggest that the heterogeneous treatment effects are similar to white male and black male and as well as white female and black female. Efficient gains are observed as before in the proposed approach.

In *Model 4* we allow the heterogeneous treatment effect only by gender and investigate whether the gender variable also determines the heterogeneous endogeneity in our sample. As shown in the right panel of Table 7, the coefficients of the interaction terms between control function and both black male and female look statistically insignificant. In contrast, the same coefficients of the interaction terms for white male and female are significant at 5% level. Thus, in our sample, we do not find evidence that the gender variable determines the endogenous and exogenous groups. Furthermore, we also find that compared to the results of the baseline approach, our method provides noticeable efficiency gains while the magnitude of the treatment effects for both male and female are similar.

5 Conclusion

In this paper, we have shown a new way to improve the precision of IV estimates and increase the power of endogeneity tests, namely by testing for and accounting for heterogeneous endogeneity when it exists. If we can identify sub-groups in the data where a potentially endogenous regressor is actually exogenous, we can reduce information loss and achieve a more precise estimate. We show this is true both asymptotically and in finite samples. The weaker is the instrumental variable being used, and the larger is the size of the non-endogenous group, the greater is the gain from modeling heterogeneous endogeneity when it exists. In our simulation exercises, efficiency gains of 50% or greater are not uncommon. However, even when the instrument is extremely strong and the non-endogenous group is relatively small, we still obtain notable efficiency gains. In our empirical example, we find around a 15% variance reduction even in this "worse case" scenario for the proposed approach.

Acknowledgements We are thankful to Christoph Rothe, Richard Startz, Lutz Kilian, Le Wang for their helpful comments and suggestions. Any errors or omissions are our own.

Appendix A

We first show the assumptions required to derive theoretical results in the paper.

Assumption 1 The data are generated as $y_{ij} = x'_{ij}\beta + u_{ij}$, $x_{ij} = Z'_{ij}\delta + \Sigma_\nu \nu^*_{ij}$ with the covariance matrix of the error terms in Eq. (3).

Assumption 2 The following limits hold for $j = 1, 2, \dots, J$, when the sample size converges to infinity:

a. (exogeneity) $Z'_j u_j / n_j \xrightarrow{p} 0$
b. (well-behaved data) $Z'_j Z_j / n_j \xrightarrow{p} Q_{Z_j Z_j}$ where $Q_{Z_j Z_j}$ is a finite, positive definite $L \times L$ matrix.
c. (relevance) $Z'_j x_j / n_j \xrightarrow{p} Q_{Z_j X_j}$ where $Q_{Z_j X_j}$ is a finite $L \times K$ matrix with rank K.

Assumption 3 As $N \to \infty$, the total number of observations in the jth group, N_j converges to infinity such that $lim_{N\to\infty} \frac{N_j}{N} = a_j$ for $j = 1, 2, \dots, J$.

Notice that the convergences in Assumption 2 hold under the weak law of large numbers and the asymptotic moments are assumed to be same across different groups under Assumption 1.

Proof of Lemma 1 Because $\hat{V}$ is a diagonal matrix, $M_{\hat{V}}$ is also a diagonal matrix of the following form:

$$M_{\hat{V}} = diag[M_{\hat{V}_1}, M_{\hat{V}_2}, \ldots, M_{\hat{V}_J}]$$

where $M_{\hat{V}_j} = I_{N_j} - \hat{V}_j(\hat{V}_j'\hat{V}_j)^{-1}\hat{V}_j'$. Therefore, we have:

$$\begin{aligned} \hat{\beta}_{CF} &= \left(X'M_{\hat{V}}X\right)^{-1}\left(X'M_{\hat{V}}Y\right) \\ &= \left(\sum_{j=1}^{J} x_j' M_{\hat{V}_j} x_j\right)^{-1}\left(\sum_{j=1}^{J} x_j' M_{\hat{V}_j} y_j\right) \\ &= \left(\sum_{j=1}^{J} x_j' P_{Z_j} x_j\right)^{-1}\left(\sum_{j=1}^{J} x_j' P_{Z_j} y_j\right) \\ &= \left(\sum_{j=1}^{J} \frac{N_j}{N}\frac{1}{N_j} x_j' P_{Z_j} x_j\right)^{-1}\left(\sum_{j=1}^{J} \frac{N_j}{N}\frac{1}{N_j} x_j' P_{Z_j} y_j\right) \\ &= \beta + \left(\frac{1}{N}\sum_{j=1}^{J} x_j' P_{Z_j} x_j\right)^{-1}\left(\frac{1}{N}\sum_{j=1}^{J} x_j' P_{Z_j} E_j\right) \end{aligned} \tag{A.1}$$

where $P_{Z_j} = Z_j(Z_j'Z_j)^{-1}Z_j'$;$Z_j = [z_{1j}\ z_{2j} \ldots\ z_{N_j j}]'$; $E_j = [\epsilon_{1j}\ \epsilon_{2j}\ \ldots\ \epsilon_{N_j j}]'$. The validity of going from the third line to the fourth line in equation (A.1) is that $M_{\hat{V}_j} x_j = P_{Z_j} x_j$. Since $N_j \to \infty$ as $N \to \infty$ and $a_j = lim_{N\to\infty}\frac{N_j}{N}$,

$$\sqrt{N}\left(\widehat{\beta}_{CF}^{*} - \beta\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma_u^2\Big(\sum_{j=1}^{J} a_j Q_{Z_jX_j}' Q_{Z_jZ_j}^{-1} Q_{Z_jX_j}\Big)^{-1}\right)$$

by the law of large numbers and the central limit theorem under the Assumption 1. □

Proof of proposition 1 As shown in Eq. (10),

$$\widehat{\beta}_{CF}^{*} = \left(X'M_{\tilde{V}}X\right)^{-1}\left(X'M_{\tilde{V}}Y\right)$$

where the residual maker matrix $M_{\tilde{V}} = I - P_{\tilde{V}}$ and $P_{\tilde{V}}$ is the usual projection matrix. We denote $\tilde{V} = \begin{bmatrix}\hat{\nu}_e \\ 0\end{bmatrix}$. Thus we get

$$P_{\tilde{V}} = \tilde{V}(\tilde{V}'\tilde{V})^{-1}\tilde{V}' = \begin{bmatrix}\hat{\nu}_e \\ 0\end{bmatrix}\left(\begin{bmatrix}\hat{\nu}_e & 0\end{bmatrix}\begin{bmatrix}\hat{\nu}_e \\ 0\end{bmatrix}\right)^{-1}\begin{bmatrix}\hat{\nu}_e & 0\end{bmatrix} = \begin{bmatrix}P_{\hat{\nu}_e} & 0 \\ 0 & 0\end{bmatrix}$$

where $P_{\hat{\nu}_e} = \hat{\nu}_e(\hat{\nu}_e'\hat{\nu}_e)^{-1}\hat{\nu}_e'$. Therefore,

$$M_{\tilde{V}} = I_N - P_{\tilde{V}} = \begin{bmatrix} M_{\hat{\nu}_e} & 0 \\ 0 & I \end{bmatrix}$$

where $M_{\hat{\nu}_e} = I - P_{\hat{\nu}_e}$. Now, we use the above $M_{\hat{\nu}_e}$ matrix to obtain the expressions $X'M_{\hat{\nu}_e}X$ and $X'M_{\hat{\nu}_e}Y$:

$$X'M_{\tilde{V}}X = \begin{bmatrix} x_e' & x_{ne}' \end{bmatrix} \begin{bmatrix} M_{\hat{\nu}_e} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x_e \\ x_b \end{bmatrix} = x_e'M_{\hat{\nu}_e}x_e + x_{ne}'x_{ne}$$

Similarly for $\tilde{X}'M_{\tilde{V}}\tilde{Y}$, we get

$$X'M_{\tilde{V}}Y = \begin{bmatrix} x_e' & x_{ne}' \end{bmatrix} \begin{bmatrix} M_{\hat{\nu}_e} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} y_e \\ y_{ne} \end{bmatrix} = x_e'M_{\hat{\nu}_e}y_e + x_{ne}'y_{ne}$$

From the standard OLS estimator we get, $x_{ne}'y_{ne} = (x_{ne}'x_{ne})\widehat{\beta}_{ne,OLS}$. For group e, we get $\hat{\beta}_{e,CF} = (x_e'M_{\hat{\nu}_e}x_e)^{-1}x_e'M_{\hat{\nu}_e}y_e$. As $M_{\hat{\nu}_e} = P_{Z_e}$, $\hat{\beta}_{e,CF} = \hat{\beta}_{e,2SLS}$. Therefore, we have $x_e'M_{\hat{\nu}_e}y_e = (x_e'M_{\hat{\nu}_e}x_e)\widehat{\beta}_{2SLS}$. Substituting $x_e'M_{\hat{\nu}_e}y_e$ and $x_{ne}'y_{ne}$ in $X'M_{\tilde{V}}Y$, we get

$$X'M_{\tilde{V}}Y = (x_e'M_{\hat{\nu}_e}x_e)\widehat{\beta}_{e,2SLS} + (x_{ne}'x_{ne})\widehat{\beta}_{ne,OLS}$$

Hence we get,

$$\begin{aligned}
\widehat{\beta}^*_{CF} &= (X'\tilde{V}X)^{-1}X'\tilde{V}Y \\
&= \left(x_e'M_{\hat{\nu}_e}x_e + x_{ne}'x_{ne}\right)^{-1}\left((x_e'M_{\hat{\nu}_e}x_e)\widehat{\beta}_{e,2SLS} + (x_{ne}'x_{ne})\widehat{\beta}_{ne,OLS}\right) \\
&= W\widehat{\beta}_{e,2SLS} + (I - W)\widehat{\beta}_{ne,OLS}
\end{aligned}$$

where

$$W = \left(x_e'M_{\hat{\nu}_e}x_e + x_{ne}'x_{ne}\right)^{-1}\left(x_e'M_{\hat{\nu}_e}x_e\right)$$

Under assumption 1, it is straightforward to show that $\widehat{Avar}[\hat{\beta}_{e,2SLS}] = \hat{\sigma}_u^2 (x_e'M_{\hat{\nu}_e}x_e)^{-1} = \hat{\sigma}_u^2(x_e'P_{Z_e}x_e)^{-1}$ and $\widehat{Avar}[\hat{\beta}_{ne,OLS}] = \sigma_u^2(x_{ne}'x_{ne})^{-1}$. Therefore,

$$W = \left(\widehat{Avar}[\widehat{\beta}_{e,2SLS}]^{-1} + \widehat{Avar}[\widehat{\beta}_{ne,OLS}]^{-1}\right)^{-1}\left(\widehat{Avar}[\widehat{\beta}_{e,2SLS}]^{-1}\right)$$

and

$$I - W = \left(\widehat{Avar}[\widehat{\beta}_{e,2SLS}]^{-1} + \widehat{Avar}[\widehat{\beta}_{ne,OLS}]^{-1})^{-1}(\widehat{Avar}[\widehat{\beta}_{ne,OLS}]^{-1}\right).$$

Proof of Lemma 2 Similar to Lemma 1,

$$
\begin{aligned}
\widehat{\beta}_{CF}^{*} &= (X' M_{\tilde{V}} X)^{-1} (X' M_{\tilde{V}} Y) \\
&= \beta + (x_e' P_{Z_e} x_e + x_{ne}' x_{ne})^{-1} (x_e' P_{Z_e} u_e + x_{ne}' u_{ne}) \\
&= \beta + \left(\frac{N_e}{N} \frac{1}{N_e} x_e' P_{Z_e} x_e + \left(1 - \frac{N_e}{N}\right) \frac{1}{N_{ne}} x_{ne}' x_{ne} \right)^{-1} \\
&\times \left(\frac{N_e}{N} \frac{1}{N_e} x_e' P_{Z_e} u_e + \left(1 - \frac{N_e}{N}\right) \frac{1}{N_{ne}} x_{ne}' u_{ne} \right)
\end{aligned}
$$

where $P_{Z_e} = Z_e (Z_e' Z_e)^{-1} Z_e'$ and $Z_e = [z_{1e}\ z_{2e} \dots z_{n_{ee}}]'$; N_e is the number of individuals in the group e. Since $N_e \to \infty$ and $N_{ne} \to \infty$ as $N \to \infty$,

$$
\sqrt{N}\left(\hat{\beta}_{CF}^{*} - \beta\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma_u^2 \left(w\ Q_{Z_e X_e}' Q_{Z_e Z_e}^{-1} Q_{Z_e X_e} + (1 - w) Q_{X_{ne} X_{ne}}\right)^{-1}\right)
$$

by the law of large numbers and the central limit theorem under assumption 1.

Proof of Proposition 2 The difference in the asymptotic variances of $\widehat{\beta}_{CF}^{*}$ and the conventional $2SLS$ estimator under homogeneous endogeneity are given by:

$$
\begin{aligned}
Avar\left[\widehat{\beta}_{2SLS}\right] - Avar\left[\widehat{\beta}_{CF}^{*}\right] &= \sigma_u^2 \frac{1}{N} plim \left(\left(\frac{1}{N} X' P_Z X \right)^{-1} \right) \\
&- \sigma_u^2 \frac{1}{N} plim \left(\left(\frac{1}{N} x_e' P_{Z_e} x_e + \frac{1}{N} x_{ne}' x_{ne} \right)^{-1} \right) \\
&= \sigma_u^2 \frac{1}{N} plim \left(\left(\frac{1}{N} X' P_Z X \right)^{-1} \right) \\
&- \sigma_u^2 \frac{1}{N} plim \left(\left(w \frac{1}{N_e} x_e' P_{Z_e} x_e + (1 - w) \frac{1}{N_{ne}} x_{ne}' x_{ne} \right)^{-1} \right) \\
&= \sigma_u^2 \frac{1}{N} plim \left(\left(\frac{1}{N} X' P_Z X \right)^{-1} \right) \\
&- \sigma_u^2 \frac{1}{N} plim \left(\left(w \frac{1}{N} X' P_Z X + (1 - w) \frac{1}{N} X' X \right)^{-1} \right) \\
&= \sigma_u^2 \frac{1}{N} plim \Big(N \big((X' P_Z X)^{-1} - (w\ X' P_Z X + (1 - w) X' X)^{-1} \big) \Big)
\end{aligned}
$$

To compare the two matrices in the brackets, we can compare their inverses:

$$
\begin{aligned}
&= w\ X'P_ZX + (1-w)X'X - X'P_ZX \\
&= w\ X'P_ZX + (1-w)X'X - wX'P_ZX - (1-w)X'P_ZX \\
&= (1-w)(X'X - X'P_ZX) \\
&= (1-w)(X'X + X'(M_Z - I)X) \\
&= (1-w)(X'M_ZX)
\end{aligned}
$$

Because M_Z is a positive definite matrix, it follows that $X'M_ZX$ is also. The matrix $(wX'P_ZX + (1-w)X'X)$ is larger in the matrix sense than $X'P_ZX$ and thus, its inverse is smaller. Therefore, $Avar[\widehat{\beta}_{2SLS}] - Avar[\widehat{\beta}^*_{CF}]$ is a positive definite matrix.

Proof of Lemma 3 Using the residual matrix estimated from the first stage regression, we obtain the following OLS estimator:

$$
\begin{aligned}
\hat{\delta} &= \left(\sum_{n=1}^{N} \hat{x}_n^* \hat{x}_n^{*'}\right)^{-1} \left(\sum_{n=1}^{N} \hat{x}_n^* y_n\right) \\
&= \left(\sum_{n=1}^{N} \hat{x}_n^* \hat{x}_n^{*'}\right)^{-1} \left\{\sum_{n=1}^{N} \hat{x}_n^* x_n^{*'} \delta + \sum_{n=1}^{N} \hat{x}_n^* \eta_n\right\} \\
&= \left(\sum_{n=1}^{N} \hat{x}_n^* \hat{x}_n^{*'}\right)^{-1} \left\{\sum_{n=1}^{N} \hat{x}_n^* \hat{x}_n^{*'} \delta + \sum_{n=1}^{N} \hat{x}_n^* m_n^{'} \delta + \sum_{n=1}^{N} \hat{x}_n^* \eta_n\right\} \\
&= \delta + \left(\frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^* \hat{x}_n^{*'}\right)^{-1} \left\{\frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^* m_n^{'} \delta + \frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^* \eta_n\right\}
\end{aligned}
$$

where $\hat{x}_n^* = [x_n' \ \ \hat{V}_n']'$; $\hat{V}_n = (D_n \otimes \hat{\nu}_n)$; $m_n = [\mathbf{0}' \ \ (V_n - \hat{V}_n)']'$. Note $x_n^* = \hat{x}_n^* + m_n$. The first term in the second bracket is reduced to:

$$
\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^* m_n^{'} \delta =& \frac{1}{N}\sum_{n=1}^{N} \begin{bmatrix} x_n \\ \hat{V}_t \end{bmatrix} \left[\mathbf{0}' \ (V_t - \hat{V}_t)'\right] \begin{bmatrix} \beta \\ \Gamma \end{bmatrix} \\
&= \frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^* \left(V_t - \hat{V}_t\right)' \Gamma \\
&= \frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^* \Gamma' \left(V_t - \hat{V}_t\right) \\
&= \left(\frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^* \Gamma' \mathbf{Z_n}\right) \left(\hat{\pi} - \pi\right)
\end{aligned}
$$

where $\mathbf{Z_n} = (D_n \otimes Z_n')$. As $\hat{\pi} \xrightarrow{p} \pi$ and $\hat{x}_n \xrightarrow{p} x_n$, we have:

$$\frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^* m_n^{'} \delta \xrightarrow{p} 0.$$

The second term in the second bracket also converges to zero as:

$$\frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^* \eta_n \xrightarrow{p} E\left[x_n^* \eta_n\right] = E\left[x_n^* E\left[\eta_n | x_n, Z_n\right]\right] = 0.$$

Because $\sum_{n=1}^{N} \hat{x}_n^* \hat{x}_n^{*'} \xrightarrow{p} E\left[x_n^* x_n^{*'}\right]$, $\hat{\delta} \xrightarrow{p} \delta$. □

Proof of Proposition 3 From Lemma 3 , we have:

$$\sqrt{N}\left(\hat{\delta} - \delta\right) = \left(\frac{1}{N}\sum_{n=1}^{N} \hat{x}_n^* \hat{x}_n^{*'}\right)^{-1} \left\{\frac{1}{\sqrt{N}}\sum_{n=1}^{N} \hat{x}_n^* m_n^{'} \delta + \frac{1}{\sqrt{N}}\sum_{n=1}^{N} \hat{x}_n^* \eta_n\right\}.$$

The first term in the second bracket is reduced to:

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N} \hat{x}_n^* m_n^{'} \delta = \left(\frac{1}{N}\sum_{n=1}^{N} \hat{x}_n \Gamma' \mathbf{Z_n}\right) \sqrt{N}(\hat{\pi} - \pi)$$

where $\mathbf{Z_n} = (D_n \otimes Z_n')$. Thus, we have:

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N} \hat{x}_n^* m_n^{'} \beta^* = E\left[\hat{x}_n \Gamma' \mathbf{Z_n}\right]\sqrt{N}(\hat{\pi} - \pi) + o_p(1).$$

Next, the second term in the second bracket is decomposed into two parts as:

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N} \hat{x}_n^* \eta_n = \frac{1}{\sqrt{N}}\sum_{n=1}^{N} x_n^* \eta_n - \frac{1}{\sqrt{N}}\sum_{n=1}^{N} m_n \eta_n.$$

And the second term in the above equation can be rewritten as:

$$\begin{aligned}
\frac{1}{\sqrt{N}}\sum_{n=1}^{N} m_n \eta_n &= \frac{1}{\sqrt{N}}\sum_{n=1}^{N} \begin{bmatrix} \mathbf{0} \\ (D_n \otimes Z_n'(\hat{\pi} - \pi))\eta_n \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{0} \\ (D_n \otimes \frac{1}{\sqrt{N}} \sum_{n=1}^{N} Z_n' \eta_n (\hat{\pi} - \pi)) \end{bmatrix}.
\end{aligned}$$

Therefore, we obtain the following result:

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N} Z_n'\eta_n(\hat{\pi}-\pi) = \left(\frac{1}{N}\sum_{n=1}^{N} Z_n'\eta_n\right)\sqrt{N}(\hat{\pi}-\pi) = o_p(1)$$

because $\sum_{n=1}^{N} Z_n'\eta_n \xrightarrow{p} E\left[Z_n'\eta_n\right] = 0$ and $\sqrt{N}(\hat{\pi}-\pi) = O_p(1)$. This leads to:

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N} \hat{x}_n^*\eta_n = \frac{1}{\sqrt{N}}\sum_{n=1}^{N} x_n^*\eta_n + o_p(1)$$

Using $\sum_{n=1}^{N} \hat{x}_n^*\hat{x}_n^{*'} \xrightarrow{p} E[x_n^* x_n^{*'}]$, we obtain:

$$\sqrt{N}(\hat{\delta}-\delta) = \left(E[x_n^* x_n^{*'}]\right)^{-1}\{H\sqrt{N}(\hat{\pi}-\pi) + \frac{1}{\sqrt{N}}\sum_{n=1}^{N} x_n^*\eta_n\} + o_p(1)$$

where $H = E\left[x_n\Gamma'\mathbf{Z_n}\right]$. Thus the asymptotic distribution of $\hat{\delta}$ is given by:

$$\sqrt{N}(\hat{\delta}-\delta) \overset{a}{\sim} N\left(\mathbf{0}, \left(E[x_n^* x_n^{*'}]\right)^{-1}\left(H\ M\ H' + Var[x_n^*\eta_n]\right)\left(\left(E[x_n^* x_n^{*'}]\right)^{-1}\right)'\right),$$

where $\sqrt{N}(\hat{\pi}-\pi) \xrightarrow{d} N(0, M)$ because $E[\eta_n|x_n, Z_n] = 0$. □

Appendix B

In Appendix B we show Monte-Carlo simulation results about the finite sample performance of the new method that combines the CF approach and the *kmeans clustering* algorithm, which is explained in Sect. 3.1. For the simulation, we employ the same data generating process and the same true values of the model parameters used in Sect. 2.3 except ρ_2 and μ^2. The true value of ρ_2 is set to be -0.9. (a strong endogeneity case) Sample sizes vary a case by a case. For the first simulation trials, we consider $\{N_1 = 10, N_2 = 10\}$ and $T = 1, 5, 10$, and 20. For the second simulation trials, we consider $\{N_1 = 100, N_2 = 100\}$ and $T = 1, 5, 10$, and 20. The ratio of the two population variances that determine the concentration parameter μ^2 is assumed $\frac{Var(\pi_1 z_{i,j})}{Var(v_{i,j})} = \frac{1}{10} = 0.1$. Thus, $\mu^2 = T(N_1 + N_2) \times 0.1$ is $\{2, 10, 20, 40\}$ for the first simulation set and is $\{20, 100, 200, 400\}$ for the second simulation set. We generate 10,000 data sets for each case. Then we apply 4 estimators, OLS, 2SLS, the CF approach based on the true group membership, and the CF approach based on the modified *kmeans clustering* algorithm. Note that the first and the second cases of the first simulation set are weak and marginally weak instrument cases in which standard 2SLS estimation does not perform well.

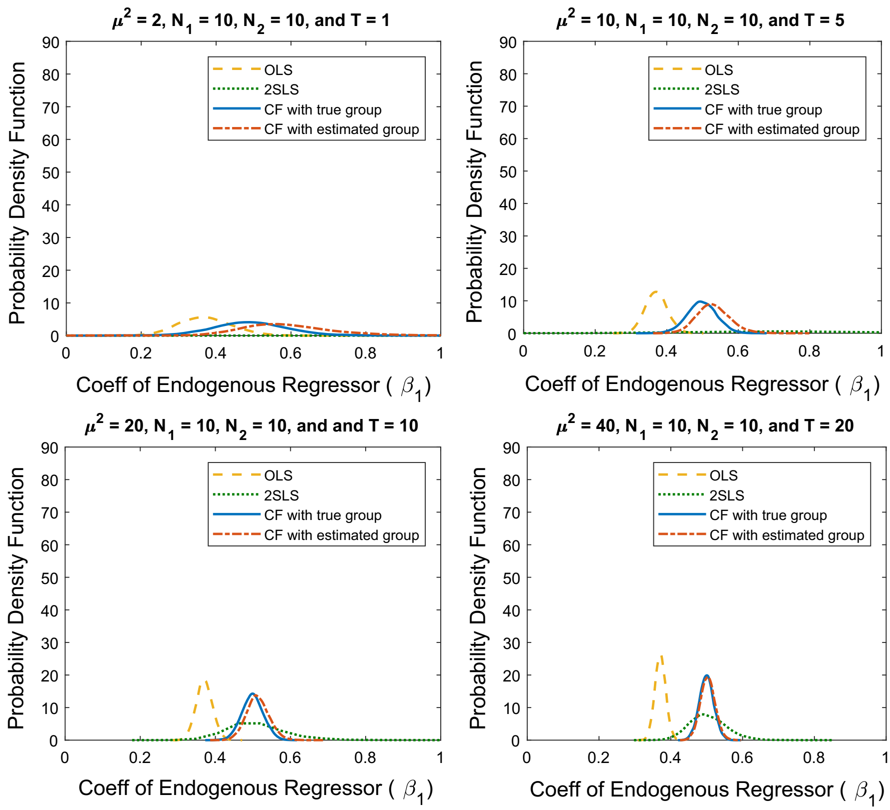

Fig. 3 Sampling distributions of β_1 from OLS, 2SLS, and two control function (CF) methods: a comparison between control function (CF) methods based on true and estimated groups

Figure 3 shows the results for the first simulation set that assumes $N_1 = 10$ and $N_2 = 10$. As well known in the literature, the sampling distribution of the 2SLS extremely disperses when $\mu^2 = 2$. This implies that 2SLS does not reveal any information for statistical inference. On the other hand, in the same case the sampling distributions of the two proposed CF estimators are close to the true value 0.5 even though there exists a slight bias in the CF estimator when the group membership is estimated by the modified *kmeans clustering* algorithm. Both proposed CF estimators perform substantially better because they extract information from the exogenous group in the data. In the second case ($T = 5$, $\mu^2 = 10$), the performance of the 2SLS estimator is not much better than the first case. However, the performance of the CF estimator that is based on the estimated groups does improve significantly. The magnitude of the bias of the CF estimator becomes almost negligible. In the third case ($T = 10$, $\mu^2 = 20$) and the fourth case ($T = 20$, $\mu^2 = 40$), the CF estimator that estimates the group membership produce almost identical results with those of the CF estimator that employs the true group membership.

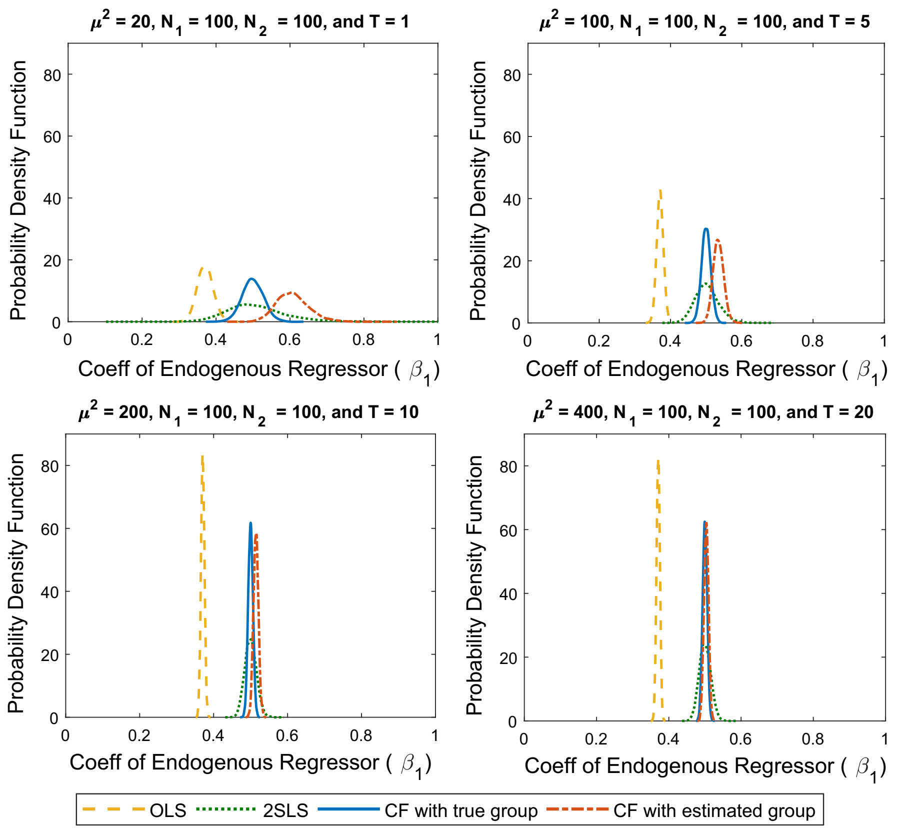

Fig. 4 Sampling distributions of β_1 from OLS, 2SLS, and two control function (CF) methods: a comparison between control function (CF) methods based on true and estimated groups

Figure 4 shows the results of the second simulation set that assumes $N_1 = 100$ and $N_2 = 100$. The important difference from the previous simulation set is that when $N = N_1 + N_2$ is large and T is small, the CF estimator that directly estimates the group membership has a larger bias. But once again the bias quickly converges to 0 when T increases. When T is larger than 10 in this case, the bias already becomes very close to 0.

References

Abadie A, Angrist J, Imbens G (2002) Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica 70:91–117

Abadie J (1991) Instrumental variables estimation of average treatment effects in econometrics and epidemiology, NBER Working Paper No. 115

Acemoglu D, Finkelstein A, Notowidigdo M (2013) Income and health spending: evidence from oil price shocks. Rev Econ Stat 95(4):1079–1095

Altonji J, Blank R (1999) Race and gender in the labor market. Handb Labor Econ 49:3143–3259

Angrist J, Imbens G, Rubin D (1996) Identification of causal effects using instrumental variables. J Am Stat Assoc 91:44–455

Angrist J, Pischke J (2009) Mostly harmless econometrics: an empiricists companion. Princeton University Press, Princeton, NJ

Autor D, Dorn D, Gordon H (2013) The china syndrome: local labor market effects of import competition in the united states. Am Econ Rev 103(6):2121–2168

Autor D, Dorn D, Gordon H, Majlesi K (2017) Importing political polarization? the electoral consequences of rising trade exposure, NBER Working Paper No. 22637

Barnow B, Cain G, Goldberger A (1981) Selection on observables. Eval Stud Rev Annu 5(1):43–59

Blundell R, Chen X, Christensen D (2007) Semi-nonparametric IV estimation of shape-invariant engel curves. Econometrica 75:1613–1669

Blundell R, Powell J (2003) Identification of unconditional partial effects in nonseparable models. In: Dewatripont M, Hansen L, Turnovsky S (eds) Advances in economics and econometrics, vol II. Cambridge University Press, Cambridge, pp 312–357

Bonhomme S, Manresa E (2015) Grouped patterns of heterogeneity in panel data. Econometrica 83(3):1147–1184

Bound J, Holzer H (1993) Industrial shifts, skill levels and the labor market for white and black males. Rev Econ Stat 75(3):387–396

Card D (1999) The causal effect of education on earnings. Handb Labor Econ 3:1801–1863

Card D (2001) Estimating the return to schooling: progress on some persistent econometric problems. Econometrica 69(5):1127–1160

Chesher A (2003) Identification in nonseparable models. Econometrica 71:1405–1441

Chesher A (2005) Nonparametric identidication under discrete variation. Econometrica 73:1525–1550

Darolles S, Florens J, Renault E (2003) Nonparametric estimation of triangular simultaneous equations models, Working Paper. Toulouse University

Florens J, Heckman J, Meghir C, Vytlacil E (2008) Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. Econometrica 76(5):1191–1206

Forgy E (1965) Cluster analysis of multivariate data: efficiency vs. interpretability of classification. Biometrics 21(3):768–769

Goldberger A, (1972) Selection bias in evaluating treatment effects: some formal illustrations., Discussion Paper. University of Wisconsin–Madison

Graddy K (1995) Testing for imperfect competition at the fulton fish market. RAND J Econ 26(1):75–92

Hall P, Horowitz J (2005) Nonparametric methods for inference in the presence of instrumental variables. Ann Stat 33:2904–2929

Hashimoto M (1987) The minimum wage law and youth crimes: time-series evidence. J Law Econ 30(2):443–464

Heckman J, Robb R (1985) Alternative methods for evaluating the impact of interventions: an overview. J Econom 30(1–2):239–267

Heckman J, Smith J, Clements N (1997) Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. Rev Econ Stud 64(4):487–535

Heckman J, Vytlacil E (1998) Instrumental variables methods for the correlated random coefficient model: estimating the average rate of return to schooling when the return is correlated with schooling. J Hum Resour 33(4):974–987

Heckman J, Vytlacil E (2005) Structural equations, treatment effects, and econometric policy evaluation. Econometrica 73(3):669–738

Heckman J, Vytlacil E (2007a) Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. Handb Econom 6:4779–4874

Heckman J, Vytlacil E (2007b) Econometric evaluation of social programs, part II: using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. Handb Econom 6:4875–5143

Holzer H (1987) Informal job search and black youth unemployment. Am Econ Rev 77(3):446–452

Imbens G, Newey W (2009) Identification and estimation of triangular simultaneous equations models without additivity. Econometrica 77:1481–1512

Imbens G, Wooldridge W (2009) Recent developments in the econometrics of program evaluation. J Econ Lit 47(1):5–86

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46:33–50

Lee S (2007) Endogeneity in quantile regression models: a control function approach. J Econom 141:1131–1158

Ma L, Koenker R (2006) Quantile regression methods for recursive structural equation models. J Econom 134:471–506
Nelson C, Startz R (1990a) The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. J Bus 63(1):S125–S140
Nelson C, Startz R (1990b) Some further results on the exact small sample properties of the instrumental variable estimator. Econometrica 58(4):967–976
Newey W, Powell J, Vella F (1999) Nonparametric estimation of triangular simultaneous equations models. Econometrica 67:563–603
Newey W, Powell J, Vella F (2003) Nonparametric estimation of triangular simultaneous equations models. Econometrica 71:1565–1578
Pinkse J (2000) Nonparametric two-step regression functions when regressors and error are dependent. Can J Stat 28:289–300
Rivers D, Vuong Q (1988) Limited information estimators and exogeneity tests for simultaneous probit models. J Econom 39(3):347–366
Smith J (1993) Affirmative action and racial wage gap. Am Econ Rev 83(2):79–84
Smith R, Blundell R (1986) An exogeneity test for a simultaneous equation tobit model with an application to labor supply. Econometrica 54(3):679–685
Stock J, Wright J, Yogo M (2002) A survey of weak instruments and weak identification in generalized method of moments. J Bus Econ Stat 20(4):518–529
Vytlacil E, Yildiz N (2007) Dummy endogenous variables in weakly separable models. Econometrica 75:757–779
Wooldridge J (2015) Control function methods in applied econometrics. J Hum Resour 50(2):420–445