ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
SCHOOL OF ECONOMIC SCIENCES
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΗΣ ΕΠΙΣΤΗΜΗΣ
DEPARTMENT OF ECONOMICS

Department of Economics

Athens University of Economics and Business

WORKING PAPER no. 08-2023

Dealing with endogenous regressors using copulas; on the problem of near multicollinearity

Dimitris Christopoulos, Dimitris Smyrnakis and Elias Tzavalis

May 2023

The Working Papers in this series circulate mainly for early presentation and discussion, as well as for the information of the Academic Community and all interested in our current research activity.

The authors assume full responsibility for the accuracy of their paper as well as for the opinions expressed therein.

Dealing with endogenous regressors using copulas; on the problem of near multicollinearity

Dimitris Christopoulos[*], Dimitris Smyrnakis[†] and Elias Tzavalis[‡]

April 30, 2023

Abstract

In this paper, in order to cope with the problem of endogenous regressors in cases that the linear regression model is non-identifiable, we suggest estimators handling the problem of multicollinearity to improve the performance of the Gaussian copula approach. This problem occurs when the endogenous regressor is nearly normally distributed and, thus, is highly correlated with its copula transformation term of the augmented regression controlling for the endogeneity problem. Based on a Monte Carlo study, we show that maximum entropy estimators can offer a solution to the problem. These estimators are found to outperform the ridge estimator, often used in practice to tackle the multicollinearity problem, and to conduct correct inference for the slope coefficients of the augmented regression.

Keywords: Copula, Endogeneity, Multicollinearity, Maximum Entropy, Ridge LS

JEL classification: C01, C12, C13, C18

[*]Department of International and European Economic Studies, Athens University of Economics and Business, email: dchristop@aueb.gr

[†]Department of Economics, Athens University of Economics and Business, email: smyrnakisd@aueb.gr

[‡]Department of Economics, Athens University of Economics and Business (Corresponding author), email: e.tzavalis@aueb.gr, tel.: (+30) 210 8203 332

We would like to thank Yiannis Dendramis and Yiannis Karavias for their comments and suggestions. Also, we thank participants at the International Conference on Refined Econometrics and Endogeneity, held at Athens University of Economics and Business (April 2023).

We acknowledge financial support by the Hellenic Foundation for Research and Innovation under the "First call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (project no: HFRI-FM17-3532).

1 Introduction

There is recently growing interest to develop econometric techniques to deal with the problem of endogenous regressors, namely the contemporaneous correlation between the regressors and the regression error term, based on methods free of instruments. The instrumental variables methods may suffer, significantly, in cases where there is no availability of valid (uncorrelated to the error term) instruments and/or the available instruments are weak, i.e., uncorrelated to the regressors (see, e.g., Hahn and Hausman (2005) and Andrews et al. (2019)). Instrument-free methods to deal with the endogeneity problem include the latent instrumental variables (LIV) approach proposed by Ebbes et al. (2005), identification through heteroscedasticity methods (e.g. Rigobon (2003), Klein and Vella (2010) and Lewbel (2012)), higher moments approaches (e.g. Cragg (1997), Dagenais and Dagenais (1997), Lewbel (1997) and Erickson and Whited (2002)), wavelet analysis (see, e.g., Gençay and Gradojevic (2011)) and granular IV (see Gabaix and Koijen (2022)). All of these methods rely heavily on distributional assumptions of the underlying variables and/or rely on a decomposition of the endogenous regressor into an exogenous part and an endogenous part, which may be hardly justified.[1]

Recently, Park and Gupta (2012) - henceforth referred to as PG - in a seminal paper suggested a Gaussian copula approach to cope with the endogenous regressor problem. This method relies on copula theory to capture the contemporaneous correlation between the potentially endogenous regressor and the error term. The method is easy to apply and has the merit that it does not rely on a decomposition of the endogenous regressor into an endogeneous and an exogenous part, compared to other instrument-free methods. Two variants of the method have been suggested for its implementation, in practice.[2] The first, originally suggested by PG, is based on the maximum likelihood (ML) approach

[1] See Rutz and Watson (2019) and Eckert and Hohberger (2022) for a comparison of these methods.

[2] Christopoulos et al. (2021, 2023a) extend the method to non-linear (threshold and smooth transition) models. Yet, panel data applications of the method can be found in Haschka (2022) and Christopoulos et al. (2022).

and employs the Gaussian copula to capture the correlation between the endogenous regressor and the error term. The second method augments the original regression model with a new regression term to control for endogeneity and employs a least squares (LS) method for estimation. The added term (referred to as copula control function CCF) is a copula transformation of the endogenous regressor, obtained by the inverse of normal cumulative distribution function of the empirical distribution of the regressor. It corrects the conditional mean of the model for the endogeneity bias. A key advantage of the second method is that it can be easily extended to the case of multiple regression, allowing for more than one endogenous regressors (see, e.g., Christopoulos et al. (2023b)). Yang et al. (2022) extends the method to the case that the regression model consists of an exogenous and an endogenous regressors which are correlated.

The PG copula approach is, however, not without limitations, as its ability to identify the true model parameters critically depends on the distinctiveness between the distributions of the endogenous regressor and the error term. This means that the endogenous regressor is required to have a non-normal distribution, assumed for the error term under the PG approach. If both the endogenous regressor and the error term are both normal (or approximately normal in distribution), then we will not be able to separate the variation of the endogenous regression from that of the CCF, and thus both the ML and LS estimates of the PG approach will suffer from identification problems. For the augmented with the CCF term regression model, this is the well known problem of multicollinearity implying biased LS estimates of parameters and large standard errors. This problem is examined thoroughly in the recent studies of Becker et al. (2022) and Eckert and Hohberger (2022), aiming to provide guidelines for the Gaussian copula approach's appropriate use. These studies show that, in order to work efficiently, the method requires sufficient (and not only significant) deviation from normality for the endogenous regressor. Popular non-normality tests, such as the Shapiro-Wilk test statistic, may not be able to identify the sufficient degree of non-normality required for

the method to work efficiently.

Given the unidentified concerns of the PG copula approach, in this paper we examine if we can improve the performance of the approach in cases that the endogenous regressor is nearly normally distributed by applying econometric methods to deal with the problem of near-multicollinearity that can potentially arise. We further show that ignoring this problem will lead to biased estimates of the slope coefficients of the model. More precisely, we investigate the performance of two well known estimators to cope with the problem of near multicollinearity of the augmented with the CCF endogenous regression; the ridge and maximum entropy estimators. The ridge estimator is obtained by minimizing a loss function (which is the traditional sum-of-squares) augmented with a penalty which warrants the identification of the regression coefficients (see, e.g., Hoerl and Kennard (1970b)). This penalty is known as the ridge penalty and it constraints the parameters space. However, the constraint imposed comes at the cost of introducing bias in the parameter estimates.

On the other hand, in the case of the entropy method the estimators are obtained by minimizing a loss function augmented by quantities which prevent the overflow of parameter estimates magnitude and their sample probabilities (see, e.g., Golan et al. (1996)). To define the magnitude of the parameter estimates and their probabilities, the two alternative estimators suggested by Paris (2004) do not rely on any a-priori information, but they use sample information. We will henceforth denote the first variant as MEE1. The alternative is extended to also consider the probabilities of the error term estimates. This variant of the estimator will be denoted as MEE2. Both of these variants of the entropy estimator are not sensitive to multicollinearity, as they are based on an estimation procedure for nonlinear programming problems with Karush-Kuhn Tucker (KKT) conditions. Compared to the ridge estimator, they have the nice properties that they are consistent and their performance does not depend on the choice of a constraint parameter value, like the ridge penalty value.

To evaluate the performance of the above estimators, we carry out a comprehensive simulation study which is focused on distributions of the endogenous regressor which are close to the normal distribution. In particular, we consider distributions, such as the Student's-t and Logistic, which are symmetric and close to the normal distribution. Our analysis can be obviously extended to other distributions. Our study provides a number of results which have useful implications for applied work. Firstly, we show that all the above estimators handling multicollinearity can considerably improve the performance of the PG method in cases that the distribution of the endogenous regressor is close to the normal. We find that they can reduce considerably the bias of the least square estimates and their root mean square error. These results hold not only for the single regression, but for a multiple extension of it also considering an exogenous regressor correlated to the endogenous one. Secondly, we find that the two variants of the maximum entropy estimator (MEE1 and MEE2) clearly outperform the ridge estimator. These entropy estimators are also found to perform well in conducting inference for the slope coefficients in the augmented linear regression with the copula transformed term based on a bootstrap procedure.

The paper is organized as follows. Section 2 presents the method, while Section 3 presents the results of the Monte Carlo study. Section 4 concludes the paper.

2 A brief overview of the Gaussian Copula method in handling endogenous regressors

For simplicity, consider the following simple linear regression model:

$$y_i = \beta_1 + \beta_2 x_i + u_i, \qquad i = 1, 2, ..., N \tag{1}$$

where $\{y_i, x_i\}_{i=1}^N$ constitute an observable sample of real valued continuous random variables, and u_i is the error term. For model (1), we initially make the following assumptions:

(a) $u_i \sim IID(0, \sigma_u^2)$,

(b) x_i is continuous and has a strictly monotonically increasing probability distribution with mean μ_x and $Var(x_i) = \sigma_x^2$, and

(c) $E(u_i|x_i) \neq 0$.

Assumptions (a)-(c) are standard in econometric textbooks concerning estimation and inference procedures for model (1) (see, e.g., Greene (2018)), meaning that the regressor x_i is endogenous. The sigma-field upon which the conditional expectation is defined is generated by x_i.

From Sklar's theorem (Sklar (1959)) it follows that there exists a unique copula function C, such that:

$$F(u_i, x_i) = C\left(F_u(u_i), F_x(x_i)\right), \tag{2}$$

where $C : [0,1]^2 \to [0,1]$, is a 2-dimension copula, $F_x(x_i)$ is the cumulative distribution function of x_i and $F_u(u_i)$ is the cumulative distribution function of u_i. Park and Gupta (2012) show that model (1) can be consistently estimated by maximizing the log-Likelihood of the probability density function corresponding to the joint cumulative distribution function of x_i and u_i, i.e. $F(u_i, x_i)$. When $u_i \sim N(0, \sigma_u^2)$ and C is the Gaussian copula, this likelihood takes the form

$$\begin{aligned}\ln L(\beta, \sigma_u, \rho) = &-\sum_{i=1}^{N}\left\{\frac{\rho^2\left[\Phi^{-1}\left(F_x(x_i)\right)^2 + \Phi^{-1}\left(F_u(u_i)\right)^2\right]}{2\left(1-\rho^2\right)} - \frac{\rho\Phi^{-1}\left(F_x(x_i)\right)\Phi^{-1}\left(F_u(u_i)\right)}{1-\rho^2}\right\} \\ &- \frac{N}{2}\ln(1-\rho^2) - \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_u^2) - \frac{1}{2\sigma_u^2}\sum_{i=1}^{N}u_i^2 - \sum_{i=1}^{N}\ln\left(f_x(x_i)\right)\end{aligned} \tag{3}$$

where $\beta = (\beta_1, \beta_2)'$ and $\rho = \rho(x_i, u_i)$. Note that the probability density function of x_i, denoted $f_x(x_i)$, is non-parametrically estimated and does not include any of the unknown parameters (β, σ_u, ρ), thus the term $-\sum_{i=1}^{N} \ln(f_x(x_i))$ can be excluded. The same holds true for the term $-\frac{N}{2}\ln(2\pi)$.[3]

Given assumptions (a)-(c) the conditional distribution of u_i on x_i can be derived based on the copula function C as follows:

$$F_{u|x}(u_i|x_i) = \frac{\partial}{\partial F_x} C\left(F_u(u_i), F_x(x_i)\right) \tag{4}$$

Based on relationship (4), it can be shown (see Appendix A) that u_i has the following single-factor correlation structure:

$$u_i = \lambda x_i^* + Var(u_i|x_i^*)^{1/2}\varepsilon_i, \text{ with } \lambda = \rho_{ux}\sigma_u. \tag{5}$$

The expectation of u_i conditional on x_i is given as

$$E(u_i|x_i) = \lambda x_i^*, \tag{6}$$

where x_i^* constitutes a transformation of the random variable x_i based on the quantile function (QF) of the distribution of u_i. When $u_i \sim N(0, \sigma_u^2)$, $x_i^* = \Phi^{-1}(F_x(x_i))$ is the quantile function of the standard normal distribution, where $\Phi^{-1} = \inf\{x_i \in \mathbb{R} : p \leq F_x(x_i)\}$, $p \in (0, 1)$. Expression (6) gives a linear relationship between u_i and copula based transformation of x_i, for the case that the error term u_i is Normally distributed, and C is the Gaussian copula, denoted as G-copula.

This result is due to the central result of the copula theory (i.e. Sklar's Theorem) decomposing the joint distribution of u_i and x_i into a part that captures the dependence structure between them through a copula and that describing the marginal distribution

[3] Alternatively, model (1) can be estimated using the conditional likelihood based on the conditional distribution of u_i on x_i.

of x_i, itself. For the Gaussian copula, this structure is linear, independently of the marginal distribution of x_i. The assumption that u_i follows the normal distribution is often used in econometrics. It is often consistent with the data after dummying out outliers and/or accounting for structural break shifts in the conditional mean of y_i on x_i, implied by regression model (1) (see Spanos (2018), for a recent survey).

Based on equation (6), we can employ the following extension of regression model (1) to control for the effects of the regressor endogeneity on the estimates of the vector of coefficients β:

$$y_i = \beta_1 + \beta_2 x_i + \lambda x_i^* + e_i, \qquad i = 1, 2, ..., N \tag{7}$$

where $e_i = u_i - \lambda x_i^*$ is a zero mean error term which is independent of x_i^*. In addition, it can be seen that $e_i = u_i - E(u_i|x_i)$, with $E(e_i|x_i) = 0$ and, hence, $E(e_i x_i|x_i) = 0$, i.e., e_i has the properties of the error of the conditional expectation decomposition of u_i on x_i. This follows from the fact that x_i^* constitutes a transformation of x_i, based on the quantile function Φ^{-1} of the normal distribution, which is independent of the error term e_i.

Model (7) adjusts the conditional mean of y_i on x_i, implied by the regression model (1), for the regressor endogeneity bias problem caused by the correlation between x_i and u_i. This is done by including in its right hand side (RHS) the copula transformed variable x_i^*, referred to as CCF (Copula Control Function). The model can be consistently estimated based on a two-step least squares procedure, provided values of the transformed variables x_i^* (see Joe (2014)). These values can be obtained, in a first step, using the transformation $x_i^* = \Phi^{-1}(F_x(x_i))$ based on a non-parametric estimation method of distribution $F_x(x_i)$ (see Silverman (1986)), or the Empirical Cumulative Distribution Function method, denoted ECDF (e.g., Rice (2007)). As N increases the distribution estimation error implied by the above methods becomes negligible, by the Glivenko - Cantelli theorem (see Cantelli (1933) and Glivenko (1933)).

Compared to the method of instrumental variables, the copula method suggested above has the interesting feature that it does not suffer from the problems of valid and/or weak instruments. Yet, with respect to the maximum likelihood full information approach, it does not require the specification of a simultaneous system of equations for y_i and x_i, and knowledge of the joint distribution of u_i and x_i to remove the bias of β's. Instead, the method constitutes a limited information approach which only requires a copula-based transformation of x_i.

The success of the above method to deal with the problem of regressor endogeneity, efficiently, depends on how well the linear correlation structure implied by (5) captures the underlying structure of the data and the distribution features of regressors x_i. For the identification of the true values of the slope coefficients of model (7), the method requires the distribution of x_i to distinctively differ from the normal (see discussion in the introduction). If, for instance, x_i is also normally distributed, then it can be easily seen that the variable x_i^*, given as $x_i^* = \Phi^{-1}\left(\Phi_{(\mu_x,\sigma_x)}(x_i)\right)$, reduces to $x_i^* = \frac{x_i-\mu_x}{\sigma_x}$, which constitutes a linear (location-scale) transformation of x_i. Using matrix notation we can write $x^* = \frac{1}{\sigma_x}x + \frac{\mu_x}{\sigma_x}\mathbf{1}$, which shows that vector x^* is a linear combination of vectors x and $\mathbf{1}$ which raises the problem of multicollinearity and identification of the slope coefficients of model (7). The fact that x_i^* is obtained with an estimation and approximation error of the true distribution $F_x(x_i)$ may mitigate this problem, but it does not eliminate it.

The problem of multicollinearity can also exist in cases where the endogenous regressor x_i is closely normally distributed.[4] Then, it is known as near multicollinearity problem and to cope with it we need to employ appropriate estimators. Two such well known estimators are the ridge LS and Maximum Entropy estimators.

The ridge LS estimator (Hoerl and Kennard (1970a,b)), denoted as RLS, minimizes

[4]In applied work, to appraise how important is the problem of near multicollinearity we can use a diagnostic, like the correlation coefficient between x_i and x_i^* and the determinant of the variance-covariance matrix of the dependent and independent variables of the model, including the transformed regressors, (see, e.g., Spanos and McGuirk (2002)), or Variance Inflation Factors and other more advanced diagnostics as suggested by Belsley et al. (1980) and Belsley (1991).

the ridge loss function, which is defined as

$$\mathcal{L}(\beta;\kappa) = \|y - Xb\|_2^2 + \kappa\|b\|_2^2 = \sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i - \lambda x_i^*)^2 + \kappa\left(\beta_1^2 + \beta_2^2 + \lambda^2\right) \quad (8)$$

where $b = (\beta_1, \beta_2, \lambda)'$ and κ denotes the penalty parameter. For $\kappa = 0$, the estimator reduces to the OLS estimator. Given a penalty parameter κ, the ridge estimator takes the form

$$b(\kappa) = \left(X'X + \kappa I\right)^{-1} X'y \quad (9)$$

where X is the matrix of the regressors' observations and y is the vector of the observations of the dependent variable. Singh et al. (1986) suggested that κ is calculated as

$$\kappa = \frac{P\hat{\sigma}^2}{\sum_{j=1}^{P} \frac{\hat{\alpha}_j^2}{1 + \sqrt{1 + \kappa_j\left(\hat{\alpha}_j^2/\hat{\sigma}^2\right)}}} \quad (10)$$

where $\hat{\alpha}_j$ and $\hat{\sigma}^2$ denote first stage OLS estimates of the standardized version of regression model (13), and $P = K - 1$ ($K = 3$ in the case of augmented model (7)).[5]

The alternative, Maximum Entropy estimators, are non-linear in nature and thus are unlikely to suffer from the problem of near multicollinearity (see Golan et al. (1996) for a discussion). Such estimators are the Maximum Entropy Leuven estimators (see Paris (2004)), denoted as MEE1 and MEE2. These estimators, contrary to other Entropy estimators, require no subjective or prior information to implement. MEE1 is defined such that it maximizes the following entropy measure

$$\mathcal{H}_1(p_b, L_b, e) = -p_b' \ln(p_b) - L_b \ln(L_b) - e'e, \quad (11)$$

[5]The penalty parameter proposed by Singh et al. (1986) showed consistent performance throughout our Monte Carlo simulation study. Other ridge estimators considered include Hoerl and Kennard (1970a), Hoerl et al. (1975), Thisted (1976), Lawless and Wang (1976), Dwivedi and Srivastava (1978), Khalaf and Shukur (2005), Khalaf (2012, 2013) and the Generalized Cross Validation approach by Golub et al. (1979), yet they performed poorly compared to the Singh et al. (1986) method.

where $L_b = b'b$, $p_b = b \odot b/L_b$, $b = (\beta_1, \beta_2, \lambda)'$, $X^* = (\mathbf{1}, x, x^*)$, $\odot$ denotes the Hadamard element by element product, and $p_{b_j} \geq 0 \ \forall j = 1, 2, ..., K$. The quantities $p_b' \ln(p_b)$ and $L_b \ln(L_b)$ added in the RHS of (11) prevent the overflow of parameter estimates magnitude and their sample probabilities based on sample information.

MEE2 extends MEE1 to include also terms which prevent overflow of the estimates of the error term and their probabilities. That is

$$\mathcal{H}_2(p_b, L_b, p_e, L_e) = -p_b' \ln(p_b) - L_b \ln(L_b) - p_e' \ln(p_e) - L_e \ln(L_e), \tag{12}$$

where $L_b = b'b$, $L_e = e'e$, $p_b = b \odot b/L_b$, $p_e = e \odot e/L_e$, $p_{b_j} \geq 0 \ \forall j = 1, 2, ..., K$, and $p_{e_i} \geq 0 \ \forall i = 1, 2, ..., N$.

3 Monte Carlo simulation

In this section, we conduct a Monte Carlo (MC) study aiming to evaluate the performance of the PG method to deal with the endogenous regressor problem in cases that the regressor has a distribution which is close to normal and thus, model identification problems arise. More specifically, our analysis has two tasks: Firstly, to investigate how severe the identification problem is by examining the biases of the slope coefficient estimates and, secondly, to examine if the ridge and maximum entropy estimators suggested in the previous section can alleviate the problem. In our analysis, we consider elliptical distributions, like the Student's-t distribution with six and nine degrees of freedom and the logistic distribution, which are close to the normal distribution. These distributions exhibit zero skewness for which the PG approach is expected to perform poorly in identifying the parameters of model (7). As shown by Becker et al. (2022), prerequisite of the PG method for the model identification is a non normal distribution of the endogenous regressor, with high levels of kurtosis and skewness.

We also consider the case of the normal distribution itself, also considered by Park

and Gupta (2012). In this case, the LS estimator of (7) will exhibit the strongest degree of multicollinearity. The identification of the model can be only achieved due to the sample estimation error of the copula regression term x^* in retrieving the ECDF of the endogenous regressor. Asymptotically, the identification of the model is infeasible, since this estimation error goes to zero. Since this case has only theoretical interest, we will present its main results in Appendix B.

We present results for the single regression case and the multiple regression case, with two regressors. For the last case, we consider the extension of the PG method suggested by Yang et al. (2022) who consider one endogenous regressor and one exogenous which is correlated with the endogenous. The presence of the exogenous regressor helps the identification of the model, if this regressor is not normally distributed. However, as aptly shown by Yang et al. (2022), its correlation with the endogenous regressor requires augmentation of the regression model not only with the copula transformation of the endogenous variable, but also with the copula transformation of the exogenous to capture the correlation across the two regressors on the slope coefficient estimates of the multiple regression. As in the single regression case, if both the endogenous and exogenous regressors are nearly normally distributed this will lead to identification problems of the model, too.

3.1 Single regression MC results

For the single linear regression case, we assume the following data generating mechanism:

$$y_i = \beta_1 + \beta_2 x_i + u_i \tag{13}$$

and assume that $\beta_1 = 1.0$ and $\beta_2 = 1.0$.[6] We also assume that $u_i \sim N(0,1)$, while for x_i, we consider (a) the normal distribution $N(\mu_x, \sigma_x)$ with $\mu_x = 0$ and $\sigma_x = 1$, (b) the Student's-t distribution with $\nu = 6$ and $\nu = 9$ degrees of freedom (i.e. a $t(6)$ and a $t(9)$), and (c) a Logistic distribution with location parameter $\mu = 0$ and scale $\sigma = 1$.

For all cases (a)-(c), we consider the following data generating process.

$$\begin{pmatrix} \tilde{x}_i \\ u_i \end{pmatrix} \sim N\left(\begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix} \right)$$

Then, we obtain the series x_i as transformation of $\tilde{x}_i$, as follows:

$$x_i = QF(\Phi(\tilde{x}_i)), \tag{14}$$

where QF denotes the corresponding quantile function of the distributions that we would like to generate x_i, namely the Normal, Student's-t (Φ^{-1}, T_6^{-1} and T_9^{-1}) and the Logistic.

We carry out 1000 iterations using samples of $T = \{50, 100, 250, 500\}$ observations. In each iteration, we generate data from model (13), under the different simulation scenarios considered, and we estimate the augmented with regressor x_i^* version of the model based on the two-step LS method suggested in the section 2. To evaluate the performance of the method, we present average values of the bias of the estimates of β_1 and β_2 from their true values (denoted as BIAS), over all iterations, and their root mean squared errors (RMSE). Finally, for the non-parametric estimation of the marginal distribution $F_x(x_i)$, employed in the first step of the estimation procedure of the method, we consider the ECDF.[7]

[6] We have also considered other pairs of values β_1 and β_2, such that they deviate substantially from one another (e.g., $\beta_1 = 3.0$ and $\beta_2 = 1.0$). The results suggest that the MEE1 estimator for very small samples ($N = 50$) exhibits a leftover bias regarding the intercept (β_1) roughly over 6.5%. The bias, however, reduces rapidly with N. MEE2 and the RLS show no such issue.

[7] We have also considered a smooth Kernel density estimator, using the Epanechnikov kernel and Silverman's bandwidth, however it performed poorly compared to the ECDF. The corresponding results are available upon request.

In our MC study, we also evaluate the size and power of the standard t-ratio tests regarding the hypotheses $\beta_1 = 1$, $\beta_2 = 1$ and $\lambda = 0$. Data generated under the null hypothesis follow the original data generating mechanism described above. To generate data for the alternative hypotheses we consider (i) the same data generating process and we alter the values of β_1 to 0.8 and β_2 to 0.8 to compute the power for the corresponding tests, and (ii) the original data generating process described above setting the correlation $\rho(x_i, u_i)$ equal to 0. For all tests we compute bootstrap standard errors, primarily because the augmented regression includes an estimated regressor, and also because MEE1 and MEE2 estimators are numerically evaluated and do not produce a closed form formula for the covariance matrix. We calculate the bootstrap covariance matrix as follows:

1. We estimate the models under the null (or alternative) hypothesis and save the residuals produced under the null hypothesis, $\hat{e}_i$.

2. For each bootstrap iteration $j = 1, 2, ..., B$, we draw a random sample with replacement from the distribution of $\hat{e}_i$ and compute $\hat{e}_i^{*(j)} = f(\hat{e}_i)$, where

$$f(\hat{e}_i) = \left(\frac{N}{N-K}\right)^{1/2} \hat{e}_i \tag{15}$$

   (see Davidson and MacKinnon (2006) for a discussion).

3. Based on sample values of the vector of regressors x_i and x_i^*, the estimates, denoted as $\hat{b}$ under H_0 and the bootstrap samples $\hat{e}_i^{*(j)}$ for all $i = 1, 2, ..., N$, we calculate the associated bootstrap samples for the dependent variable, denoted by $y_i^{(j)}$. Then, we estimate the regression under the null (or alternative) hypothesis and calculate new bootstrap estimates for vector $b = (\beta_1, \beta_2, \lambda)'$.

4. Finally, we calculate the bootstrap Covariance matrix

$$Cov_B(\hat{b}) = \frac{1}{B-1}\sum_{j=1}^{B}\left(\hat{b}^{(j)} - \bar{\hat{b}}\right)\left(\hat{b}^{(j)} - \bar{\hat{b}}\right)', \tag{16}$$

where $\bar{\hat{b}} = \frac{1}{B}\sum_{j=1}^{B}\hat{b}^{(j)}$.

The results of our MC experiments are reported in Tables 1, 2, 3A-3C and 4A-4C. More specifically, Table 1 presents results of the correlation coefficients between the endogenous regressor and its copula transformation to see how severe is the problem of multicollinearity. This is done across all the distributions considered (see (a)-(c)), namely the Student's-t distribution with six and nine degrees of freedom, respectively and the Logistic distribution, as well as, the normal distribution. In this table, we also present estimates of the correlation coefficient between the ECDF estimation error, given as $\hat{v} = x^* - \hat{x}^*$, and the endogenous regressor and its copula transformation to see if this error can cause any significant contemporaneous correlation (endogeneity) regression problem. Yet, we present estimates of the mean and standard deviation of $\hat{v}$ to see if it is important.

Table 2 presents the rejection frequencies of tests statistics for the normality assumption of regressor x of the PG method. A similar exercise is conducted by Becker et al. (2022). Tables 3A-3C present results of the bias and root mean square (RMSE) error of the following alternative estimators employed, across the distributions (see (a)-(c)) considered: the ML estimator (based on equation (3)), LS estimators ignoring the endogenous regressor problem, the LS estimator of the augmented regression with the copula term, and the three estimators employed to deal with the problem of multicollinearity of the augmented regression, namely MEE1, MEE2 and RLS. Finally, Tables 4A-4C present results of the size and power of the t-ratio test statistics to conduct inference on the vector of coefficients $b = (\beta_1, \beta_2, \lambda)'$.

Table 1: Estimation error - Summary statistics

	$N = 50$	$N = 100$	$N = 250$	$N = 500$	$N = 1000$
			$x_i \sim N(0,1)$		
$\rho(x, \hat{x}^*)$	0.98855	0.99357	0.99696	0.99844	0.99914
$\rho(x, \hat{v})$	0.00086	0.00057	-0.00015	-0.00006	-0.00006
$\rho(x^*, \hat{v})$	0.00086	0.00057	-0.00015	-0.00006	-0.00006
st.dev.$(\hat{v})$	0.17767	0.13039	0.08761	0.06286	0.04622
mean$(\hat{v})$	0.00167	0.00075	0.00051	0.00106	-0.00047
			$x_i \sim t(6)$		
$\rho(x, \hat{x}^*)$	0.97348	0.97967	0.98314	0.98566	0.98712
$\rho(x, \hat{v})$	0.00036	0.00037	0.00053	0.0004	0.00067
$\rho(x^*, \hat{v})$	0.00033	0.00033	0.00045	0.00038	0.00065
st.dev.$(\hat{v})$	0.17628	0.13	0.08689	0.06366	0.04621
mean$(\hat{v})$	-0.0023	-0.00005	0.0016	0.00046	-0.00011
			$x_i \sim t(9)$		
$\rho(x, \hat{x}^*)$	0.98106	0.98669	0.99096	0.99328	0.99441
$\rho(x, \hat{v})$	0.00071	-0.00048	0.00043	-0.00015	0.00052
$\rho(x^*, \hat{v})$	0.0007	-0.0005	0.00037	-0.00014	0.00053
st.dev.$(\hat{v})$	0.17754	0.13037	0.08721	0.06319	0.04565
mean$(\hat{v})$	0.00143	0.00047	0.00048	-0.00024	0.00096
			$x_i \sim Logistic(0,1)$		
$\rho(x, \hat{x}^*)$	0.97959	0.9855	0.99047	0.99258	0.99386
$\rho(x, \hat{v})$	0.00082	-0.00023	0.0005	0.00053	0.00039
$\rho(x^*, \hat{v})$	0.00081	-0.00026	0.00046	0.00051	0.00036
st.dev.$(\hat{v})$	0.17676	0.13058	0.08651	0.06272	0.04596
mean$(\hat{v})$	-0.00598	0.00256	-0.00343	0.00039	-0.00105

Notes: The table presents the Pearson correlation coefficients among the regressor x, the true and estimated CCF, denoted as x^* and $\hat{x}^*$ respectively, and the estimation error of the CCF, i.e. $\hat{v} = x^* - \hat{x}^*$. Summary statistics regarding the CCF estimation error $\hat{v}$ are also presented, namely the mean and standard deviation.

Table 2: Rejection frequencies for Normality tests

	Anderson-Darling	Jarque-Berra	Lilliefors	Shapiro-Wilk	Cramér-von Mises
			$x_i \sim t(6)$		
$N = 50$	24.0%	35.6%	17.2%	34.6%	4.5%
$N = 100$	34.7%	54.1%	21.6%	51.1%	6.1%
$N = 250$	68.6%	84.3%	46.5%	82.6%	10.8%
$N = 500$	88.9%	97.0%	72.8%	96.5%	27.9%
$N = 1000$	99.8%	100.0%	96.7%	100.0%	64.6%
			$x_i \sim t(9)$		
$N = 50$	12.3%	22.5%	9.8%	20.8%	5.1%
$N = 100$	17.8%	35.7%	11.8%	33.4%	5.1%
$N = 250$	36.2%	59.5%	20.4%	56.9%	6.4%
$N = 500$	59.2%	82.0%	34.8%	79.0%	9.1%
$N = 1000$	87.3%	97.2%	62.1%	96.1%	18.3%
			$x_i \sim Logistic(0, 1)$		
$N = 50$	15.1%	28.3%	11.0%	26.8%	4.5%
$N = 100$	25.5%	40.1%	15.2%	38.6%	6.2%
$N = 250$	45.4%	66.9%	28.6%	64.2%	7.5%
$N = 500$	77.2%	90.4%	53.8%	88.6%	13.0%
$N = 1000$	95.7%	99.0%	82.5%	98.5%	32.1%

Notes: The table presents rejection frequencies of popular tests statistics for the normality assumption of regressor x of the PG method. We consider the Anderson and Darling (1952), Jarque and Bera (1987), Shapiro and Wilk (1965)/Shapiro and Francia (1972), Lilliefors (1967) and Cramér-von Mises (Cramér (1928) and von Mises (1928)) tests. All tests are conducted at the 5% significance level.

Table 3A: Simulation Results $x_i \sim t(6)$

	OLS		CCF		MLE		MEE1		MEE2		RLS	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
$N = 50$												
$\beta_1 = 1$	0.008	0.118	0.011	0.16	0.009	0.133	-0.036	0.14	0.006	0.146	0.008	0.158
$\beta_2 = 1$	0.495	0.505	0.356	0.726	0.365	0.495	0.016	0.169	0.011	0.178	-0.037	0.314
$\sigma = 1$	-0.198	0.214	0.055	0.381	-0.121	0.152	0.009	0.123	0.019	0.126	0.065	0.196
$\rho_{ux} = 0.6$			-0.438	0.686	-0.382	0.594	-0.039	0.158	-0.031	0.168	-0.05	0.279
$N = 100$												
$\beta_1 = 1$	0.000	0.082	-0.006	0.104	-0.004	0.092	-0.023	0.1	-0.002	0.102	-0.003	0.109
$\beta_2 = 1$	0.496	0.501	0.28	0.553	0.313	0.428	0.036	0.173	0.033	0.18	-0.049	0.247
$\sigma = 1$	-0.2	0.208	-0.009	0.262	-0.115	0.142	-0.005	0.121	0.002	0.125	0.058	0.159
$\rho_{ux} = 0.6$			-0.344	0.566	-0.323	0.509	-0.047	0.161	-0.044	0.172	-0.012	0.215
$N = 250$												
$\beta_1 = 1$	0.001	0.052	0.003	0.062	0.003	0.058	-0.006	0.062	0.003	0.063	0.003	0.068
$\beta_2 = 1$	0.487	0.489	0.174	0.356	0.222	0.32	0.047	0.173	0.046	0.182	-0.071	0.194
$\sigma = 1$	-0.198	0.201	-0.045	0.174	-0.1	0.127	-0.017	0.115	-0.014	0.118	0.057	0.133
$\rho_{ux} = 0.6$			-0.218	0.391	-0.225	0.373	-0.061	0.166	-0.061	0.179	0.018	0.149
$N = 500$												
$\beta_1 = 1$	-0.001	0.037	0.000	0.044	0.000	0.042	-0.004	0.045	0.000	0.045	0.000	0.049
$\beta_2 = 1$	0.485	0.486	0.1	0.245	0.143	0.214	0.037	0.154	0.036	0.158	-0.09	0.164
$\sigma = 1$	-0.195	0.197	-0.036	0.148	-0.077	0.106	-0.013	0.108	-0.011	0.11	0.069	0.122
$\rho_{ux} = 0.6$			-0.125	0.25	-0.131	0.226	-0.049	0.139	-0.049	0.144	0.042	0.106
$N = 1000$												
$\beta_1 = 1$	0.001	0.025	0.001	0.031	0.001	0.03	-0.001	0.031	0.001	0.031	0.001	0.034
$\beta_2 = 1$	0.485	0.486	0.06	0.17	0.082	0.141	0.03	0.13	0.028	0.133	-0.092	0.144
$\sigma = 1$	-0.196	0.197	-0.027	0.116	-0.049	0.078	-0.012	0.094	-0.011	0.095	0.07	0.11
$\rho_{ux} = 0.6$			-0.069	0.156	-0.067	0.135	-0.036	0.112	-0.035	0.114	0.051	0.09

Notes: Simulation results for model (13), when $x_i \sim t(6)$. Bias and RMSE of the alternative estimators employed are presented; namely OLS, the copula control function approach (CCF), the maximum likelihood estimator (MLE) proposed by Park and Gupta (2012), the two maximum entropy estimators, denoted as MEE1 and MEE2, and the ridge LS estimator as suggested by Singh et al. (1986), denoted as RLS. Values equal to 0.000 correspond to values less than 5×10^{-4}.

Table 3B: Simulation Results $x_i \sim t(9)$

	OLS		CCF		MLE		MEE1		MEE2		RLS	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
$N = 50$												
$\beta_1 = 1$	-0.003	0.113	-0.003	0.158	-0.002	0.126	-0.043	0.135	-0.002	0.138	-0.002	0.146
$\beta_2 = 1$	0.532	0.542	0.42	0.855	0.412	0.554	-0.008	0.176	-0.013	0.182	-0.012	0.36
$\sigma = 1$	-0.204	0.218	0.087	0.435	-0.124	0.152	0.02	0.121	0.029	0.124	0.056	0.191
$\rho_{ux} = 0.6$			-0.489	0.739	-0.411	0.626	-0.019	0.144	-0.012	0.153	-0.073	0.322
$N = 100$												
$\beta_1 = 1$	0.000	0.08	0.001	0.108	0.000	0.093	-0.021	0.101	0.000	0.102	0.000	0.108
$\beta_2 = 1$	0.532	0.537	0.357	0.674	0.384	0.513	0.017	0.179	0.015	0.19	-0.022	0.289
$\sigma = 1$	-0.199	0.208	0.013	0.293	-0.114	0.139	0.01	0.118	0.015	0.123	0.043	0.157
$\rho_{ux} = 0.6$			-0.42	0.652	-0.396	0.589	-0.032	0.153	-0.032	0.168	-0.043	0.249
$N = 250$												
$\beta_1 = 1$	0.001	0.049	0.002	0.059	0.002	0.054	-0.007	0.059	0.002	0.06	0.002	0.064
$\beta_2 = 1$	0.529	0.531	0.273	0.503	0.313	0.425	0.035	0.17	0.035	0.181	-0.044	0.217
$\sigma = 1$	-0.2	0.203	-0.044	0.218	-0.111	0.135	-0.008	0.11	-0.006	0.114	0.043	0.132
$\rho_{ux} = 0.6$			-0.323	0.512	-0.317	0.48	-0.046	0.151	-0.047	0.164	-0.006	0.178
$N = 500$												
$\beta_1 = 1$	0.001	0.036	0.000	0.044	0.001	0.041	-0.003	0.045	0.001	0.045	0.001	0.049
$\beta_2 = 1$	0.527	0.528	0.206	0.404	0.238	0.343	0.043	0.18	0.044	0.189	-0.059	0.199
$\sigma = 1$	-0.198	0.2	-0.049	0.187	-0.094	0.123	-0.012	0.113	-0.01	0.116	0.051	0.125
$\rho_{ux} = 0.6$			-0.244	0.413	-0.238	0.383	-0.055	0.16	-0.058	0.173	0.01	0.16
$N = 1000$												
$\beta_1 = 1$	-0.001	0.025	0.000	0.03	0.000	0.029	-0.002	0.03	0.000	0.031	0.000	0.033
$\beta_2 = 1$	0.527	0.527	0.125	0.281	0.135	0.228	0.038	0.169	0.039	0.174	-0.082	0.165
$\sigma = 1$	-0.199	0.2	-0.045	0.156	-0.064	0.097	-0.013	0.108	-0.012	0.11	0.06	0.113
$\rho_{ux} = 0.6$			-0.146	0.27	-0.124	0.236	-0.048	0.142	-0.05	0.149	0.037	0.108

Notes: Simulation results for model (13), when $x_i \sim t(9)$. Bias and RMSE of the alternative estimators employed are presented; namely OLS, the copula control function approach (CCF), the maximum likelihood estimator (MLE) proposed by Park and Gupta (2012), the two maximum entropy estimators, denoted as MEE1 and MEE2, and the ridge LS estimator as suggested by Singh et al. (1986), denoted as RLS. Values equal to 0.000 correspond to values less than 5×10^{-4}.

Table 3C: Simulation Results $x_i \sim Logistic(0,1)$

	OLS		CCF		MLE		MEE1		MEE2		RLS	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
$N = 50$												
$\beta_1 = 1$	0.002	0.116	0.000	0.153	0.001	0.127	-0.043	0.133	0.000	0.138	-0.002	0.166
$\beta_2 = 1$	0.33	0.337	0.263	0.498	0.259	0.348	0.074	0.16	0.063	0.171	-0.085	0.256
$\sigma = 1$	-0.205	0.221	0.053	0.362	-0.126	0.156	-0.05	0.135	-0.029	0.131	0.146	0.279
$\rho_{ux} = 0.6$			-0.49	0.738	-0.421	0.637	-0.132	0.254	-0.119	0.284	-0.002	0.295
$N = 100$												
$\beta_1 = 1$	0.000	0.08	0.003	0.108	0.002	0.092	-0.02	0.096	0.002	0.1	0.004	0.122
$\beta_2 = 1$	0.332	0.335	0.201	0.403	0.228	0.308	0.074	0.153	0.064	0.163	-0.099	0.229
$\sigma = 1$	-0.203	0.211	0.005	0.302	-0.118	0.144	-0.051	0.133	-0.032	0.133	0.148	0.27
$\rho_{ux} = 0.6$			-0.393	0.623	-0.369	0.562	-0.126	0.242	-0.116	0.268	0.036	0.239
$N = 250$												
$\beta_1 = 1$	-0.002	0.05	-0.003	0.061	-0.003	0.056	-0.012	0.059	-0.004	0.06	-0.005	0.072
$\beta_2 = 1$	0.33	0.331	0.151	0.306	0.186	0.259	0.075	0.151	0.068	0.159	-0.1	0.199
$\sigma = 1$	-0.199	0.202	-0.033	0.223	-0.105	0.131	-0.05	0.128	-0.038	0.13	0.141	0.239
$\rho_{ux} = 0.6$			-0.297	0.494	-0.305	0.472	-0.13	0.241	-0.124	0.259	0.053	0.19
$N = 500$												
$\beta_1 = 1$	0.001	0.037	0.001	0.044	0.001	0.042	-0.003	0.044	0.001	0.045	0.001	0.053
$\beta_2 = 1$	0.33	0.331	0.106	0.217	0.131	0.192	0.066	0.137	0.059	0.14	-0.105	0.175
$\sigma = 1$	-0.199	0.201	-0.05	0.17	-0.091	0.119	-0.046	0.119	-0.037	0.12	0.139	0.215
$\rho_{ux} = 0.6$			-0.199	0.351	-0.198	0.33	-0.111	0.212	-0.103	0.219	0.074	0.158
$N = 1000$												
$\beta_1 = 1$	0.000	0.027	-0.001	0.032	-0.001	0.031	-0.003	0.032	-0.001	0.032	-0.001	0.037
$\beta_2 = 1$	0.329	0.33	0.073	0.163	0.085	0.139	0.053	0.122	0.048	0.123	-0.101	0.156
$\sigma = 1$	-0.198	0.199	-0.044	0.143	-0.066	0.099	-0.038	0.112	-0.032	0.113	0.129	0.194
$\rho_{ux} = 0.6$			-0.133	0.25	-0.122	0.225	-0.09	0.179	-0.084	0.181	0.078	0.134

Notes: Simulation results for model (13), when $x_i \sim Logistic(0,1)$. Bias and RMSE of the alternative estimators employed are presented; namely OLS, the copula control function approach (CCF), the maximum likelihood estimator (MLE) proposed by Park and Gupta (2012), the two maximum entropy estimators, denoted as MEE1 and MEE2, and the ridge LS estimator as suggested by Singh et al. (1986), denoted as RLS. Values equal to 0.000 correspond to values less than 5×10^{-4}.

The results of the tables lead to the following interesting conclusions. Firstly, Table 1 indicate that all three distributions considered for the endogenous regressor imply a close to unity correlation coefficient between the endogenous regressor (x) and its transformation (x^*), which can lead to a serious multicollinearity problem. The problem can become more severe as the sample size increases, since the variance of the estimation error of the ECDF (st.dev.($\hat{v}$)) reduces. In addition, it can be also more profound for the Student's-t distribution with 9 degrees of freedom which is more close to the normal distribution. The decline of the standard deviation of the estimation error $\hat{v}$ with N, observed in the table, is in accordance to the Glivenko - Cantelli theorem. Finally, another interesting conclusion that can be drawn from the table is that the correlation coefficients between the estimation error $\hat{v}$ and either of variables x or x^* are almost zero, which does not raise any concern for endogeneity of the CCF x^*. This result is true for all distributions considered.

Secondly, the rejection frequency of the normality tests, reported in Table 2, indicate that we need large sizes of N so that the tests to have very good power (e.g. bigger than 90%) to reject the normality of x. This is true across the alternative distributions considered. Note that, for the Cramér-von Mises test, this power is not achieved even for sample sizes of $N = 1000$.

Thirdly, regarding the performance of the alternative estimators considered, the results of Tables 3A-3C clearly indicate that, as was expected, ignoring endogeneity leads to seriously biased and highly RMSE estimates of the slope coefficients of the linear regression (see the first two columns of the tables), especially the slope coefficient β_2. The bias does not substantially reduce with N and becomes worst (almost 50% of the true value of β_2) for the case of the Student's-t distribution which is closer to the normal. For the two other distributions considered, the bias is also severe ranging from 50% to 30% of the true value of β_2. Similar conclusions can be drawn for the RMSE of the slope estimates, reported in the tables. Moreover the bias exists even for cases that

the normality tests, reported in Table 2, reject the normality hypothesis at very high rejection frequency.

The ML and LS estimators suggested by PG to control for the endogenous regression problem, based on the CCF, are not able to substantially remove the bias (or reduce the RMSE) in the estimates of β_2 even in moderate or large sizes of N, i.e., $N = \{500, 1000\}$ and for cases that the normality hypothesis of x is clearly rejected. The performance of these two estimators is worst and improves very slowly with N in the case of the Student's-t distribution, which is closer to the normal. As in Becker et al. (2022), we have found that we need a very large size of sample (i.e., $N \geq 1000$), not often met, in practice) so that the bias of the two estimators to be eliminated. Evidence of non-normality of x is not sufficient to guarantee satisfactory performance of the method.

The results of the tables clearly indicate that the two maximum entropy estimators and the ridge estimator considered can offer a solution to the above problems. The use of these estimators reduces substantially the bias and RMSE of the estimates of coefficient β_2 even for small samples, e.g. $N = \{50, 100\}$. These results hold across all the distributions considered. Note that the performance of the above estimators is very satisfactory even for the case that x is normally distributed, reported in the appendix, for which x and x^* are fully multicollinear asymptotically. Another interesting conclusion that can be drawn from the results of tables 3A-3C concerns the performance of the above estimators themselves. The results suggest that the two maximum entropy estimators (MEE1 and MEE2) perform equally well to each other and clearly outperform the ridge estimator, in terms of both the bias and the RMSE metrics reported in the tables.

Finally, regarding inference about the slope coefficient estimates, the results of Tables 4A-4C do not change our conclusion about the superiority of the MEE1 and MEE2 estimators. For the two entropy estimators, the t-ratio statistics used to test the following null hypotheses $H_0 : \beta_1 = 1$ and $H_0 : \beta_2 = 1$, are found to be slightly oversized in smaller samples, especially for $H_0 : \beta_1 = 1$. The t-ratio statistic for hypothesis $H_0 : \beta_2 = 1$ at-

Table 4A: Power and Size, $x_i \sim t(6)$

		$\beta_1 = 1$		$\beta_2 = 1$		$\lambda = 0$	
		size	power	size	power	size	power
N=50	MEE1	0.135	0.517	0.102	0.201	0.311	0.829
	MEE2	0.115	0.382	0.099	0.226	0.317	0.822
	RLS	0.118	0.395	0.236	0.741	0.142	0.617
N=100	MEE1	0.112	0.72	0.091	0.163	0.213	0.818
	MEE2	0.108	0.656	0.086	0.175	0.213	0.789
	RLS	0.139	0.605	0.392	0.095	0.755	0.886
N=250	MEE1	0.11	0.962	0.09	0.137	0.127	0.812
	MEE2	0.11	0.946	0.086	0.144	0.128	0.779
	RLS	0.141	0.938	0.652	0.556	0.81	0.983
N=500	MEE1	0.11	0.999	0.079	0.184	0.089	0.894
	MEE2	0.104	0.999	0.071	0.191	0.097	0.864
	RLS	0.139	0.998	0.387	0.789	0.698	0.985
N=1000	MEE1	0.101	1.000	0.073	0.279	0.065	0.962
	MEE2	0.1	1.000	0.069	0.275	0.076	0.954
	RLS	0.14	1.000	0.61	0.919	0.834	1.000

Notes: The table presents the power and size of the t-ratio test statistics of the null hypotheses $H_0 : \beta_1 = 1$ and $H_0 : \beta_2 = 1$ and $H_0 : \lambda = 0$ against the alternatives defined in the text. Values equal to 1.000 correspond to values greater or equal than 0.9995.

Table 4B: Power and Size, $x_i \sim t(9)$

		$\beta_1 = 1$		$\beta_2 = 1$		$\lambda = 0$	
		size	power	size	power	size	power
N=50	MEE1	0.121	0.517	0.076	0.229	0.34	0.802
	MEE2	0.088	0.412	0.083	0.259	0.363	0.789
	RLS	0.115	0.417	0.461	0.452	0.426	0.854
N=100	MEE1	0.119	0.729	0.058	0.179	0.326	0.758
	MEE2	0.107	0.651	0.063	0.199	0.299	0.736
	RLS	0.133	0.643	0.431	0.472	0.509	0.834
N=250	MEE1	0.077	0.95	0.046	0.149	0.197	0.75
	MEE2	0.079	0.936	0.046	0.155	0.164	0.716
	RLS	0.109	0.923	0.42	0.458	0.491	0.934
N=500	MEE1	0.105	0.999	0.06	0.131	0.135	0.761
	MEE2	0.11	0.998	0.053	0.135	0.131	0.725
	RLS	0.144	0.998	0.459	0.802	0.754	0.97
N=1000	MEE1	0.101	1.000	0.077	0.131	0.102	0.885
	MEE2	0.092	1.000	0.076	0.134	0.102	0.856
	RLS	0.125	1.000	0.545	0.829	0.779	0.996

Notes: The table presents the power and size of the t-ratio test statistics of the null hypotheses $H_0 : \beta_1 = 1$ and $H_0 : \beta_2 = 1$ and $H_0 : \lambda = 0$ against the alternatives defined in the text. Values equal to 1.000 correspond to values greater or equal than 0.9995.

Table 4C: Power and Size, $x_i \sim Logistic(0, 1)$

		$\beta_1 = 1$		$\beta_2 = 1$		$\lambda = 0$	
		size	power	size	power	size	power
N=50	MEE1	0.105	0.535	0.074	0.159	0.239	0.498
	MEE2	0.089	0.405	0.069	0.196	0.266	0.516
	RLS	0.159	0.439	0.526	0.528	0.705	0.82
N=100	MEE1	0.1	0.743	0.071	0.147	0.176	0.487
	MEE2	0.086	0.658	0.066	0.161	0.185	0.481
	RLS	0.178	0.671	0.512	0.762	0.621	0.85
N=250	MEE1	0.08	0.961	0.083	0.158	0.121	0.467
	MEE2	0.076	0.938	0.073	0.179	0.121	0.475
	RLS	0.146	0.925	0.236	0.802	0.798	0.645
N=500	MEE1	0.096	0.996	0.066	0.205	0.098	0.568
	MEE2	0.101	0.995	0.06	0.222	0.105	0.56
	RLS	0.192	0.994	0.672	0.935	0.776	0.991
N=1000	MEE1	0.105	1.000	0.093	0.286	0.083	0.746
	MEE2	0.106	1.000	0.081	0.302	0.098	0.733
	RLS	0.171	1.000	0.597	0.966	0.868	0.994

Notes: The table presents the power and size of the t-ratio test statistics of the null hypotheses $H_0 : \beta_1 = 1$ and $H_0 : \beta_2 = 1$ and $H_0 : \lambda = 0$ against the alternatives defined in the text. Values equal to 1.000 correspond to values greater or equal than 0.9995.

tends a size closer to its nominal 5% value, even for the small size $N = 100$. The power of the statistics is always higher than their size, meaning that they are unbiased. Note that the power also increases with N, as is expected. In contrast, for the ridge estimator, the size performance of the above statistics is much worse, especially for $H_0 : \beta_2 = 1$. For this case the size reached levels of 50%, or above.

Regarding the t-ratio statistic of the null hypothesis $H_0 : \lambda = 0$, meaning that regressor x is exogenous, the results of the tables demonstrate that, for both the two entropy and ridge estimators the above statistics are critically oversized in small samples of size $N = \{50, 100, 250\}$. However, the size approaches its nominal size as N increases to $N = 1000$ for the two maximum entropy estimators. For the ridge estimator, the size performance of the test always remains unsatisfactory, and it is not improved with N. This may be attributed to the fact that the ridge estimator is biased. As a final note that both the size and power of the t-ratio statistics improve considerably, as the distribution of the endogenous regressor deviates from the normal. This can be easily seen by comparing, for instance, the results of the Student's-t distributions with nine and six degrees of freedom, respectively.

Summing up, the results of this section indicate that the two maximum entropy estimators, MEE1 and MEE2, coping with the problem of multicollinearity in linear regressions, can be successfully employed to improve the performance of the copula approach to control for the endogenous regressor problem in linear regressions cases where regressors are close to normal distribution. These estimators clearly outperform the ridge estimator, especially in conducting inference about the slope coefficients of the regression and the CCF term.

3.2 Multiple regression MC results

Next, we consider a multiple regression model with two regressors:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i, \tag{17}$$

where x_{i2} and x_{i3} follow the same distribution and they are correlated with each other, but only one of them is endogenous (say x_{i2}). By modelling the correlation structure among x_{i2}, x_{i3} and u_i via a Gaussian copula, we can obtain a multiple factor representation similar to equation (5), such that the transformed regression becomes

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \lambda_1 x_{i2}^* + \lambda_2 x_{i3}^* + e_i, \tag{18}$$

where

$$\lambda_1 = \sigma_u \frac{\rho(u_i, x_{i2}) - \rho(x_{i2}, x_{i3})\rho(u_i, x_{i3})}{1 - \rho(x_{i2}, x_{i3})^2} \tag{19}$$

$$\lambda_2 = \sigma_u \frac{\rho(u_i, x_{i3}) - \rho(x_{i2}, x_{i3})\rho(u_i, x_{i2})}{1 - \rho(x_{i2}, x_{i3})^2} \tag{20}$$

From equation (20) it is clear that even though x_{i3} is strictly exogenous, i.e. $\rho(x_{i3}, u_i) = 0$, it holds that $\lambda_2 \neq 0$, since $\rho(x_{i2}, x_{i3}) \neq 0$. See, also, Yang et al. (2022) for a discussion.

For this simulation scenario, we generate 1000 samples of size $N = 50, 100, 200, 500$ from a joint normal distribution:

$$\begin{pmatrix} u_i \\ \tilde{x}_{i2} \\ \tilde{x}_{i3} \end{pmatrix} \sim N \left(\begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 1 & 0.6 & 0.0 \\ 0.6 & 1 & 0.3 \\ 0.0 & 0.3 & 1 \end{bmatrix} \right), \tag{21}$$

and obtain series for x_{i2} and x_{i3} as transformations of $\tilde{x}_{i2}$ and $\tilde{x}_{i3}$ respectively, as:

$$x_{ik} = QF(\Phi(\tilde{x}_{ik})), \text{ for } k = 2, 3. \tag{22}$$

where QF denotes the corresponding quantile function of the considered distribution. Then, we construct y_i based on equation (17), with $\beta_1 = 0.5$, $\beta_2 = 1$ and $\beta_3 = -0.5$.

The results of the multiple regression model are reported in Tables 5A-5C. The tables report results on the bias and RMSE metrics for all six alternative estimators considered in our single regression analysis (see Tables 3A-3C). Again, this is done across the three distributions considered (see (a)-(c)). Two main conclusions can be drawn from the results of the tables. Firstly, ignoring the endogeneity of a regressor can also lead to biased estimates of the exogenous regressor, when the two regressors are correlated. This result is consistent with Yang et al. (2022) analysis. Note that the bias of the exogenous regressor remains important as a percentage term of its true value, even if we control for the regressor endogeneity by adding in the RHS of the linear regression CCF terms of both the endogenous and exogenous regressors. As in the single regression, these biases can be attributed to identification problems related to the Gaussian transformation of both regressors of the model which are closely normally distributed.

Secondly, as in the single regression case, our results indicate that the maximum entropy estimators can save the copula approach from the above problems. MEE1 and MEE2 can substantially reduce the bias and RMSE of the slope coefficients of all the regressors of the model, including the endogenous and exogenous ones. The performance of these estimators is better than the ridge estimator. This is very satisfactory even for small samples, and increases fast with N. These results hold for all distributions of the regressors considered.

Table 5A: Simulation Results $x_{i2} \sim t(6)$ and $x_{i3} \sim t(6)$

	OLS		CCF		MLE		MEE1		MEE2		RLS	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
$N = 50$												
$\beta_1 = 0.5$	0.001	0.113	0.000	0.183	-0.498	0.498	-0.014	0.137	0.004	0.146	0.004	0.159
$\beta_2 = 1$	0.54	0.549	0.349	0.719	0.541	0.552	0.024	0.167	0.021	0.17	-0.065	0.254
$\beta_3 = -0.5$	-0.163	0.189	-0.126	0.587	-0.164	0.193	0.029	0.179	0.024	0.189	0.092	0.218
$N = 100$												
$\beta_1 = 0.5$	-0.002	0.078	-0.004	0.113	-0.497	0.497	-0.013	0.098	-0.004	0.101	-0.004	0.11
$\beta_2 = 1$	0.538	0.542	0.295	0.556	0.537	0.543	0.049	0.166	0.048	0.169	-0.068	0.206
$\beta_3 = -0.5$	-0.157	0.173	-0.088	0.469	-0.156	0.174	0.014	0.184	0.009	0.193	0.092	0.186
$N = 250$												
$\beta_1 = 0.5$	-0.002	0.05	-0.002	0.067	-0.496	0.497	-0.005	0.064	-0.002	0.065	-0.001	0.07
$\beta_2 = 1$	0.533	0.535	0.17	0.334	0.533	0.535	0.055	0.164	0.056	0.167	-0.091	0.166
$\beta_3 = -0.5$	-0.156	0.162	-0.072	0.315	-0.157	0.163	-0.001	0.168	-0.005	0.175	0.092	0.148
$N = 500$												
$\beta_1 = 0.5$	0.001	0.036	0.000	0.045	-0.494	0.496	-0.001	0.045	0.000	0.045	0.000	0.05
$\beta_2 = 1$	0.532	0.533	0.107	0.238	0.532	0.533	0.047	0.15	0.048	0.152	-0.102	0.151
$\beta_3 = -0.5$	-0.157	0.16	-0.026	0.215	-0.157	0.161	0.003	0.157	0.001	0.161	0.095	0.134
$N = 1000$												
$\beta_1 = 0.5$	0.001	0.025	0.001	0.031	-0.486	0.491	0.000	0.031	0.001	0.031	0.001	0.034
$\beta_2 = 1$	0.532	0.533	0.063	0.167	0.53	0.532	0.034	0.129	0.034	0.13	-0.104	0.14
$\beta_3 = -0.5$	-0.158	0.159	-0.016	0.147	-0.157	0.161	0.001	0.123	-0.001	0.125	0.091	0.117

Notes: Simulation results for model (17), when $x_{i2}, x_{i3} \sim t(6)$ and x_{i3} is exogenous. Bias and RMSE of the alternative estimators employed are presented; namely OLS, the copula control function approach (CCF), the maximum likelihood estimator (MLE) proposed by Park and Gupta (2012), the two maximum entropy estimators, denoted as MEE1 and MEE2, and the ridge LS estimator as suggested by Singh et al. (1986), denoted as RLS. Values equal to 0.000 correspond to values less than 5×10^{-4}.

Table 5B: Simulation Results $x_{i2} \sim t(9)$ and $x_{i3} \sim t(9)$

	OLS		CCF		MLE		MEE1		MEE2		RLS	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
$N = 50$												
$\beta_1 = 0.5$	0.005	0.112	0.003	0.184	-0.497	0.498	-0.015	0.135	0.002	0.144	0.000	0.15
$\beta_2 = 1$	0.582	0.592	0.455	0.848	0.58	0.591	0.01	0.169	0.009	0.172	-0.027	0.263
$\beta_3 = -0.5$	-0.172	0.203	-0.136	0.707	-0.174	0.208	0.046	0.193	0.04	0.205	0.091	0.245
$N = 100$												
$\beta_1 = 0.5$	-0.002	0.078	0.000	0.116	-0.498	0.498	-0.009	0.098	0.000	0.101	0.000	0.105
$\beta_2 = 1$	0.58	0.585	0.394	0.677	0.58	0.586	0.03	0.164	0.03	0.168	-0.036	0.216
$\beta_3 = -0.5$	-0.173	0.188	-0.13	0.596	-0.173	0.19	0.023	0.197	0.016	0.209	0.082	0.211
$N = 250$												
$\beta_1 = 0.5$	-0.003	0.051	-0.003	0.068	-0.497	0.498	-0.006	0.063	-0.003	0.063	-0.003	0.066
$\beta_2 = 1$	0.579	0.581	0.295	0.52	0.579	0.581	0.052	0.179	0.054	0.183	-0.049	0.189
$\beta_3 = -0.5$	-0.174	0.181	-0.096	0.44	-0.175	0.182	0.014	0.188	0.008	0.197	0.086	0.172
$N = 500$												
$\beta_1 = 0.5$	0.002	0.035	0.003	0.045	-0.494	0.496	0.001	0.044	0.003	0.044	0.003	0.047
$\beta_2 = 1$	0.579	0.58	0.219	0.394	0.578	0.579	0.061	0.179	0.062	0.183	-0.067	0.158
$\beta_3 = -0.5$	-0.174	0.177	-0.069	0.323	-0.174	0.178	0.006	0.178	0.001	0.186	0.089	0.146
$N = 1000$												
$\beta_1 = 0.5$	0.002	0.025	0.001	0.031	-0.484	0.49	0.000	0.031	0.001	0.032	0.002	0.035
$\beta_2 = 1$	0.578	0.579	0.123	0.275	0.577	0.578	0.041	0.163	0.042	0.165	-0.095	0.148
$\beta_3 = -0.5$	-0.173	0.174	-0.041	0.255	-0.173	0.177	0.003	0.171	0.000	0.176	0.093	0.136

Notes: Simulation results for model (17), when $x_{i2}, x_{i3} \sim t(9)$ and x_{i3} is exogenous. Bias and RMSE of the alternative estimators employed are presented; namely OLS, the copula control function approach (CCF), the maximum likelihood estimator (MLE) proposed by Park and Gupta (2012), the two maximum entropy estimators, denoted as MEE1 and MEE2, and the ridge LS estimator as suggested by Singh et al. (1986), denoted as RLS. Values equal to 0.000 correspond to values less than 5×10^{-4}.

Table 5C: Simulation Results $x_{i2} \sim Logistic(0,1)$ and $x_{i3} \sim Logistic(0,1)$

	OLS		CCF		MLE		MEE1		MEE2		RLS	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
$N = 50$												
$\beta_1 = 0.5$	-0.002	0.112	-0.005	0.18	-0.497	0.498	-0.02	0.129	-0.001	0.138	-0.002	0.171
$\beta_2 = 1$	0.362	0.368	0.266	0.501	0.362	0.37	0.083	0.156	0.079	0.158	-0.116	0.228
$\beta_3 = -0.5$	-0.11	0.129	-0.086	0.447	-0.109	0.132	-0.006	0.152	-0.011	0.162	0.103	0.194
$N = 100$												
$\beta_1 = 0.5$	0.008	0.079	0.011	0.119	-0.498	0.499	0.001	0.093	0.011	0.098	0.012	0.12
$\beta_2 = 1$	0.362	0.365	0.228	0.41	0.363	0.366	0.091	0.153	0.088	0.155	-0.117	0.208
$\beta_3 = -0.5$	-0.106	0.115	-0.064	0.335	-0.107	0.119	-0.011	0.145	-0.014	0.154	0.107	0.171
$N = 250$												
$\beta_1 = 0.5$	0.002	0.049	0.001	0.067	-0.495	0.497	-0.003	0.061	0.001	0.063	0.001	0.076
$\beta_2 = 1$	0.362	0.363	0.161	0.291	0.361	0.363	0.088	0.152	0.086	0.153	-0.124	0.184
$\beta_3 = -0.5$	-0.107	0.11	-0.059	0.252	-0.107	0.111	-0.019	0.14	-0.022	0.147	0.096	0.145
$N = 500$												
$\beta_1 = 0.5$	0.000	0.034	0.000	0.047	-0.49	0.493	-0.002	0.044	0.000	0.045	-0.001	0.054
$\beta_2 = 1$	0.362	0.363	0.119	0.221	0.362	0.363	0.082	0.139	0.08	0.14	-0.121	0.167
$\beta_3 = -0.5$	-0.107	0.109	-0.031	0.189	-0.107	0.109	-0.011	0.13	-0.012	0.135	0.097	0.133
$N = 1000$												
$\beta_1 = 0.5$	0.001	0.025	0.001	0.031	-0.47	0.482	0.000	0.03	0.001	0.03	0.001	0.036
$\beta_2 = 1$	0.361	0.361	0.078	0.158	0.358	0.365	0.061	0.119	0.06	0.119	-0.124	0.158
$\beta_3 = -0.5$	-0.107	0.108	-0.02	0.134	-0.105	0.115	-0.008	0.109	-0.009	0.112	0.092	0.119

Notes: Simulation results for model (17), when $x_{i2}, x_{i3} \sim Logistic(0,1)$ and x_{i3} is exogenous. Bias and RMSE of the alternative estimators employed are presented; namely OLS, the copula control function approach (CCF), the maximum likelihood estimator (MLE) proposed by Park and Gupta (2012), the two maximum entropy estimators, denoted as MEE1 and MEE2, and the ridge LS estimator as suggested by Singh et al. (1986), denoted as RLS. Values equal to 0.000 correspond to values less than 5×10^{-4}.

4 Conclusion

Since the seminar work of Park and Gupta (2012), there exist a growing number of empirical applications of the Gaussian copula method to deal with the endogenous regressor problem in marketing and management research. Results in the literature reveal critical limitations of the method when the distribution of the endogenous regression is close to the normal, implying a near multicollinearity problem between the endogenous regressor and its copula transformation term added to the regression as a control function regressor to deal with the endogeneity. This raises concerns about the ability of the method to identify the true regression slope coefficient values.

In this paper, to improve upon the performance of the method in the above cases, we suggest using estimators often employed in the literature to cope with the issue of near-multicollinearity. These include the maximum entropy and ridge estimators. We consider two variants of the maximum entropy estimator. The first takes into account the magnitude of the parameter estimates and their probabilities, and the second also considers magnitude and the probability specification of the error term estimates. To evaluate the performance of the above suggested estimators, we carry out a comprehensive simulation study which considers distributions of the endogenous regressor which are close to the normal distribution. In particular, we consider the Student's-t distributions with six and nine degrees of freedom and the Logistic which are close to the normal. These distributions lead to serious multicollinearity problems between the endogenous regressor and its copula transformation. Our analysis can be also extended to other distributions, which may also lead to near-multicollinearity problems.

We provide a number of very useful results for applied work. Firstly, we show that all of the above estimators can considerably improve the performance of the PG method in cases that the distribution of the endogenous regressor is close to the normal. We find that the maximum entropy estimators can reduce substantially the bias of the least

square estimates and their root mean square error (RMSE). These results hold not only for the single regression, but also for a multiple extension of it considering an exogenous regressor correlated to the endogenous one. Secondly, we find that the two variants of the maximum entropy estimators perform equally well and clearly outperform the ridge estimator, in terms of both bias and RMSE reductions. These two estimators are also found to perform well in conducting inference for the slope coefficients of the augmented linear regression with the copula transformed control function as an additional regressor to deal with the endogeneity problem, based on a bootstrap procedure.

Appendices

A Derivation of the conditional expectation

In this appendix we derive the conditional expectation $E(u_i|x_i)$. Based on equation (4) the conditional density $f_{u|x}(u_i|x_i)$ can be derived as follows

$$\begin{aligned} f_{u|x}(u_i|x_i) &= \frac{\partial}{\partial u_i} F_{u|x}(u_i|x_i) \\ &= \frac{\partial}{\partial u_i} \Phi_2 \left(\frac{\frac{u_i}{\sigma_u} - \rho_{ux} \Phi^{-1}\left(F_x(x_i)\right)}{\left(1-\rho_{ux}^2\right)^{1/2}} \right) \\ &= \frac{1}{\sigma_u \left(1-\rho_{ux}^2\right)^{1/2}} \phi \left(\frac{\frac{u_i}{\sigma_u} - \rho_{ux} \Phi^{-1}\left(F_x(x_i)\right)}{\left(1-\rho_{ux}^2\right)^{1/2}} \right), \end{aligned} \tag{23}$$

where $\phi(\cdot)$ denotes the univariate standard normal density function. Then $E(u_i|x_i)$ can be derived analytically as

$$\begin{aligned} E(u_i|x_i) &= \int_{\mathbb{R}} u_i f_{u|x}(u_i|x_i) du_i \\ &= \int_{\mathbb{R}} u_i \frac{1}{\sigma_u \left(1-\rho_{ux}^2\right)^{1/2}} \phi \left(\frac{\frac{u_i}{\sigma_u} - \rho_{ux} \Phi^{-1}\left(F_x(x_i)\right)}{\left(1-\rho_{ux}^2\right)^{1/2}} \right) du_i \\ &= \rho_{ux} \sigma_u x_i^* \end{aligned} \tag{24}$$

where $x_i^* = \Phi^{-1}\left(F_x(x_i)\right)$.

B Simulation results for the Normal case

In this appendix we present results for the bias and RMSE of the estimators for (β_1, β_2) for the single regression case with a normally distributed regressor (see Table B1). Note that the estimation error involved, in the first step, to obtain the CCF term x_i^* implies that x_i and x_i^* are near (and not perfect) multicollinear. This allows us to implement the MEE1, MEE2 and RLS methods in finite samples. If we used the true CDF, $\Phi(\cdot)$,

to obtain x_i^*, estimation would not be feasible due to perfect multicollinearity between x and x^*.

The results of this table are similar to those reported in Tables 3A-3C on the main text, for the other distributions. They show that, even for this most sever multicollinearity case, the above estimators perform well, with the two MEE estimators exhibiting superior performance. Similar conclusions can be drawn for the multiple regression case (see Table B2).

Table B1: Simulation Results $x_i \sim N(0,1)$

	OLS		CCF		MLE		MEE1		MEE2		RLS	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
$N = 50$												
$\beta_1 = 1$	-0.003	0.112	-0.004	0.171	-0.003	0.127	-0.043	0.138	-0.002	0.141	-0.002	0.144
$\beta_2 = 1$	0.601	0.612	0.593	1.118	0.515	0.673	-0.052	0.178	-0.054	0.181	0.058	0.454
$\sigma = 1$	-0.205	0.219	0.152	0.521	-0.122	0.15	0.046	0.117	0.053	0.12	0.036	0.195
$\rho_{ux} = 0.6$			-0.597	0.848	-0.466	0.681	0.013	0.116	0.019	0.124	-0.128	0.381
$N = 100$												
$\beta_1 = 1$	0.000	0.08	0.002	0.126	0.001	0.094	-0.021	0.103	0.001	0.104	0.001	0.107
$\beta_2 = 1$	0.604	0.609	0.608	1.024	0.538	0.687	-0.033	0.17	-0.032	0.178	0.077	0.404
$\sigma = 1$	-0.201	0.209	0.098	0.401	-0.106	0.13	0.039	0.11	0.042	0.113	0.014	0.167
$\rho_{ux} = 0.6$			-0.604	0.844	-0.507	0.704	0.008	0.116	0.009	0.127	-0.127	0.355
$N = 250$												
$\beta_1 = 1$	0.001	0.048	0.000	0.067	0.002	0.055	-0.007	0.063	0.002	0.064	0.002	0.064
$\beta_2 = 1$	0.598	0.6	0.594	0.946	0.513	0.664	-0.022	0.167	-0.018	0.178	0.077	0.368
$\sigma = 1$	-0.202	0.205	0.043	0.332	-0.101	0.125	0.023	0.1	0.024	0.103	0.005	0.136
$\rho_{ux} = 0.6$			-0.589	0.804	-0.498	0.683	-0.001	0.118	-0.005	0.133	-0.116	0.336
$N = 500$												
$\beta_1 = 1$	0.000	0.035	0.001	0.047	0.001	0.04	-0.003	0.045	0.001	0.045	0.001	0.044
$\beta_2 = 1$	0.6	0.601	0.621	0.961	0.477	0.646	-0.007	0.164	-0.002	0.176	0.101	0.363
$\sigma = 1$	-0.199	0.201	0.044	0.317	-0.086	0.113	0.016	0.1	0.016	0.103	-0.006	0.137
$\rho_{ux} = 0.6$			-0.631	0.836	-0.479	0.672	-0.009	0.111	-0.015	0.127	-0.13	0.332
$N = 1000$												
$\beta_1 = 1$	0.001	0.026	0.001	0.034	0.001	0.03	-0.001	0.032	0.001	0.032	0.001	0.032
$\beta_2 = 1$	0.6	0.601	0.594	0.907	0.209	0.403	-0.004	0.17	0.003	0.184	0.092	0.352
$\sigma = 1$	-0.2	0.201	0.019	0.278	-0.048	0.073	0.014	0.099	0.012	0.103	-0.003	0.127
$\rho_{ux} = 0.6$			-0.594	0.802	-0.202	0.426	-0.013	0.122	-0.02	0.139	-0.119	0.326

Notes: Simulation results for model (13), when $x_i \sim N(0,1)$. Bias and RMSE of the alternative estimators employed are presented; namely OLS, the copula control function approach (CCF), the maximum likelihood estimator (MLE) proposed by Park and Gupta (2012), the two maximum entropy estimators, denoted as MEE1 and MEE2, and the ridge LS estimator as suggested by Singh et al. (1986), denoted as RLS. Values equal to 0.000 correspond to values less than 5×10^{-4}.

Table B2: Simulation Results $x_{i2} \sim N(0,1)$ and $x_{i3} \sim N(0,1)$

	OLS		CCF		MLE		MEE1		MEE2		RLS	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
$N = 50$												
$\beta_1 = 0.5$	0.000	0.115	-0.014	0.215	-0.497	0.498	-0.02	0.142	-0.003	0.15	-0.004	0.152
$\beta_2 = 1$	0.659	0.67	0.681	1.126	0.659	0.67	-0.028	0.158	-0.028	0.161	0.062	0.356
$\beta_3 = -0.5$	-0.196	0.232	-0.209	0.939	-0.196	0.234	0.075	0.199	0.068	0.211	0.063	0.304
$N = 100$												
$\beta_1 = 0.5$	-0.003	0.08	-0.007	0.134	-0.495	0.496	-0.013	0.101	-0.004	0.104	-0.005	0.104
$\beta_2 = 1$	0.658	0.663	0.625	0.99	0.657	0.663	-0.023	0.149	-0.022	0.152	0.041	0.286
$\beta_3 = -0.5$	-0.195	0.212	-0.167	0.82	-0.195	0.213	0.073	0.198	0.066	0.208	0.074	0.254
$N = 250$												
$\beta_1 = 0.5$	0.001	0.052	0.001	0.08	-0.498	0.498	0.000	0.066	0.003	0.067	0.003	0.067
$\beta_2 = 1$	0.663	0.665	0.668	0.975	0.664	0.666	0.004	0.152	0.006	0.156	0.066	0.284
$\beta_3 = -0.5$	-0.196	0.203	-0.183	0.691	-0.196	0.203	0.058	0.183	0.05	0.194	0.059	0.22
$N = 500$												
$\beta_1 = 0.5$	0.000	0.034	0.001	0.054	-0.495	0.496	-0.003	0.044	-0.001	0.045	-0.001	0.044
$\beta_2 = 1$	0.659	0.66	0.659	0.96	0.658	0.659	0.009	0.154	0.013	0.158	0.07	0.28
$\beta_3 = -0.5$	-0.197	0.201	-0.192	0.722	-0.198	0.201	0.052	0.198	0.045	0.21	0.05	0.228
$N = 1000$												
$\beta_1 = 0.5$	0.000	0.025	0.001	0.037	-0.484	0.49	-0.002	0.033	-0.001	0.033	-0.001	0.033
$\beta_2 = 1$	0.66	0.66	0.644	0.92	0.656	0.661	0.024	0.161	0.028	0.167	0.075	0.274
$\beta_3 = -0.5$	-0.197	0.199	-0.213	0.696	-0.199	0.211	0.038	0.201	0.029	0.213	0.041	0.221

Notes: Simulation results for model (17), when $x_{i2}, x_{i3} \sim N(0,1)$ and x_{i3} is exogenous. Bias and RMSE of the alternative estimators employed are presented; namely OLS, the copula control function approach (CCF), the maximum likelihood estimator (MLE) proposed by Park and Gupta (2012), the two maximum entropy estimators, denoted as MEE1 and MEE2, and the ridge LS estimator as suggested by Singh et al. (1986), denoted as RLS. Values equal to 0.000 correspond to values less than 5×10^{-4}.

References

Anderson, T. W. and Darling, D. A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193–212.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753.

Becker, J. M., Proksch, D., and Ringle, C. M. (2022). Revisiting Gaussian copulas to handle endogenous regressors. *Journal of the Academy of Marketing Science*, 50:46–66.

Belsley, D. A. (1991). A Guide to Using the Collinearity Diagnostics. *Computer Science in Economics and Management*, 4:33–50.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, Inc., New York.

Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, 4:421–424.

Christopoulos, D., McAdam, P., and Tzavalis, E. (2021). Dealing With Endogeneity in Threshold Models Using Copulas. *Journal of Business & Economic Statistics*, 39(1):166–178.

Christopoulos, D., McAdam, P., and Tzavalis, E. (2023a). Exploring Okun's law asymmetry: An endogenous threshold logistic smooth transition regression approach. *Oxford Bulletin of Economics and Statistics*, 85(1):123–158.

Christopoulos, D., Smyrnakis, D., and Tzavalis, E. (2022). Human capital threshold effects in economic development: A panel data approach with endogenous threshold. Working Paper 17-2022, Department of Economics, Athens University of Economics & Business.

Christopoulos, D., Smyrnakis, D., and Tzavalis, E. (2023b). A multiple regression extension of the Gaussian copula approach in dealing with endogenous regressors. mimeo.

Cragg, J. G. (1997). Using Higher Moments to Estimate the Simple Errors-in-Variables Model. *The RAND Journal of Economics*, 28:S71–S91.

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Dagenais, M. G. and Dagenais, D. L. (1997). Higher moment estimators for linear regression models with errors in the variables. *Journal of Econometrics*, 76(1):193–221.

Davidson, R. and MacKinnon, J. G. (2006). *Bootstrap Methods in Ecovometrics*, chapter 25. Palgrave Handbooks of Econometrics: Vol. 1 Econometric Theory.

Dwivedi, T. D. and Srivastava, V. K. (1978). On the minimum mean squared error estimators in a regression model. *Communications in Statistics - Theory and Methods*, 7(5):487–494.

Ebbes, P., Wedel, M., Böckenholt, U., and Steerneman, T. (2005). Solving and Testing for Regressor-Error (in)Dependence When no Instrumental Variables are Available: With New Evidence for the Effect of Education on Income. *Quantitative Marketing and Economics*, 3(4):365–392.

Eckert, C. and Hohberger, J. (2022). Addressing Endogeneity Without Instrumental Variables: An Evaluation of the Gaussian Copula Approach for Management Research. *Journal of Management*, 49(4):1–36.

Erickson, T. and Whited, T. M. (2002). Two-Step GMM Estimation of the Errors-in-Variables Model Using High-Order Moments. *Econometric Theory*, 18(3):776–799.

Gabaix, X. and Koijen, R. S. J. (2022). Granular Instrumental Variables. NBER Working Papers 28204, National Bureau of Economic Research, Inc.

Gençay, R. and Gradojevic, N. (2011). Errors-in-variables estimation with wavelets. *Journal of Statistical Computation and Simulation*, 81(11):1545–1564.

Glivenko, V. I. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, 4:92–99.

Golan, A., Judge, G. G., and Miller, D. (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley & Sons, Inc., Chichester UK.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21(2):215–223.

Greene, W. H. (2018). *Econometric Analysis, 8th ed.* Pearson, New York.

Hahn, J. and Hausman, J. (2005). Estimation with Valid and Invalid Instruments. *Annals of Economics and Statistics*, (79-80):25–57.

Haschka, R. E. (2022). Handling Endogenous Regressors Using Copulas: A Generalization to Linear Panel Models with Fixed Effects and Correlated Regressors. *Journal of Marketing Research*, 59(4):860–881.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12(1):69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge regression:some simulations. *Communications in Statistics - Theory and Methods*, 4(2):105–123.

Jarque, C. M. and Bera, A. K. (1987). A Test for Normality of Observations and Regression Residuals. *International Statistical Review*, 55(2):163–172.

Joe, H. (2014). *Dependence Modelling with Copulas*. Chapman & Hall, London.

Khalaf, G. (2012). A Proposed Ridge Parameter to Improve the Least Square Estimator. *Journal of Modern Applied Statistical Methods*, 11(2):443–449.

Khalaf, G. (2013). A Comparison between Biased and Unbiased Estimators in Ordinary Least Squares Regression. *Journal of Modern Applied Statistical Methods*, 12(2):293–303.

Khalaf, G. and Shukur, G. (2005). Choosing Ridge Parameter for Regression Problems. *Communications in Statistics - Theory and Methods*, 34(5):1177–1182.

Klein, R. and Vella, F. (2010). Estimating a class of triangular simultaneous equations models without exclusion restrictions. *Journal of Econometrics*, 154(2):154–164.

Lawless, J. F. and Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics - Theory and Methods*, 5(4):307–323.

Lewbel, A. (1997). Constructing Instruments for Regressions With Measurement Error When no Additional Data are Available, with An Application to Patents and R&D. *Econometrica*, 65(5):1201–1213.

Lewbel, A. (2012). Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models. *Journal of Business and Economic Statistics*, 30(1):67–80.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, 62(318):399–402.

Paris, Q. (2004). Maximun Entropy Leuven Estimators and Multicollinearity. *Statistica*, 64(1):3–22.

Park, S. and Gupta, S. (2012). Handling Endogenous Regressors by Joint Estimation Using Copulas. *Marketing Science*, 31(4):567–586.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis, 3rd ed.* Cengage Learning.

Rigobon, R. (2003). Identification Through Heteroskedasticity. *The Review of Economics and Statistics*, 85(4):777–792.

Rutz, O. J. and Watson, G. F. (2019). Endogeneity and marketing strategy research: an overview. *Journal of the Academy of Marketing Science*, 47:479–498.

Shapiro, S. S. and Francia, R. S. (1972). An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association*, 67(337):215–216.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Singh, B., Chaubey, Y. P., and Dwivedi, T. D. (1986). An Almost Unbiased Ridge Estimator. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 48(3):342–346.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231.

Spanos, A. (2018). Misspecification testing in retrospect. *Journal of Economic Surveys*, 32(2):541–577.

Spanos, A. and McGuirk, A. (2002). The problem of near-multicollinearity revisited: erratic vs systematic volatility. *Journal of Econometrics*, 108(2):365–393.

Thisted, R. A. (1976). *Ridge Regression, Minimax Estimation, and Empirical Bayes Methods*. PhD thesis, Department of Statistics, Stanford University.

von Mises, R. E. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*. Julius Springer.

Yang, F., Qian, Y., and Xie, H. (2022). Addressing Endogeneity Using a Two-stage Copula Generated Regressor Approach. Working Paper 29708, National Bureau of Economic Research.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
SCHOOL OF ECONOMIC SCIENCES

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΗΣ ΕΠΙΣΤΗΜΗΣ
DEPARTMENT OF ECONOMICS

Department of Economics
Athens University of Economics and Business

List of Recent Working Papers

2021

01-21 **Historical Cycles of the Economy of Modern Greece From 1821 to the Present, George Alogoskoufis**

02-21 **Greece Before and After the Euro: Macroeconomics, Politics and the Quest for Reforms, George Alogoskoufis**

03-21 **Commodity money and the price level, George C. Bitros. Published in: *Quarterly Journal of Austrian Economics*, 2022**

04-21 **Destabilizing asymmetries in central banking: With some enlightenment from money in classical Athens, George C. Bitros. Published in: *Journal of Economic Asymmetries*, 2021**

05-21 **Exploring the Long-Term Impact of Maximum Markup Deregulation, Athanasios Dimas and Christos Genakos**

06-21 **A regularization approach for estimation and variable selection in high dimensional regression models, Y. Dendramis, L. Giraitis, G. Kapetanios**

07-21 **Tax Competition in the Presence of Environmental Spillovers, Fabio Antoniou, Panos Hatzipanayotou, Michael S. Michael, Nikos Tsakiris**

08-21 **Firm Dynamics by Age and Size Classes and the Choice of Size Measure, Stelios Giannoulakis and Plutarchos Sakellaris**

09-21 **Measuring the Systemic Importance of Banks, Georgios Moratis, Plutarchos Sakellaris**

10-21 **Firms' Financing Dynamics Around Lumpy Capacity Adjustments, Christoph Görtz, Plutarchos Sakellaris, John D. Tsoukalas**

11-21 **On the provision of excludable public goods General taxes or user prices? George Economides and Apostolis Philippopoulos**

12-21 **Asymmetries of Financial Openness in an Optimal Growth Model, George Alogoskoufis**

13-21 **Evaluating the impact of labour market reforms in Greece during 2010-2018 , Georgios Gatopoulos, Alexandros Louka, Ioannis Polycarpou, Nikolaos Vettas**

14-21 **From the Athenian silver to the bitcoin standard: Private money in a state-enforced free banking model, George C. Bitros**

15-21 **Ordering Arbitrage Portfolios and Finding Arbitrage Opportunities. Stelios Arvanitis and Thierry Post**

16-21 **Inconsistency for the Gaussian QMLE in GARCH-type models with infinite variance, Stelios Arvanitis and Alexandros Louka**

17-21 **Competition and Pass-Through: Evidence from Isolated Markets, Christos Genakos and Mario Pagliero**

18-21 **Exploring Okun's Law Asymmetry: An Endogenous Threshold LSTR Approach, Dimitris Christopoulos, Peter McAdam and Elias Tzavalis**

19-21 **Limit Theory for Martingale Transforms with Heavy-Tailed Multiplicative Noise, Stelios Arvanitis and Alexandros Louka**

20-21 **Optimal taxation with positional considerations, Ourania Karakosta and Eleftherios Zacharias**

21-21 The ECB's policy, the Recovery Fund and the importance of trust: The case of Greece, Vasiliki Dimakopoulou, George Economides and Apostolis Philippopoulos

2022

01-22 Is Ireland the most intangible intensive economy in Europe? A growth accounting perspective, Ilias Kostarakos, KieranMcQuinn and Petros Varthalitis
02-22 Common bank supervision and profitability convergence in the EU, Ioanna Avgeri, Yiannis Dendramis and Helen Louri
03-22 Missing Values in Panel Data Unit Root Tests, Yiannis Karavias, Elias Tzavalis and Haotian Zhang
04-22 Ordering Arbitrage Portfolios and Finding Arbitrage Opportunities, Stelios Arvanitis and Thierry Post
05-22 Concentration Inequalities for Kernel Density Estimators under Uniform Mixing, Stelios Arvanitis
06-22 Public Sector Corruption and the Valuation of Systemically Important Banks, Georgios Bertsatos, Spyros Pagratis, Plutarchos Sakellaris
07-22 Finance or Demand: What drives the Responses of Young and Small Firms to Financial Crises? Stelios Giannoulakis and Plutarchos Sakellaris
08-22 Production function estimation controlling for endogenous productivity disruptions, Plutarchos Sakellaris and Dimitris Zaverdas
09-22 A panel bounds testing procedure, Georgios Bertsatos, Plutarchos Sakellaris, Mike G. Tsionas
10-22 Social policy gone bad educationally: Unintended peer effects from transferred students, Christos Genakos and Eleni Kyrkopoulou
11-22 Inconsistency for the Gaussian QMLE in GARCH-type models with infinite variance, Stelios Arvanitis and Alexandros Louka
12-22 Time to question the wisdom of active monetary policies, George C. Bitros
13-22 Investors' Behavior in Cryptocurrency Market, Stelios Arvanitis, Nikolas Topaloglou and Georgios Tsomidis
14-22 On the asking price for selling Chelsea FC, Georgios Bertsatos and Gerassimos Sapountzoglou
15-22 Hysteresis, Financial Frictions and Monetary Policy, Konstantinos Giakas
16-22 Delay in Childbearing and the Evolution of Fertility Rates, Evangelos Dioikitopoulos and Dimitrios Varvarigos
17-22 Human capital threshold effects in economic development: A panel data approach with endogenous threshold, Dimitris Christopoulos, Dimitris Smyrnakis and Elias Tzavalis
18-22 Distributional aspects of rent seeking activities in a Real Business Cycle model, Tryfonas Christou, Apostolis Philippopoulos and Vanghelis Vassilatos

2023

01-23 Real interest rate and monetary policy in the post Bretton Woods United States, George C. Bitros and Mara Vidali
02-23 Debt targets and fiscal consolidation in a two-country HANK model: the case of Euro Area, Xiaoshan Chen, Spyridon Lazarakis and Petros Varthalitis
03-23 Central bank digital currencies: Foundational issues and prospects looking forward, George C. Bitros and Anastasios G. Malliaris
04-23 The State and the Economy of Modern Greece. Key Drivers from 1821 to the Present, George Alogoskoufis
05-23 Sparse spanning portfolios and under-diversification with second-order stochastic dominance, Stelios Arvanitis, Olivier Scaillet, Nikolas Topaloglou

06-23 What makes for survival? Key characteristics of Greek incubated early-stage startup(per)s during the Crisis: a multivariate and machine learning approach, Ioannis Besis, Ioanna Sapfo Pepelasis and Spiros Paraskevas

07-23 The Twin Deficits, Monetary Instability and Debt Crises in the History of Modern Greece, George Alogoskoufis

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ SCHOOL OF ECONOMIC SCIENCES
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΗΣ ΕΠΙΣΤΗΜΗΣ DEPARTMENT OF ECONOMICS

Department of Economics
Athens University of Economics and Business

The Department is the oldest Department of Economics in Greece with a pioneering role in organising postgraduate studies in Economics since 1978. Its priority has always been to bring together highly qualified academics and top quality students. Faculty members specialize in a wide range of topics in economics, with teaching and research experience in world-class universities and publications in top academic journals.

The Department constantly strives to maintain its high level of research and teaching standards. It covers a wide range of economic studies in micro-and macroeconomic analysis, banking and finance, public and monetary economics, international and rural economics, labour economics, industrial organization and strategy, economics of the environment and natural resources, economic history and relevant quantitative tools of mathematics, statistics and econometrics.

Its undergraduate program attracts high quality students who, after successful completion of their studies, have excellent prospects for employment in the private and public sector, including areas such as business, banking, finance and advisory services. Also, graduates of the program have solid foundations in economics and related tools and are regularly admitted to top graduate programs internationally. Three specializations are offered:1. Economic Theory and Policy, 2. Business Economics and Finance and 3. International and European Economics. The postgraduate programs of the Department (M.Sc and Ph.D) are highly regarded and attract a large number of quality candidates every year.

For more information:

https://www.dept.aueb.gr/en/econ/