NBER WORKING PAPER SERIES

ADDRESSING ENDOGENEITY USING A TWO-STAGE COPULA GENERATED REGRESSOR APPROACH

Fan Yang
Yi Qian
Hui Xie

Working Paper 29708
http://www.nber.org/papers/w29708

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2022, Revised July 2022

We thank colleagues for helpful comments. We acknowledge the support by the Social Sciences and Humanities Research Council of Canada [grant 435-2018-0519], Natural Sciences and Engineering Research Council of Canada [grant RGPIN-2018-04313] and US National Institute of Health [grants R01CA178061] . All inferences, opinions, and conclusions drawn in this study are those of the authors, and do not reflect the opinions or policies of the funding agencies and data stewards. No personal identifying information was made available as part of this study. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Addressing Endogeneity Using a Two-stage Copula Generated Regressor Approach
Fan Yang, Yi Qian, and Hui Xie
NBER Working Paper No. 29708
January 2022, Revised July 2022
JEL No. C01,C1,C13,C18,C4

ABSTRACT

A prominent challenge when drawing causal inference using observational data is the ubiquitous presence of endogenous regressors. The classical econometric method to handle regressor endogeneity requires instrumental variables that must satisfy the stringent condition of exclusion restriction, making it infeasible to use in many settings. We propose new instrument-free methods using copulas to address the endogeneity problem. Existing copula correction methods require non-normal endogenous regressors: normally or nearly normally distributed endogenous regressors cause model non-identification or significant finite-sample bias. Furthermore, existing copula control function methods presume the independence of exogenous regressors and the copula control function. Our proposed two-stage copula endogeneity correction (2sCOPE) method simultaneously relaxes the two key identification requirements, and we prove that 2sCOPE yields consistent causal-effect estimates with correlated endogenous and exogenous regressors as well as normally distributed endogenous regressors. Besides relaxing identification requirements, 2sCOPE has superior finite-sample performance and addresses the significant finite sample bias problem due to insufficient regressor non-normality. 2sCOPE employs generated regressors derived from existing regressors to control for endogeneity, and is straightforward to use and broadly applicable. Overall, 2sCOPE can greatly increase the ease and broaden the applicability of using instrument-free methods to handle regressor endogeneity. We further demonstrate the performance of 2sCOPE via simulation studies and an empirical application.

Fan Yang
Sauder School of Business
2053 Main Mall
Vancouver, BC V6T 1Z2
Canada
fan.yang@sauder.ubc.ca

Yi Qian
Sauder School of Business
University of British Columbia
2053 Main Mall
Vancouver, BC V6T 1Z2
and NBER
yi.qian@sauder.ubc.ca

Hui Xie
Department of Biostatistics
School of Public Health
University of Illinois at Chicago
huixie@uic.edu

Causal inference is central to many problems faced by academics and practitioners, and becomes increasingly important as rapidly-available observational data in this digital era promise to offer real-world evidence on cause-and-effect relationships for better decision makings. However, a prominent challenge faced by empirical researchers to draw valid causal inferences from these data is the presence of endogenous regressors that are correlated with the structural error in the population regression model representing the causal relationship of interest. For example, omitted variables such as ability would cause endogeneity of schooling when examining schooling's effect on wages (Angrist and Krueger 1991).

Regressor endogeneity poses great empirical challenges to researchers and demands special handling of the issue in order to draw valid causal inferences. One classical method to deal with the endogeneity issue is using instrumental variables (IV). The ideal IV has to meet two requirements: it is correlated with the endogenous regressor via an explainable and validated relationship (i.e., relevance restriction), yet uncorrelated with the structural error (i.e., exclusion restriction). Although the theory of IVs is well-developed, researchers often face the challenge of finding good IVs satisfying these two requirements. Potential IVs often suffer from either weak relevance or challenging justification for exclusion restriction, which hampers using IVs to correct for the underlying endogeneity concerns (Rossi 2014).

To address the lack of suitable IVs, there has been a growing interest in developing and applying IV-free endogeneity-correction methods. Several instrument-free approaches have been developed, including identification via higher moments (Lewbel 1997), heteroscedasticity (Rigobon 2003, Hogan and Rigobon 2003), and latent instrumental variables (Ebbes et al. 2005). All three IV-free methods decompose the endogenous regressor into an exogenous part and an endogenous part. The assumption of the endogenous regressor containing an exogenous component is akin to the stringent condition of exclusion restriction for IVs, and thus can be difficult to justify.

Park and Gupta (2012) propose an alternative instrument-free method that uses the copula model (Danaher 2007; Danaher and Smith 2011; Christopoulos, McAdam, and Tzavalis

2021) to capture the regressor-error dependence.[1] Compared with the three IV-free methods above, their copula method does not impose the exogeneity assumption as it directly models the association between the structural error and the endogenous regressor via copula. Furthermore, the copula method can handle discrete endogenous regressors better than other IV-free methods. These features considerably increase the feasibility of endogeneity correction, as evidenced by the rapidly increasing use of the copula correction method (see examples of recent applications in the next section on literature review). However, similar to other IV-free methods, the copula correction methods also require the distinctiveness between the distributions of the endogenous regressor and the structural error. This means that the endogenous regressor is required to have a non-normal distribution for model identification with the commonly assumed normal structural error distribution (Park and Gupta 2012; Papies, Ebbes, and Van Heerde 2017; Becker, Proksch, and Ringle 2021; Haschka 2022; Eckert and Hohberger 2022; Qian, Xie, and Koschmann 2022). Furthermore, we show that the existing copula control function correction method implicitly requires all exogenous regressors to be uncorrelated with the linear combination of copula transformations of endogenous regressors (henceforth referred to as copula control function (CCF)) used to control for endogeneity, and may yield significant bias when there are noticeable correlations between the CCF and exogenous regressors.

In practical applications, both requirements of sufficient regressor non-normality and no correlation between CCF and exogenous regressors can be too strong, and pose significant challenges and limitations for applying the copula correction method. We often encounter endogenous regressors or include transformations of endogenous variables as regressors that have close-to-normality distributions. Examples of such regressors in economics and marketing management studies include stock market returns (Sorescu, Warren, and Ertekin 2017), corporate social responsibility (Eckert and Hohberger 2022), the organizational intelligent

[1] In statistics, a copula is a multivariate cumulative distribution function where the marginal distribution of each variable is a uniform distribution on $[0, 1]$. Copulas permit modeling dependence without imposing assumptions on marginal distributions.

quotient (Mendelson 2000), and the logarithm of price (see Figure 4 in the Application). Theoretically, the endogenous regressor and the structural error can contain a common set of unobservables that collectively have a normal distribution, which can lead to a close-to-normal distribution of the endogenous regressor. In these situations, even if the model is identified asymptotically, close-to-normality of endogenous regressors can cause estimation bias even in moderate sample size and can require a large sample size to mitigate the finite-sample bias (Becker, Proksch, and Ringle 2021). Correlations between the CCF and exogenous regressors are also quite common in practical applications, especially when the exogenous regressors are included to control for observed confounders. Examples of such exogenous control variables abound in marketing and management studies, such as customer-specific variables (age, household size, income, past purchase behaviors, etc.) when estimating the returns of consumer targeting strategies on product sales (Papies, Ebbes, and Van Heerde 2017) and firms' similarity when estimating the effect of competition on innovation (Aghion et al. 2005). Although regressor normality or insufficient regressor non-normality leads to more severe identification issues, including model non-identifiability and poor finite sample performance (Table 1), correlations between CCF and exogenous regressors may occur more frequently than close-to-normality of endogenous regressors. Thus, we consider the two requirements of sufficient regressor non-normality and no correlation between CCF and exogenous regressors as being equally stringent, which call for more general and flexible copula correction methods that relax both requirements.

In this paper, we develop a generalized two-stage copula endogeneity correction method, denoted as 2sCOPE, that relaxes the above two requirements. Similar to the existing copula methods, 2sCOPE requires neither IVs nor the assumption of exclusion restriction. The 2sCOPE method corrects for endogeneity by adding residuals, obtained from regressing latent copula data for each endogenous regressor on the latent copula data for the exogenous regressors, as generated regressors in the structural regression model. Unlike the original copula method (Park and Gupta 2012; Becker, Proksch, and Ringle 2021; Eckert and Hohberger 2022, henceforth denoted as Copula$_{\text{Origin}}$), 2sCOPE can account for the dependence between endogenous and exogenous regressors. Under a Gaussian copula model for the endogenous regressors, correlated exogenous regressors and the structural error, we prove that 2sCOPE can identify causal effects under weaker assumptions than Copula$_{\text{Origin}}$ and overcomes the above two key limitations of Copula$_{\text{Origin}}$ as shown in Table 1. Copula$_{\text{Origin}}$ can be viewed as a special case of 2sCOPE. To demonstrate the benefits of 2sCOPE, we also consider a benchmark method, denoted as COPE, which is a direct extension of Copula$_{\text{Origin}}$ and corrects endogeneity by adding latent copula data themselves as generated regressors.

The contributions of this work are three folds. *First*, to our knowledge, this work is among the first in the literature to provide formal proofs for theoretical properties of copula correction methods. These theoretical results are needed because model identifiability is central to address the endogeneity issue. Recent work notes the lack of rigorous proofs of required model identification conditions and estimation properties (consistency and efficiency) for copula correction as one major area requiring further research (Becker, Proksch, and Ringle 2021; Haschka 2022)[2]. The theoretical results presented here can fill in this important knowledge gap, and contribute to better understanding of the properties of the copula correction methods and guiding their use.

Two novel theoretical findings emerge from this study. First, we identify an implicit assumption required for Copula$_{\text{Origin}}$ to yield consistent estimation, and provide conditions to verify this implicit assumption. This helps improve the effectiveness of the rapidly adopted method for addressing the endogeneity issue. A useful result is that the existence of the correlations between endogenous and exogenous regressors alone does not automatically introduce bias to Copula$_{\text{Origin}}$. Instead, we show that the implicit assumption is the uncorrelatedness of the exogenous regressors with the CCF, the *linear combination* of copula transformations of endogenous regressors used to control for endogeneity. The difference between the im-

[2] For instance, owing to the complex form of the estimation method, Haschka (2022) notes the lack of theoretical proofs of required model identification conditions and estimation consistency as one limitation of the copula correction method developed there, and thus has to rely solely on simulation studies to evaluate its empirical properties.

Features	Park and Gupta (2012)	Haschka (2022)	2sCOPE
Nonnormality of Endogenous Regressors[1]	Required	Required	Not Required[2]
No Correlated Exogenous Regressors[3]	Required	Not Required	Not Required
Intercept Included	YES	NO[4]	YES
Theoretical Proof	YES	NO	YES
Estimation Method	Control Function & MLE	MLE	Control Function
Structural Model	Linear Regression RCL Slope Endogeneity	LPM-FE	Linear Regression LPM-FE, LPM-RE, LPM-ME RCL, Slope Endogeneity

Table 1: A Comparison of Copula Correction Methods

Note: [1]: When required, normality of any endogenous regressor leads to non-identifiable models. Insufficient non-normality of endogenous regressors can also cause poor finite sample performance (finite sample bias and large standard errors) and require extremely large sample size to perform well.
[2]: Non-normality of endogenous regressors is not required as long as at least one correlated exogenous regressor is not normally distributed.
[3]: In our paper, correlated exogenous regressors refer to those exogenous regressors correlated with the CCF (copula control function) used to control for endogeneity.
[4]: The approach cannot estimate the intercept term, which is removed from the panel model prior to estimation using first-difference or fix-effects transformation (Web Appendix A.8 of Haschka (2022)). Becker, Proksch, and Ringle (2021) shows the importance of including intercept in marketing applications.
LPM: Linear Panel Model; FE: Fixed Effects for individual-specific intercepts with common slope coefficients; RE: Random Effects; ME: Mixed-Effects (including both fixed-effects and random coefficients); RCL: Random Coefficient Logit

plicit assumption and the condition of zero pairwise correlations between endogenous and exogenous regressors can be substantial, especially with multiple endogenous regressors.[3] We prove that the proposed 2sCOPE yields consistent causal-effect estimates when the implicit assumption above is violated, which can cause biased causal effect estimates for $\text{Copula}_{\text{Origin}}$.

The second novel finding of our theoretical investigation is as follows. Although the

[3] Although Haschka (2022) explains why correlated regressors can cause potential bias for $\text{Copula}_{\text{Origin}}$, no condition of when bias can occur is given. Specifically, it is possible that with multiple endogenous regressors, the CCF is uncorrelated with exogenous regressors when pairwise correlations between endogenous and exogenous regressors are non-zeros. Even if there is only one endogenous regressor and CCF reduces to be proportional to the copula transformation of the endogenous regressor, the correlation coefficient is not invariant to nonlinear transformations and thus changes after the copula transformation of the endogenous regressor (Danaher and Smith 2011).

exogenous regressors that are correlated with the CCF require special handling for consistent causal-effect estimation, we prove that they can be exploited by 2sCOPE to relax the model identification requirement of non-normality of endogenous regressors. Furthermore, we prove that when both COPE and 2sCOPE methods yield consistent estimates, 2sCOPE improves the efficiency (i.e., precision) of the structural model estimation by exploiting the correlations between the endogenous and exogenous regressors. The efficiency gain is substantial and can be up to ∼50% in our empirical evaluation, meaning that sample size can be reduced by ∼50% to achieve the same estimation efficiency.

Second, the proposed 2sCOPE method is the first copula-correction method that simultaneously relaxes the non-normality assumption of endogenous regressors and handles correlated endogenous and exogenous regressors (Table 1). Existing copula correction methods do not account for correlated endogenous and exogenous regressors. An exception is Haschka (2022), which generalizes Park and Gupta (2012) to fixed-effects linear panel models with correlated regressors by jointly modeling the structural error, endogenous and exogenous regressors using copulas and maximum likelihood estimation (MLE). However, as noted in Haschka (2022), Haschka's approach still requires non-normality of endogenous regressors. Thus, all existing copula correction methods require sufficient non-normality assumption of endogenous regressors for model identification (Park and Gupta 2012; Haschka 2022; Becker, Proksch, and Ringle 2021; Eckert and Hohberger 2022); even when the model is identified, insufficient regressor non-normality can cause significant finite sample bias in sample size of less than 2,000 (Haschka 2022; Becker, Proksch, and Ringle 2021; Eckert and Hohberger 2022). Becker, Proksch, and Ringle (2021) suggest a minimum absolute skewness of 2 for an endogenous regressor to ensure good performance of Gaussian copula correction methods in sample size of less than 1000 (Figure 8 in Becker, Proksch, and Ringle 2021). These requirements can significantly limit the use of copula correction methods in practical applications.

Our proposed 2sCOPE method overcomes these important restrictions of existing copula correction methods. First, we prove that the structural model with normally distributed endogenous regressors can be identified using the 2sCOPE method as long as one of the exogenous regressors correlated with endogenous ones is nonnormal, which is considerably more feasible in many practical applications. Second, consistent with the above theoretical result, our evaluation in Case 3 of the simulation studies demonstrates superior finite-sample performance of 2sCOPE and shows that 2sCOPE eliminates or substantially reduces the significant problem of finite sample bias due to insufficient regressor non-normality raised in Becker, Proksch, and Ringle (2021) and Eckert and Hohberger (2022). In fact, even when the endogenous regressor is normal or close-to-normal with skewness of 0, 2sCOPE is still capable of reducing substantial estimation bias to be negligible for sample size as small as 200 as shown in Figure 2. Third, we develop a novel bootstrap re-sampling method to detect and quantify the finite sample bias due to insufficient regressor non-normality. The bootstrap method directly informs the specific size of finite sample bias when applying 2sCOPE to the data at hand, and thus complements the indirect diagnosis method such as tests of normality or skewness of endogenous regressors. These merits of 2sCOPE have important implications in practical applications. In our empirical application, results from comparison with the IV method and from applying the bootstrap method show that 2sCOPE eliminates the large finite sample bias caused by insufficient non-normality of the endogenous regressor (logarithm of the price). Overall, the proposed 2sCOPE method can greatly broaden the applicability of the instrument-free methods for dealing with endogeneity issues in practice.

Third, 2sCOPE employs generated regressors to address endogeneity. Despite that the vast majority of applications of the copula correction method have used the generated-regressor approach (Becker, Proksch, and Ringle 2021; Eckert and Hohberger 2022), no copula control function method exists that can handle endogenous regressors having insufficient non-normality and being correlated with exogenous regressors. The proposed 2sCOPE overcomes this hurdle and provides a versatile and feasible copula control function method to handle regressor endogeneity. By including generated regressors in the structural model to control endogeneity, 2sCOPE enjoys a number of benefits associated with using the control function to address endogeneity as compared with the alternative MLE approach, such as incurring little extra computational and modeling burden to be integrated with complex outcome models, broader applicability with weaker assumptions, and increased robustness to model mis-specifications.[4] We demonstrate that 2sCOPE retains and enhances these desirable properties of the control function approach for a range of commonly used models in marketing studies, as shown in Table 1. In many of these models, the MLE approach becomes much more difficult or computationally infeasible, but 2sCOPE is straightforward. We present an example with Footnote 9 showing that extending the MLE approach of Haschka (2022) to random coefficient linear panel models (RC-LPMs) with correlated endogenous and exogenous regressors requires numerically evaluating potentially high-dimensional integrals of complicated functions containing the product of copula density functions evaluated at repeated measurement occasions, whereas 2sCOPE involves none of these integrals and can be implemented using standard software programs for RC-LPMs assuming all regressors are exogenous. Furthermore, although 2sCOPE assumes the normal error distribution, we show that 2sCOPE is robust to symmetric non-normal error distributions (Web Appendix E.4), in contrast to the sensitivity to these error mis-specifications for the existing copula methods (Becker, Proksch, and Ringle 2021). Thus, the 2sCOPE control function approach together with correlated exogenous regressors included in 2sCOPE can increase robustness to model mis-specifications. Last, the generated-regressor approach facilitates studying theoretical properties of the 2sCOPE procedure.

The remainder of this paper unfolds as follows. It begins with a review of the related literature on methods for causal inference with endogenous regressors. We then propose the 2sCOPE method that identifies causal effects under weaker assumptions than Copula$_{\text{Origin}}$, providing theoretical proofs for the consistency of 2sCOPE as well as for efficiency gain and model identifiability with normally distributed regressors under 2sCOPE. Next, we evaluate the performance of the proposed 2sCOPE method using simulation studies and compare it

[4] As shown in Becker, Proksch, and Ringle (2021), Gaussian copula control function approach is more robust against error term mis-specifications than the Gaussian copula MLE approach.

with $\text{Copula}_{\text{Origin}}$ and its direct extension COPE under different scenarios. We further apply the proposed 2sCOPE method to estimate price elasticity using store purchase databases. Proofs and a set of robustness checks are in the Web Appendix.

LITERATURE REVIEW

The marketing, economics, and statistics literatures develop a rich set of methods to draw causal inferences. The gold standard to estimate causal effects is randomized assignment such as controlled lab experiments and field experiments (Johnson, Lewis, and Nubbemeyer 2017, Anderson and Simester 2004, Godes and Mayzlin 2009). When controlled experiments are not feasible, quasi-experimental designs such as regression discontinuity, difference-in-difference, and synthetic control are used to mimic randomized experiments and to enable the identification of causal effects with observational data (Hartmann, Nair, and Narayanan 2011, Narayanan and Kalyanam 2015, Athey and Imbens 2006, Shi et al. 2017, Kim, Lee, and Gupta 2020). However, these quasi-experimental designs have special data and design requirements, and are not aimed for coping with the general issue of endogenous regressors when estimating causal effects using observational data.

There is a large literature focusing on approaches to addressing endogenous regressors when inferring causal effects. Papies, Ebbes, and Van Heerde (2017), Rutz and Watson (2019), and Park and Gupta (2012) provide an overview of addressing endogeneity in marketing. Three broad classes of solutions are discussed, and the most commonly used solution is the instrumental variable approach (Kleibergen and Zivot 2003, Qian 2008, Novak and Stern 2009, Ataman, Van Heerde, and Mela 2010, Van Heerde et al. 2013, Li and Ansari 2014). Angrist and Krueger (2001) and Rossi (2014) provide a survey of literature that uses the instrumental variables approach. Rossi (2014) surveyed 10 years of publications in *Marketing Science* and *Quantitative Marketing and Economics*, revealing that the most commonly used instrumental variables are lagged variables, costs, fixed effects and Hausman-style variables from other markets. However, the survey found that the strength of the instruments is rarely measured or reported, which is needed to detect the weak instrument problem. Moreover, one generally cannot test the exclusion restriction condition and verify the validity of instruments. The survey also found that most papers lack a discussion of why the instruments used are valid. In a word, though the theory of instrumental variables is well-developed, good instruments are difficult to find, making the IV approach hard to implement in practice.

The second class of solutions to mitigate endogeneity is to specify the economic structure that generates the observational data including endogenous regressors (e.g., a supply-side model for marketing-mix variables). Doing so allows researchers not only to recover parameters of interest and make causal inferences, but also to perform counterfactual analysis (Chintagunta et al. 2006). Some other examples of this approach in the marketing literature are Berry (1994), Sudhir (2001), Yang, Chen, and Allenby (2003), Sun (2005), Dotson and Allenby (2010) and Otter, Gilbride, and Allenby (2011). The key concern with this approach is that the performance highly depends on model assumptions of supply side. Incorrect assumptions or insufficient information of the supply-side can lead to biased estimates (Chintagunta et al. 2006, Hartmann, Nair, and Narayanan 2011).

The third class of solutions in the domain of endogeneity correction is instrument-free methods. This is a more recent stream of methodological development. Three extant instrument-free approaches are discussed in Ebbes, Wedel, and Böckenholt (2009): the higher moments (HM) approach (Lewbel 1997), the identification through heteroscedasticity (IH) estimator (Rigobon 2003, Hogan and Rigobon 2003), the latent instrumental variables (LIV) method (Ebbes et al. 2005). Recently Wang and Blei (2019) proposed a deconfounder approach that has some flavor of the LIV approach. All these approaches divide the endogenous regressor P into an endogenous and an exogenous part, $P = f(Z) + v$, where $f(Z)$ is treated as an exogenous random variable with unique structures imposed for model identification in different methods. However, the assumption of $f(Z)$ being exogenous is hard to guarantee. Park and Gupta (2012) introduce another instrument-free method that doesn't require the exogeneity of $f(Z)$. It directly models the association between the structural error and the endogenous regressor via copula.

The copula method has been rapidly adopted by researchers to deal with the endogeneity problem because of its feasibility to use in that no instruments are needed (Becker, Proksch, and Ringle 2021; Haschka 2022; Eckert and Hohberger 2022; Qian, Xie, and Koschmann 2022). For example, the copula method has been used to study the effects of marketing activities such as promotion, advertising and loyalty programs (Datta, Foubert, and Van Heerde 2015, Keller, Deleersnyder, and Gedenk 2019); to study product design and brand equity (Heitmann et al. 2020); to study sales force training (Atefi et al. 2018); to study healthy food consumption (Elshiewy and Boztug 2018). Haschka (2022) develops an MLE method that extends Park and Gupta (2012) to linear panel models with fixed-effect intercepts and constant slope coefficients in the presence of correlated regressors. In our paper, we delineate the precise and verifiable condition for $\text{Copula}_{\text{Origin}}$ to yield consistent estimates with correlated endogenous and exogenous regressors. When this condition fails, we develop a new two-stage endogeneity correction method using copula control functions (2sCOPE) that simultaneously relaxes two key assumptions imposed in existing copula correction methods: (1) all endogenous regressors must have non-normal distributions and (2) exogenous regressors must be uncorrelated with the CCF used to control for the endogeneity. We provide proofs of the theoretical properties of 2sCOPE, and derive 2sCOPE for a variety of types of structural models, including the random coefficients models commonly used in marketing studies. As a result, 2sCOPE is applicable in more general settings with the capability to exploit exogenous regressors to improve model identification and estimation.

METHODS

In this section, we develop a copula-based instrument-free method to handle endogenous regressors with insufficient non-normality and correlated with exogenous regressors. We first review the $\text{Copula}_{\text{Origin}}$ method and show that $\text{Copula}_{\text{Origin}}$ implicitly assumes no correlations between the exogenous regressors and the CCF, as well as the bias in the structural model parameter estimates that may arise from the violation of the assumption. Then we present a new proposed method to deal with the problem and the detailed estimation procedure. We also show how exogenous regressors correlated with endogenous regressors can sharpen structural model parameter estimates and permit the identification of the structural model containing normally distributed endogenous regressors, known to cause the model non-identifiability issue for $\text{Copula}_{\text{Origin}}$.

Assumptions of Existing Copula Endogeneity-Correction Method (Copula$_{Origin}$)

Consider the following linear structural regression model with an endogenous regressor and a vector of exogenous regressors [5]:

$$Y_t = \mu + P_t\alpha + W_t'\beta + \xi_t, \tag{1}$$

where $t = 1, 2, ..., T$ indexes either time or cross-sectional units, Y_t is a (1×1) dependent variable, P_t is a (1×1) continuous endogenous regressor, W_t is a (k×1) vector of exogenous regressors, ξ_t is the structural error term, and (μ, α, β) are model parameters. P_t is correlated with ξ_t, and this correlation generates the endogeneity problem. W_t is exogenous, which means it is not correlated with ξ_t, but can be correlated with the endogenous variable P_t.

The key idea of $\text{Copula}_{\text{Origin}}$ (Park and Gupta 2012) is to use a copula to jointly model the correlation between the endogenous regressor P_t and the error term ξ_t. The advantage of using copula is that marginals are not restricted by the joint distribution. Using information contained in the observed data, marginals of the endogenous regressor and the error term are first obtained respectively. Then the copula model enables researchers to construct a flexible multivariate joint distribution that captures the correlation among variables.

Let $F(P, \xi)$ be the joint cumulative distribution function (CDF) of the endogenous regressor P_t and the structural error ξ_t with marginal CDFs $H(P)$ and $G(\xi)$, respectively. For notational simplicity, we may omit the index t in P_t and ξ_t below when appropriate. According to Sklar's theorem (Sklar 1959), there exists a copula function $C(\cdot, \cdot)$ such that

[5] As shown in Becker, Proksch, and Ringle (2021), it is important to include the intercept term when evaluating the copula correction method.

$$F(P, \xi) = C(H(P), G(\xi)) = C(U_p, U_\xi), \tag{2}$$

where $U_p = H(P)$ and $U_\xi = G(\xi)$, and they both follow uniform(0,1) distributions. Thus, the copula maps the marginal CDFs of the endogenous regressor and the structural error to their joint CDF, and makes it possible to separately model the marginals and correlations of these random variables.

To capture the association between the endogenous regressor P and the error ξ, Park and Gupta (2012) uses the following Gaussian copula for its desirable properties (Danaher 2007; Danaher and Smith 2011):

$$\begin{aligned} F(P, \xi) &= C(U_p, U_\xi) = \Psi_\rho(\Phi^{-1}(U_p), \Phi^{-1}(U_\xi)) \\ &= \frac{1}{2\pi(1-\rho^2)^{1/2}} \int_{-\infty}^{\Phi^{-1}(U_p)} \int_{-\infty}^{\Phi^{-1}(U_\xi)} \exp\left[\frac{-(s^2 - 2\rho \cdot s \cdot t + t^2)}{2(1-\rho^2)}\right] dsdt, \end{aligned} \tag{3}$$

where $\Phi(\cdot)$ denotes the univariate standard normal distribution function and $\Psi_\rho(\cdot, \cdot)$ denotes the bivariate standard normal distribution with the correlation coefficient ρ. With emprical marginal CDFs, the above Gaussian copula model depends on the rank-order of raw data only, and is invariant to strictly monotonic transformations of variables in (P, ξ). Thus, the above Gaussian copula model is considered general and robust for most marketing applications (Danaher and Smith 2011). In the Gaussian copula model, ρ captures the endogeneity of the regressor P, and a non-zero value of ρ corresponds to P being endogenous.

Under the above copula model for (P_t, ξ_t) and the commonly-assumed normal distribution for the structural error ξ_t, Park and Gupta (2012) develop the following generated regressor procedure to correct for regressor endogeneity. Let $P_t^* = \Phi^{-1}(U_p)$ and $\xi_t^* = \Phi^{-1}(U_\xi)$, the above Gaussian copula assumes $[P_t^*, \xi_t^*]'$ follow the standard bivariate normal distribution with the correlation coefficient ρ as follows:

$$\begin{pmatrix} P_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right). \tag{4}$$

Under the assumption that the structural error ξ_t follows $N(0, \sigma_\xi^2)$, Park and Gupta (2012) show that the structural error can be split into two parts as follows:

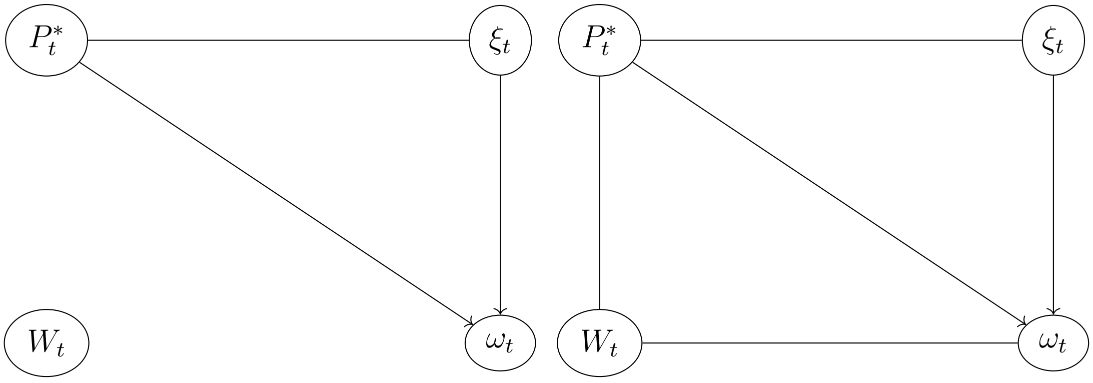
$$\xi_t = \sigma_\xi \xi_t^* = \sigma_\xi \rho P_t^* + \sigma_\xi \sqrt{1-\rho^2}\omega_t, \tag{5}$$

where the first part $\sigma_\xi \rho P_t^*$ captures the correlation between ξ_t and the endogenous regressor, and the other part $\sigma_\xi \cdot \sqrt{1-\rho^2}\omega_t$ being an independent new error term. Equation (1) can be rewritten as follows:

$$Y_t = \mu + P_t\alpha + W_t\beta + \sigma_\xi \cdot \rho \cdot P_t^* + \sigma_\xi \cdot \sqrt{1-\rho^2} \cdot \omega_t. \tag{6}$$

Based on the above representation, Park and Gupta (2012) suggest the following generated regressor approach to correcting for the endogeneity of P_t: the ordinary least squares (OLS) estimation of Equation (6) with $P_t^* = \Phi^{-1}(U_p)$ included as an additional regressor will yield consistent model estimates. Park and Gupta (2012) also pointed out that in order for the above approach to work, P_t needs to have a non-normal distribution. Suppose P_t is normally distributed, $P_t = P_t^* \cdot \sigma_p$, resulting in perfect collinearity between P_t and P_t^* and violating the full rank assumption required for identifying the linear regression model in Equation (6). Thus, P_t should follow a different distribution from the normal error term so that the causal effect of P that is independent of all other regressors can be identified.

However, we show here that an implicit assumption for the above generated regressor approach to yield consistent model estimates is the uncorrelatedness between P_t^* and W_t. For the OLS estimation to yield consistent estimation, the error term ω_t in Equation (6) is required to be uncorrelated with all the regressors on the right-hand side of the equation: P_t, W_t, P_t^*. Figure 1 shows how the correlation between W_t and the new error term ω_t changes when W_t becomes correlated with P_t^*. Absence of a line between two variables in Figure 1 means that the two variables are not correlated. When W_t is not correlated with P_t^*, W_t should also be uncorrelated with ω_t, which is determined by ξ_t and P_t^*, because of the exogenous feature of W_t (Figure 1 (a)). However, when W_t is correlated with P_t^*, it would become correlated with ω_t because (1) ω_t is a linear combination of ξ_t and P_t^* (Equation 5), and (2) W_t is uncorrelated with ξ_t. The induced correlation between the exogenous regressor



(a) No Correlation Between P_t^* and W_t **(b)** Has Correlation Between P_t^* and W_t

Figure 1: Correlation Between W_t and New Error ω_t.

Note: Presence (absence) of a solid line between two variables means the two variables are correlated (uncorrelated). A line without an arrow represents stochastic association between two nodes. A line with an arrow represents a deterministic relationship. Specifically, ω_t is determined jointly by P_t^* and ξ_t.

W_t and the new error term ω_t is intuitively shown in Figure 1 (b) and formally proved in Theorem 1 below. Thus, the correlation between the exogenous regressor W_t and the generated regressor P_t^* would cause biased OLS estimates of Equation (6) using Copula$_{\text{Origin}}$ because of the induced correlation between the error term ω_t and W_t. That is, W_t becomes endogenous in Equation (6) when W_t and P_t^* are correlated.

Theorem 1. *Assuming (1) the error term is normal, (2) a Gaussian Copula for the structural error term and P_t, and (3) P_t is endogenous: $\rho \neq 0$, $Cov(\omega_t, W_t) = -\frac{\rho}{\sqrt{1-\rho^2}} Cov(W_t, P_t^*) \neq 0$ if P_t^* and W_t are correlated.*

Proof: See Web Appendix A.1, Proof of Theorem 1.

To summarize, the generated regressor procedure based on Equation (6) makes the following set of assumptions.

Assumption 1. *The structural error follows a normal distribution;*

Assumption 2. *P_t and the structural error follow a Gaussian copula;*

Assumption 3. *Non-normality of the endogenous regressor P_t;*

Assumption 4. *W_t and P_t^* are uncorrelated.*

As shown in Web Appendix A.2, **Assumption 4** can be extended to **Assumption 4(b)** below for the case of multiple endogenous regressors.

Assumption 4(b). *For multiple endogenous regressors, W_t is uncorrelated with the CCF, i.e., the linear combination of P_t^* used to control for endogenous regressors. Specifically, $Cov(W_t, \frac{\rho_{\xi 1}-\rho_{12}\rho_{\xi 2}}{1-\rho_{12}^2} \cdot P_{1,t}^* + \frac{\rho_{\xi 2}-\rho_{12}\rho_{\xi 1}}{1-\rho_{12}^2} \cdot P_{2,t}^*) = 0$ is required in the 2-endogenous regressors case.*[6]

Assumptions 4 and 4(b) are verifiable and provide users with criteria to check whether $\text{Copula}_{\text{Origin}}$ would provide consistent estimation when there exist exogenous regressors that may be correlated with the CCF. With only one endogenous regressor, one can simply check the correlations between the copula transformation of this endogenous regressor with each exogenous regressor. For multiple endogenous regressors, one should check the correlations between the CCF (i.e., the linear combination of copula transformations of these endogenous regressors used to control for endogeneity) in $\text{Copula}_{\text{Origin}}$ with each exogenous regressor. If there exists at least one exogenous regressor in W_t that fails the Assumptions 4 or 4(b), $\text{Copula}_{\text{Origin}}$ yields biased estimates, and our proposed 2sCOPE can be used, which is derived in the next subsection.

For Assumptions 1 to 3, Park and Gupta (2012) have shown reasonable robustness of their copula method to non-normal distributions of the structural error (Assumption 1) and alternative forms of copula functions (Assumption 2), although it is not surprising to observe sensitivity of $\text{Copula}_{\text{Origin}}$ to gross violations of these assumptions, such as highly skewed error distributions (Eckert and Hohberger 2022). By contrast, the assumption that the endogenous regressor P_t follows a non-normal distribution (Assumption 3) is critical. An endogenous regressor following a normal distribution can cause the model to be unidentifiable regardless of sample size; a nearly normally distributed endogenous regressor may require a very large sample size for the method to perform well and may cause the method to have poor performance for a finite sample size. Moreover, we have shown above that for their method to work best, there should be no exogenous regressors that are correlated with P_t^*

[6] It is clear that this requirement is not the same as either $Cov(W_t, P_{1,t}^*) = 0, Cov(W_t, P_{2,t}^*) = 0$ or $Cov(W_t, P_{1,t}) = 0, Cov(W_t, P_{2,t}) = 0$.

(Assumption 4). Both the Assumptions (3 and 4) can be too strong and substantially limit the applicability of the instrument-free copula method in practice.

Proposed Method: Two-stage Copula Endogeneity-correction (2sCOPE)

In this subsection, we propose a two-stage Copula (2sCOPE) method and show that it can relax both the uncorrelatedness assumption between the copula-transformed endogenous regressor and the exogenous regressors (Assumption 4) and the key identification assumption of non-normality on the endogenous regressors (Assumption 3). The 2sCOPE method jointly models the endogenous regressor, P_t, the correlated exogenous variable, W_t, and the structural error term, ξ_t, using the Gaussian copula model, which implies that $[P_t^*, W_t^*, \xi_t^*]$ follows the multivariate normal distribution:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{p\xi} \\ \rho_{pw} & 1 & 0 \\ \rho_{p\xi} & 0 & 1 \end{bmatrix}\right), \tag{7}$$

where $P_t^* = \Phi^{-1}(H(P_t))$, $W_t^* = \Phi^{-1}(L(W_t))$, and $\xi_t^* = \Phi^{-1}(G(\xi_t))$, and $H(\cdot)$, $L(\cdot)$ and $G(\cdot)$ are marginal CDFs of P_t, W_t and ξ_t respectively.

Under the above Gaussian copula model in Equation (7), one can develop a direct extension of Copula$_{\text{Origin}}$, which adds generated regressors P_t^* and W_t^* into the structural regression model to correct for endogeneity bias (Web Appendix A.3). The resulting method, denoted as COPE, is shown to yield consistent causal effect estimates without requiring Assumption 4 needed for Copula$_{\text{Origin}}$. However, COPE requires endogenous regressors P_t and exogenous regressors W_t to be both non-normally distributed (Theorem A1 in Web Appendix A.3). To overcome the limitations of COPE, below we derive the 2sCOPE method that relaxes both assumptions and is shown to be more efficient than COPE.

Under the above Gaussian copula model, we have the following system of equations that are similar to two-stage least-squares method using IVs. However, we do not require any variable that satisfies the exclusion restriction.

$$Y_t = \mu + P_t\alpha + W_t\beta + \xi_t \tag{8}$$

$$P_t^* = W_t^*\gamma + \epsilon_t, \tag{9}$$

where the two error terms ϵ_t and ξ_t are correlated because of the endogeneity of P_t. Under the assumption that ξ_t follows a normal distribution, ϵ_t and ξ_t follow a bivariate normal distribution, since they are a linear combination of tri-normal variate (ξ_t^*, P_t^*, W_t^*) under the Gaussian copula assumption. Equation (9) expresses the copula transformation of the endogenous regressor, determined by the rank-order of P_t, as a linear combination of observed and unobserved variables.

The main idea of 2sCOPE is to make use of the fact that, by conditioning on ϵ_t, the structural error term ξ_t becomes independent of both P_t and W_t. That is, by conditioning on the component of P_t causing the endogeneity of P_t (i.e, ϵ_t here), the structural error is not correlated with both P_t and W_t, thereby ensuring the consistency of standard estimation methods. In this sense, ϵ_t serves as a (scaled) control function to address the endogeneity bias. [7] To demonstrate this point, note that the Gaussian copula model in Equation (7) can be rewritten as follows:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{pw} & \sqrt{1-\rho_{pw}^2} & 0 \\ \rho_{p\xi} & \frac{-\rho_{pw}\rho p\xi}{\sqrt{1-\rho_{pw}^2}} & \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} \end{pmatrix} \cdot \begin{pmatrix} \omega_{1,t} \\ \omega_{2,t} \\ \omega_{3,t} \end{pmatrix},$$

$$\begin{pmatrix} \omega_{1,t} \\ \omega_{2,t} \\ \omega_{3,t} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right). \tag{10}$$

Given the above joint normal distribution for (P_t^*, W_t^*, ξ_t^*) and $\xi_t^* = \sigma_\xi \xi_t$, we have

[7] An alternative approach to obtaining ϵ_t is assuming $P_t = W_t\gamma + \epsilon_t$, where (ξ_t, ϵ_t) follow a Gaussian Copula model and $\epsilon_t = P_t - W_t\gamma$. This approach is based on the same idea that the first stage error term ϵ_t captures the endogenous part of P_t that is uncorrelated with exogenous regressors, but uses the copula transformation of ϵ_t, $\Phi^{-1}(H(\epsilon_t))$, as the generated regressor. Identification under this model involves a different set of modeling assumptions. In particular, the assumption of (ξ_t, ϵ_t) following a Gaussian Copula model means the model becomes unidentifiable if the first-stage error term ϵ_t is normally distributed, which is not uncommon in practice.

$$P_t^* \;=\; \rho_{pw} W_t^* + \sqrt{(1-\rho_{pw}^2)} \cdot \omega_{2,t} = \rho_{pw} W_t^* + \epsilon_t, \tag{11}$$

which shows γ in Equation (9) is ρ_{pw} and $\epsilon_t = \sqrt{(1-\rho_{pw}^2)} \cdot \omega_{2,t}$, and

$$\begin{aligned}
Y_t &= \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* + \frac{-\sigma_\xi \rho_{pw} \rho_{p\xi}}{1-\rho_{pw}^2} W_t^* + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t} \\
&= \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} (P_t^* - \rho_{pw} W_t^*) + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}, \\
&= \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} \epsilon_t + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}. \qquad (12)
\end{aligned}$$

Equation (12) suggests adding the estimate of the error term ϵ_t from the first stage regression as a generated regressor to the outcome regression instead of using P_t^* and W_t^*. As shown in Theorem 2, the new error term $\omega_{3,t}$ is uncorrelated with all the regressors in Equation (12), ensuring the consistency of model estimates. This two-step procedure, named as 2sCOPE, adds the first-stage residual term $\widehat{\epsilon}_t$ to control for endogeneity and in this aspect is similar to the control function approach of Petrin and Train (2010). However, unlike Petrin and Train (2010), 2sCOPE requires no use of instrumental variables.

Theorem 2. ***Estimation Consistency.*** *Assuming (1) the error is normal, (2) the endogenous regressor P_t or correlated regressors W_t is nonnormal, and (3) a Gaussian Copula for the error, P_t and W_t, $Cov(\omega_{3,t}, W_t) = Cov(\omega_{3,t}, P_t) = Cov(\omega_{3,t}, \epsilon_t) = 0$ in Equation (12).*
Proof: See Web Appendix B.1, Proof of Theorem 2.

According to Theorem 2, the proposed method 2sCOPE can yield consistent estimates when assumptions are met. Specifically, Assumption 4 is relaxed because 2sCOPE can handle the case when the model includes exogenous regressors correlated with the endogenous regressor. Theorem 3 below further show that 2sCOPE relaxes Assumption 3 (the non-normality assumption on endogenous regressors), a critical model identification condition required in all other copula correction methods.

Theorem 3. ***Non-normality Assumption Relaxed.*** *Assuming (1) the error term is normal, (2) one of the correlated exogenous regressors W_t is nonnormal, and (3) a Gaussian*

Copula for the error term, P_t and W_t, 2sCOPE estimator $\widehat{\theta}_{\mathbf{2}}$ is consistent when P_t follows a normal distribution while the COPE estimator $\widehat{\theta}_{\mathbf{1}}$ is not consistent.

Proof: See Web Appendix B.2, Proof of Theorem 3.

Theorem 3 shows that as long as one of the exogenous regressors that are correlated with the endogenous regressor P_t is nonnormally distributed, 2sCOPE can correct for endogeneity for a normal regressor P_t while COPE cannot. Intuitively, when P_t (or W_t) is normal, P_t^* (or W_t^*) becomes a linear function of P_t (or W_t) under the Gaussian copula assumption, rendering COPE to fail the full rank assumption and become unidentified. Thus, COPE cannot deal with normal endogenous/exogenous regressors. For 2sCOPE in Equation (12), adding the first stage residual $\widehat{\epsilon}_t$ as the generated regressor improves model identification. As long as not all W_t are normal, $\widehat{\epsilon}_t$ would not be a linear function of P_t and W_t and thus the second stage model (Equation 12) in 2sCOPE would satisfy the full rank requirement for model identification. Thus, 2sCOPE can relax the non-normality assumption on the endogenous regressor required in Park and Gupta (2012) as long as one of the W_t is non-normally distributed.

Theorem 4 below shows that when both COPE and 2sCOPE yield consistent estimates, 2sCOPE outperforms COPE, the direct extension of $\text{Copula}_{\text{Origin}}$ to more general settings, by reducing the variance of the estimates and improving estimation efficiency.

Theorem 4. ***Variance Reduction.*** *Assuming (1) the error term is normal, (2) the endogenous variable P_t and correlated regressors W_t are nonnormal, and (3) a Gaussian Copula for the error term, P_t and W_t, $\mathbf{Var}(\widehat{\theta}_{\mathbf{2}}) \leq \mathbf{Var}(\widehat{\theta}_{\mathbf{1}})$, where $\widehat{\theta}_1$ and $\widehat{\theta}_2$ denote parameter estimates from COPE and 2sCOPE, respectively.*

Proof: See Web Appendix B.3, Proof of Theorem 4.

To sum up, we have proved the consistency of 2sCOPE (Theorem 2). Theorems 3 and 4 further establish that the 2sCOPE method outperforms the COPE method, the extended $\text{Copula}_{\text{Origin}}$, in terms of estimation efficiency gain and relaxing the non-normality assumption on the endogenous regressors required in $\text{Copula}_{\text{Origin}}$ by satisfying a very loose condition.

Multiple Endogenous Regressors

In this subsection, we extend 2sCOPE to the general case of multiple endogenous regressors. Consider the following structural linear regression model with two endogenous regressors ($P_{1,t}$ and $P_{2,t}$) that are potentially correlated with the exogenous regressor W_t:

$$Y_t = \mu + P_{1,t} \cdot \alpha_1 + P_{2,t} \cdot \alpha_2 + W_t\beta + \xi_t. \tag{13}$$

Under the multivariate Gaussian distribution assumption on $(\xi_t, P^*_{1,t}, P^*_{2,t}, W^*_t)$, the system of equations for the 2sCOPE method in Equations (8, 9) are readily extended to the case with two endogenous regressors as

$$Y_t = \mu + P_{1,t}\alpha_1 + P_{2,t}\alpha_2 + W_t\beta + \xi_t, \tag{14}$$

$$P^*_{1,t} = \rho_{wp1}W^*_t + \epsilon_{1,t}, \tag{15}$$

$$P^*_{2,t} = \rho_{wp2}W^*_t + \epsilon_{2,t}, \tag{16}$$

where Equations (15) and (16) can be directly derived from the Gaussian copula assumption; $(\xi_t, \epsilon_{1,t}.\epsilon_{2,t})$ are a linear transformation of $(\xi_t, P^*_{1,t}, P^*_{2,t}, W^*_t)$, and thus also follow a multivariate Gaussian distribution. As a result, we can decompose the structural error ξ_t as additive terms for $\epsilon_{1,t}$, $\epsilon_{2,t}$ and a remaining independent error term $\omega_{4,t}$ as follows

$$Y_t \;\; = \mu + P_{1,t}\alpha_1 + P_{2,t}\alpha_2 + W_t\beta + \eta_1\epsilon_{1,t} + \eta_2\epsilon_{2,t} + \sigma_\xi \cdot m \cdot \omega_{4,t}, \tag{17}$$

where $\epsilon_{1,t} = P^*_{1,t} - \rho_{wp1}W^*_t$ and $\epsilon_{2,t} = P^*_{2,t} - \rho_{wp2}W^*_t$, m is a constant depending only on the correlation coefficients in the Gaussian copula, η_1, η_2 and $\omega_{4,t}$ are the same as those defined in Equation (W6) in Web Appendix A.3 for describing COPE for multiple endogenous regressors and thus the new (scaled) error term $\omega_{4,t}$ is independent of latent copula data $(P^*_{1,t}, P^*_{2,t}, W^*_t)$ as well as all functions of these latent data including $P_{1,t}, P_{2,t}, W_t, \epsilon_{1,t}, \epsilon_{2,t}$. Because $\omega_{4,t}$ is independent of all regressors on the right-hand side of Equation (17), the OLS estimation of Equation (17) yields consistent estimation of structural model parameters. Note that Equation (17) can also be obtained from Equation (W7) in Web Appendix A.3 for describing COPE for multiple endogenous regressors by noting that $\epsilon_{1,t} = P^*_{1,t} - \rho_{wp1}W^*_t$

and $\epsilon_{2,t} = P^*_{2,t} - \rho_{wp2} W^*_t$. However, 2sCOPE adds only two residual terms $(\epsilon_{1,t}, \epsilon_{2,t})$ as generated regressors instead of three copula transformations of regressors $(P^*_{1,t}, P^*_{2,t}, W^*_t)$ as generated regressors, as COPE does (Equation W7 in Web Appendix A.3). Thus, 2sCOPE adds a smaller number of generated regressors than COPE, and provides higher estimation efficiency. In addition, by adding residual terms as the generated regressors, 2sCOPE relaxes the assumption of regressor non-normality required by COPE as long as not all W_t are normal. The proof for the estimation consistency of 2sCOPE, estimation efficiency gain and relaxation of the regressor-nonnormality assumption for 2sCOPE can be found in Web Appendix B under the related Theorems 2, 3, 4.

$\text{Copula}_{\text{Origin}}$	COPE	2sCOPE
• The structural error follows a normal distribution (Asm. 1); • P_t and the structural error follow a Gaussian copula (Asm. 2); • All regressors in P_t are nonnormally distributed (Asm. 3); • W_t is uncorrelated with the CCF (copula control function which is the linear combination of all P^*_t used to control for endogeneity) (Asm. 4, 4(b)).	• The structural error follows a normal distribution; • P_t, W_t and the structural error follow a Gaussian copula; • All regressors in P_t and W_t are nonnormally distributed.	• The structural error follows a normal distribution; • P_t, W_t and the structural error follow a Gaussian copula; • P_t can be normally distributed as long as one of W_t is nonnormal.

Table 2: Summary of Assumptions for the Three Methods

Table 2 summarizes the assumptions for the three methods: our proposed 2sCOPE method, the existing copula method $\text{Copula}_{\text{Origin}}$ and $\text{Copula}_{\text{Origin}}$'s direct extension, COPE. The 2sCOPE can handle the case when there are exogenous regressors that are correlated with endogenous regressors. Moreover, 2sCOPE can further relax the regressor-nonnormality assumption. Table 3 summarizes the estimation procedures of COPE and 2sCOPE.

COPE	2sCOPE
	Stage 1:
• Obtain empirical CDFs for each regressor in P_t and W_t, denoted as $\widehat{H}(P_t)$ and $\widehat{L}(W_t)$;	• Obtain empirical CDFs for each regressor in P_t and W_t, $\widehat{H}(P_t)$ and $\widehat{L}(W_t)$;
• Compute $P_t^* = \Phi^{-1}(\widehat{H}(P_t))$ and $W_t^* = \Phi^{-1}(\widehat{L}(W_t))$;	• Compute $P_t^* = \Phi^{-1}(\widehat{H}(P_t))$ and $W_t^* = \Phi^{-1}(\widehat{L}(W_t))$;
• Add P_t^* and W_t^* to the outcome structural regression model as generated regressors.	• Regress each endogenous regressor in P_t^* separately on W_t^* and obtain residual $\widehat{\epsilon}_t$;
	Stage 2:
	• Add $\widehat{\epsilon}_t$ to the outcome structural regression model as generated regressors.
• Standard errors of parameter estimates are estimated using bootstrap in both methods.	

Table 3: Estimation Procedure

2sCOPE for Random Coefficient Linear Panel Models

We consider the following random coefficient model for linear panel data

$$Y_{it}|\mu_i, \alpha_i, \beta_i \quad = \quad \bar{\mu} + \mu_i + P_{it}'\alpha_i + W_{it}'\beta_i + \xi_{it}, \tag{18}$$

where $i = 1, \cdots, N$ indexes cross-sectional units and $t = 1, \cdots, T$ indexes occasions. P_{it} (W_{it}) denotes a vector of endogenous (exogenous) regressors. P_{it} and W_{it} can be correlated. The error term $\xi_{it} \overset{iid}{\sim} N(0, \sigma_\xi^2)$, which is correlated with P_{it} due to the endogeneity of P_{it} but is uncorrelated with the exogenous regressors in W_{it}. The individual-specific intercept μ_i and individual-specific slope coefficients (α_i, β_i) permit heterogeneity in both intercepts and regressor effects across cross-sectional units. Extant marketing studies have shown the ubiquitous presence of heterogeneous consumers' responses to marketing mix variables (e.g., price sensitivity) and substantial bias associated with ignoring such heterogeneity in slope coefficients. Thus, it is important to permit individual-specific slope coefficients.

The linear panel data model as specified in Equation (18) is general and includes the linear panel model with only individual-specific intercepts considered in Haschka (2022) as a special case. Specifically, Haschka (2022) fixes (α_i, β_i) to be the same value (α, β) across all units, assuming all cross-sectional units have the same slope coefficients. In contrast, the model in Equation (18) relaxes this strong assumption and can generate unit-specific slope parameters, which can be used for targeting purposes.

A random coefficient model typically assumes $(\mu_i, \alpha_i, \beta_i)$ follows a multivariate normal distribution. When all regressors are exogenous, estimation algorithms for such random coefficient models are well-established and computationally feasible even for a high-dimensional vector of random effects $(\mu_i, \alpha_i, \beta_i)$: with the normal conditional distribution for $Y_{it}|(\mu_i, \alpha_i, \beta_i)$ in Equation (18) and the multivariate normal prior distribution for random effects $(\mu_i, \alpha_i, \beta_i)$, marginally Y_{it} follows a normal distribution with a closed-form expression containing no integrals with respect to random effects $(\mu_i, \alpha_i, \beta_i)$, leading to an easy-to-evaluate likelihood function (Greene 2003). For instance, R function `lme()` can be used to obtain MLEs of population parameters and empirical Bayes estimates of individual random effects. Alternatively, one can assume a mixed-effect model where μ_i is a fixed effect parameter with μ_i's allowed to be correlated with the regressors P_{it} and W_{it}. To avoid the potential incidental parameter problem, one often uses the first-difference or fixed-effects transformation to eliminate the incidental intercept parameters as follows

$$\tilde{y}_{it}|\alpha_i, \beta_i \quad = \quad \tilde{P}_{it}'\alpha_i + \tilde{W}_{it}'\beta_i + \tilde{\xi}_{it}, \tag{19}$$

where $\tilde{y}_{it}$, $\tilde{P}_{it}$, $\tilde{W}_{it}$ and $\tilde{\xi}_{it}$ denote new variables obtained from the first-difference or fixed-effect transformation. Haschka (2022) considered a special case of Equation (19) by fixing (α_i, β_i) to be constants.

It is straightforward to apply 2sCOPE to address regressor endogeneity in the general random coefficient model for linear panel data in Equation (18) and the transformed one without intercepts in Equation (19).[8] Assuming $(P_{it}, W_{it}, \xi_{it})$ follow a Gaussian copula, COPE adds

[8] Similar to Haschka (2022), a GLS transformation can be applied to both sides of Equation (19), resulting in a pooled regression for which 2sCOPE can be directly applied.

the generated regressor $P_{it}^* = \Phi^{-1}(\widehat{H}(P_{it}))$ and $W_{it}^* = \Phi^{-1}(\widehat{L}(W_{it}))$ into Equation (18) to control for regressor endogeneity. The 2sCOPE procedure adds the residuals obtained from regressing P_{it}^* on W_{it}^*. Thus, 2sCOPE method can be implemented using standard software programs for random coefficient linear panel models assuming all regressors are exogenous (see Simulation Study Case 4) for an illustration using the R function `lme()`). By contrast, the MLE approach for copula correction in the random coefficients model accounting for correlated endogenous and exogenous regressors is not available yet and would require constructing complicated joint likelihood on the error term, P_t and W_t, which involves newly appearing numerical integrals with respect to random effects and cannot be maximized by standard estimation algorithms for random coefficient models.[9] Finally, current applications applying Copula$_{\text{Origin}}$ do not consider the role of exogenous regressors. Our analysis shows that this may yield bias if any exogenous regressor is correlated with the CCF added to control endogeneity, for which 2sCOPE should be used to address regressor endogeneity.

2sCOPE for Slope Endogeneity and Random Coefficient Logit Model

In Web Appendices C and D, we derive the 2sCOPE method to tackle the slope endogeneity problem and address endogeneity bias in random coefficient logit models with correlated and normally distributed regressors. In these two cases, we show how to apply 2sCOPE to correct for the endogeneity bias, which can avoid the potential bias of Copula$_{\text{Origin}}$ due to the potential correlations between the exogenous regressors and CCF, as well as make use of the correlated exogenous regressors to relax the non-normality assumption of endogenous regressors, improve model identification and sharpen model estimates. As shown there, 2sCOPE can be implemented using standard estimation methods by adding generated regressors to control for endogenous regressors. By contrast, the maximum likelihood approach can re-

[9] With endogenous regressors, the individual random effects parameters enter into both the density function for the outcome $Y_{it}|(\mu_i, \alpha_i, \beta_i)$ and the density of copula function $C(U_{\xi,it}, U_{P,it}, U_{W,it})$ via $U_{\xi,it}$, and thus cannot be integrated out in closed-form any more from the likelihood function even with the normal structural error term and normal random effects. Therefore, numerical integration is required for obtaining MLEs in random coefficient models with endogenous regressors, which cannot be performed with standard software programs for random coefficient model estimation.

quire constructing a complicated joint likelihood that is not what the standard estimation method uses and thus requires separate development and significantly more computation involving numerical integration.

SIMULATION STUDY

In this section, we conduct Monte Carlo simulation studies for the following goals: (a) to assess the performance of the proposed method for correlated regressors, (b) to assess the performance of the proposed method under regressor normality and near normality, (c) to assess the performance of the proposed method under various types of structural models, and (d) to assess the robustness of the proposed method to violations of model assumptions. Following Park and Gupta (2012), we measure the estimation bias using t_{bias} calculated as the ratio of the absolute difference between the mean of the sampling distribution and the true parameter value to the standard error of the parameter estimate. As defined above, t_{bias} represents the size of bias relative to the sampling error.

Case 1: Non-normal Regressors

We first examine the case when P and W are correlated. The data-generating process (DGP) is summarized below:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{p\xi} \\ \rho_{pw} & 1 & 0 \\ \rho_{p\xi} & 0 & 1 \end{bmatrix} \right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right), \tag{20}$$

$$\xi_t = G^{-1}(U_{\xi,t}) = G^{-1}(\Phi(\xi_t^*)) = \Phi^{-1}(\Phi(\xi^*)) = 1 \cdot \xi_t^*, \tag{21}$$

$$P_t = H^{-1}(U_{P,t}) = H^{-1}(\Phi(P_t^*)), \quad W_t = L^{-1}(U_{W,t}) = L^{-1}(\Phi(W_t^*)), \tag{22}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + \xi_t = 1 + 1 \cdot P_t + (-1) \cdot W_t + \xi_t, \tag{23}$$

where ξ_t^* and P_t^* are correlated ($\rho_{p\xi} = 0.5$), generating the endogeneity problem; W_t^* is exogenous and uncorrelated with ξ_t^*; W_t^* and P_t^* are correlated ($\rho_{pw} = 0.5$), and thus W_t and P_t are correlated. We consider four different estimation methods: (1) OLS, (2) $\text{Copula}_{\text{Origin}}$

in the form of Equation (6), (3) the extended method COPE in the form of Equation (W3) in Web Appendix A, and (4) the proposed 2sCOPE in the form of Equation (12). We set the sample size $T = 1000$, and generate 1000 data sets as replicates using the DGP above. In the simulation, we use the gamma distribution $Gamma(1, 1)$ with shape and rate equal to 1 for P_t and the exponential distribution $Exp(1)$ with rate 1 for W_t. Models are estimated on all generated data sets, providing the empirical distributions of parameter estimates.

			OLS			$\text{Copula}_{\text{Origin}}$			COPE			2sCOPE		
ρ_{pw}	Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
0.5	μ	1	0.689	0.045	6.964	1.231	0.081	2.849	1.012	0.093	0.129	1.009	0.059	0.157
	α	1	1.571	0.036	15.75	1.055	0.069	0.791	0.985	0.072	0.213	0.986	0.070	0.197
	β	-1	-1.259	0.031	8.236	-1.289	0.031	9.169	-0.997	0.067	0.038	-0.995	0.042	0.123
	$\rho_{p\xi}$	0.5	-	-	-	0.570	0.047	1.504	0.505	0.055	0.090	0.504	0.038	0.097
	σ_ξ	1	0.862	0.020	6.902	1.011	0.043	0.244	1.008	0.041	0.206	1.006	0.040	0.143
	D-error			-			-			0.002613			0.001614	
0.7	μ	1	0.730	0.041	6.629	1.307	0.076	4.037	1.011	0.085	0.124	1.005	0.053	0.088
	α	1	1.800	0.041	19.67	1.260	0.068	3.838	0.988	0.078	0.148	0.991	0.075	0.118
	β	-1	-1.529	0.037	14.21	-1.567	0.037	15.36	-0.997	0.071	0.041	-0.994	0.056	0.110
	$\rho_{p\xi}$	0.5	-	-	-	0.633	0.043	3.130	0.503	0.057	0.048	0.500	0.026	0.000
	σ_ξ	1	0.799	0.018	11.18	0.980	0.044	0.468	1.007	0.041	0.160	1.003	0.040	0.084
	D-error			-			-			0.002902			0.001760	

Table 4: Results of the Simulation Study Case 1: Non-normal Regressors

Note: Mean and SE denote the average and standard deviation of parameter estimates over all the 1,000 simulated samples.

Table 4 reports estimation results. As expected, OLS estimates of both α and β are biased ($t_{bias} = 15.75/8.24$) due to the regressor endogeneity. $\text{Copula}_{\text{Origin}}$ reduces the bias, but still shows significant bias for the coefficient estimates of P_t and W_t. The bias of $\text{Copula}_{\text{Origin}}$ depends on the strength of the correlation between W and P. Stronger correlations between P^* and W^* can cause a larger bias of $\text{Copula}_{\text{Origin}}$ estimates. For example, when the correlation between W^* and P^* increases from 0.5 to 0.7, the bias of estimated α increases by around five times (from 0.055 to 0.260 in Table 4 under the column "$\text{Copula}_{\text{Origin}}$"). The bias confirms our derivation in the model section, demonstrating that using the existing copula method may not solve the endogeneity problem completely with correlated regressors.

The proposed 2sCOPE method provides consistent estimates without using instruments. The average estimates of $\rho_{p\xi}$ is close to the true value 0.5 and is significantly different from 0, implying regressor endogeneity detected correctly using 2sCOPE. Moreover, 2sCOPE shows greater estimation efficiency. The standard error of $\alpha(\beta)$ in 2sCOPE is 0.070 (0.042), which is 2.78% (37.31%) smaller than the corresponding standard errors using COPE. We further calculate the estimation precision of COPE and 2sCOPE using the D-error measure $|\Sigma|^{1/K}$ (Arora and Huber 2001, Qian and Xie 2022), where Σ is the covariance matrix of the regression coefficient estimates, and K is the number of explanatory variables in the structural model. A smaller D-error means greater estimation efficiency and improved estimation precision. When $\rho_{pw} = 0.5$, the D-error measure is 0.002613 for COPE and 0.001614 for 2sCOPE (Table 4), and thus 2sCOPE increases estimation precision by 38.2%, meaning that for 2sCOPE to achieve the same precision with COPE, sample size can be reduced by 38.2%. A 39.3% of efficiency gain for 2sCOPE is observed for $\rho_{pw} = 0.7$ (Table 4).

We perform a further simulation study for a small sample size. Specifically, we use the same DGP as described above to generate synthetic data, except with the sample size T=200. Web Appendix E Table W1 reports the results and shows that OLS estimates have endogeneity bias and Copula$_{\text{Origin}}$ reduces the endogeneity bias but significant bias remains. Our proposed method, 2sCOPE, performs well and has unbiased estimates for the small sample size T=200. The efficiency gain of 2sCOPE relative to COPE appears to be greater when sample size becomes smaller. When the correlation between P^* and W^* is 0.5, the D-error measures are 0.0166 and 0.0091 for COPE and 2sCOPE (Web Appendix Table W1), respectively, meaning that 2sCOPE increases estimation precision by 1-0.0091/.0166=46% compared with COPE, and thus sample size can be reduced by almost a half (~50%) for 2sCOPE to achieve the same estimation precision as that achieved by COPE. A similar magnitude of efficiency gain for 2sCOPE relative to COPE (~50%) is observed when the correlation between P^* and W^* is 0.7 (Web Appendix Table W1).

Case 2: Normal Regressors

Next, we examine the case when the endogenous regressor and (or) the correlated exogenous regressor are normally distributed. We pay special attention to this case because normality is not allowed for endogenous regressors in Park and Gupta (2012). We use the same DG as described in Equations (20) to (23) to generate the data, except that the marginal CDFs for regressors, $H(\cdot)$ and $L(\cdot)$, are chosen according to the distributions listed in the first two columns in Table 5.

Table 5 summarizes the estimation results. As expected, OLS estimates are biased. $\text{Copula}_{\text{Origin}}$ produces biased estimates whenever the endogenous regressor P follows a normal distribution. The estimates of $\text{Copula}_{\text{Origin}}$ are biased when P follows a gamma distribution (first row of Table 5) for a different reason: P and W are correlated. Similar to $\text{Copula}_{\text{Origin}}$, the COPE estimators are biased in all the three scenarios when either P_t or W_t is normal. When W_t is normal, β is 0.323 away from the true value -1; when P_t is normally distributed, α is 0.684 away from the true value; when both P_t and W_t are normal, α is 0.663 away from the true value 1 and β is 0.324 away from the true value -1. This is expected because COPE adds P_t^* and W_t^*, the copula transformation of regressors, as additional regressors, and will cause perfect co-linearity and model non-identification problem whenever at least one of these regressors is normally distributed.

By contrast, the proposed 2sCOPE method provides consistent estimates as long as P_t and W_t are not both normally distributed. Both α and β are tightly distributed near the true value whenever P_t or W_t is nonnormally distributed. Unlike $\text{Copula}_{\text{Origin}}$ and COPE, 2sCOPE adds the residual term obtained from regressing P_t^* on W_t^* as the generated regressor. Thus, as long as P_t and W_t are not both normally distributed, the residual term is not perfectly co-linear with the original regressors, permitting model identification. Only when both P_t and W_t are normally distributed (the last scenario in Table 5), the residual term added into the structural regression model becomes a linear combination of P_t and W_t, causing perfect

Distribution				OLS			Copula$_{\text{Origin}}$			COPE			2sCOPE		
P	W	Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
Gamma	Normal	μ	1	0.431	0.045	12.63	1.018	0.078	0.227	1.017	0.080	0.217	1.015	0.077	0.190
		α	1	1.569	0.037	15.40	0.979	0.070	0.302	0.979	0.070	0.296	0.985	0.070	0.212
		β	-1	-1.259	0.030	8.619	-1.333	0.028	11.78	-1.323	0.433	0.746	-0.997	0.045	0.067
		$\rho_{p\xi}$	0.5	-	-	-	0.640	0.039	3.556	0.589	0.141	0.631	0.506	0.036	0.151
		σ_{ξ}	1	0.861	0.019	7.240	1.064	0.046	1.394	1.135	0.162	0.837	1.005	0.038	0.134
Normal	Exp	μ	1	1.286	0.042	6.777	1.286	0.045	6.374	0.994	0.073	0.081	1.023	0.070	0.334
		α	1	1.628	0.031	20.36	1.532	0.462	1.152	1.684	0.437	1.568	1.048	0.126	0.381
		β	-1	-1.286	0.032	8.956	-1.287	0.032	8.960	-0.992	0.066	0.127	-1.024	0.062	0.383
		$\rho_{p\xi}$	0.5	-	-	-	0.089	0.419	0.980	-0.167	0.384	1.738	0.465	0.074	0.473
		σ_{ξ}	1	0.829	0.018	9.492	0.940	0.151	0.394	0.981	0.151	0.129	0.980	0.063	0.318
Normal	Normal	μ	1	1.001	0.026	0.046	1.002	0.030	0.052	1.001	0.033	0.024	1.002	0.028	0.057
		α	1	1.668	0.030	22.38	1.663	0.450	1.474	1.663	0.460	1.441	1.655	0.395	1.657
		β	-1	-1.335	0.029	11.44	-1.335	0.029	11.42	-1.324	0.438	0.740	-1.328	0.197	1.668
		$\rho_{p\xi}$	0.5	-	-	-	0.006	0.412	1.198	0.001	0.412	2.426	0.010	0.303	1.616
		σ_{ξ}	1	0.816	0.019	9.687	0.917	0.155	0.534	1.003	0.211	0.016	0.879	0.092	1.317

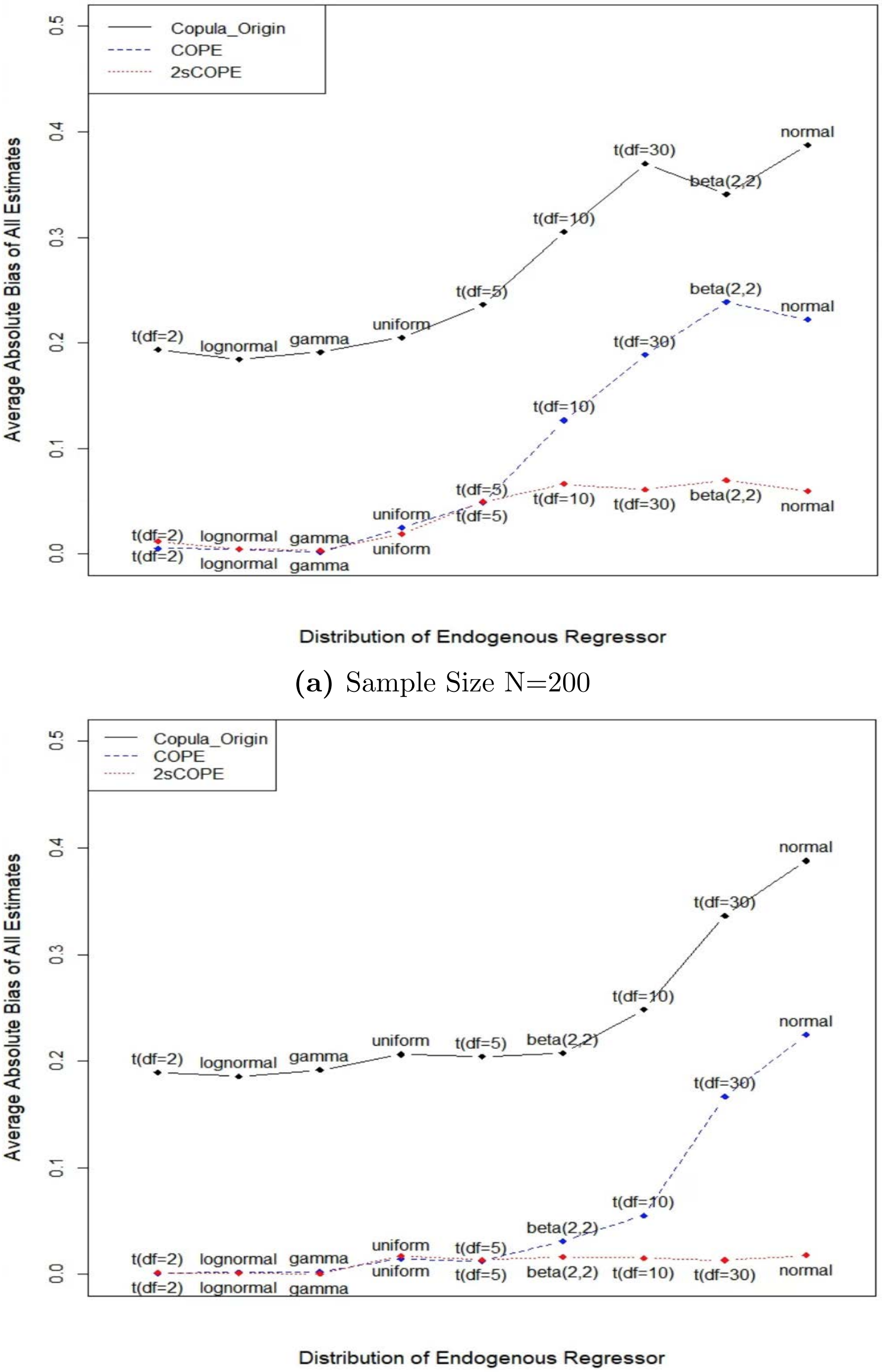
Table 5: Results of Case 2: Normal Regressors

co-linearity and model non-identification. Overall, this simulation study demonstrates the capability of the proposed 2sCOPE to relax the non-normality assumption in Copula$_{\text{Origin}}$ as long as one of P_t and W_t is nonnormally distributed.

Case 3: Insufficient Non-Normality of Endogenous Regressors

The above case shows that the proposed 2sCOPE can deal with normal endogenous regressors, while Copula$_{\text{Origin}}$ and COPE cannot. In this case, we examine the performance of these methods in the more common situation of close-to-normal regressors. Although models are identified asymptotically (i.e., infinite sample size), appreciable finite sample bias can occur with realistic sample size commonly seen in marketing studies, if the endogenous regressor is too close to a normal distribution (Becker, Proksch, and Ringle 2021; Haschka 2022; Eckert and Hohberger 2022). Becker, Proksch, and Ringle (2021) suggest a minimum absolute skewness of 2 for an endogenous regressor in order for Copula$_{\text{Origin}}$ to have good performance in sample size less than 1000. This requirement can significantly limit the use of copula correction methods in practical applications. Given that 2sCOPE can handle normal endogenous regressors, we expect that 2sCOPE can handle much better the finite sample bias caused by insufficient regressor non-normality than the existing copula correction methods. Thus, in this case, we examine the finite sample performance of those methods when the distribution of the endogenous regressor has various closeness to normality. We use the DGP as described in Equations (20) to (23) to generate data, except that the marginal CDF for the endogenous regressor ($H(\cdot)$) is varied from some common distributions with varying closeness to normality. Specifically, we consider uniform, log normal, t, gamma, beta and normal distributions, and use the average absolute estimation bias of all the regression parameters (μ, α, β) in the structural model to measure the performance.

Figure 2 plots the estimation bias with different distributions of the endogenous regressor P. Results show estimates of Copula$_{\text{Origin}}$ are biased with correlated endogenous and exogenous regressors, consistent with our theoretical proof. COPE performs well when P has sufficient non-normality ($t(2)$, log normal, gamma) and has no bias even for sample size as small as 200. However, COPE cannot handle a normal endogenous regressor and yields large estimation bias that remains unchanged as the sample size increases, consistent with our theoretical proof and the simulation result in Case 2. Furthermore, COPE suffers from finite-sample bias when the endogenous regressor P has distributions with insufficient non-normality (e.g., beta(2,2), $t(\text{df} = 30)$). Moreover, the estimation bias of COPE is larger when the sample size is smaller or the distribution of the endogenous regressor P is closer to normal. For instance, t-distribution with a degree of freedom 30 is closer to normal than the t distribution with degree of freedom 10, 5 and 2, resulting in larger estimation bias. For $t(\text{df} = 30)$ which is very close to normal, increasing sample size from T=200 to 1000 barely change the size of estimation bias. By contrast, our proposed 2sCOPE method yields consistent estimates for all normal and close-to-normal regressor distributions and has negligible finite sample bias even for sample size as small as 200 (bias $<$ 5% of parameter values).



(a) Sample Size N=200

(b) Sample Size N=1000

Figure 2: Average absolute estimation bias of all the regression parameters (μ, α, β) in the structural model for different distributions of endogenous regressor.

Note: 'lognormal' is lognormal(0,1), 'uniform' is U[0,1], and 'gamma' is $Gamma(1, 1)$.

Case 4: Random Coefficient Linear Panel Model

We investigate the performance of 2sCOPE in random coefficient linear panel model. We use the copula function and marginal distributions of $[P_{it}, W_{it}, \xi_{it}]$ as specified in Case 1 (Equations 20-22). We assign $\rho_{pw} = 0.7$ as an example. We then generate the outcome Y_{it} using the following standard random coefficient linear panel model:

$$Y_{it} = \bar{\mu} + \mu_i + P_{it}(\bar{\alpha} + a_i) + W_{it}(\bar{\beta} + b_i) + \xi_{it} = 1 + \mu_i + P_{it}(1 + a_i) + W_{it}(-1 + b_i) + \xi_{it},$$

where $[\mu_i, a_i, b_i] \sim N(0, I_3)$, $t = 1, ..., 50$ indexes occasions for repeated measurements, and $i = 1, ..., 500$ indexes the individual units. The above random coefficients model permits individual units to have heterogeneous baseline preferences (μ_i) and heterogeneous responses to regressors (a_i, b_i). Such random coefficients models are frequently used in marketing studies to capture individual heterogeneity and to profile and target individuals. The correlation between ξ_{it} and P_{it} creates the regressor endogeneity problem, which can cause biased estimates for standard linear random coefficient estimation methods ignoring the regressor-error correlation. We generate individual-level panel data as described above for 1000 times and use the data for estimation. Estimation results are in Table 6. LME is the standard estimation method for linear mixed models assuming all regressors are exogenous, as implemented in the R function `lme()`. LME and Copula$_{\text{Origin}}$ are biased because of endogeneity and correlated exogenous regressors, respectively. Our proposed method 2sCOPE provides unbiased estimates that are tightly distributed around the true values for all parameters.

Additional Results on Robustness Checking

Web Appendix E provides simulation results on a small sample size (E.1), model estimation with multiple endogenous regressors (E.2) and multiple exogenous control covariates (E.3), and the robustness of 2sCOPE to mis-specifications of the structural error distribution and the copula dependence structure (E.4 & E.5). Overall, these results demonstrate 2sCOPE performs well in these situations and is robust to reasonable violations of normal error and Gaussian copula assumptions.

		LME			Copula$_{\text{Origin}}$			COPE			2sCOPE		
Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
$\bar{\mu}$	1	0.722	0.046	6.052	1.314	0.049	6.399	1.001	0.054	0.016	1.004	0.048	0.091
$\bar{\alpha}$	1	1.853	0.045	18.83	1.293	0.045	6.469	1.000	0.045	0.009	1.000	0.046	0.008
$\bar{\beta}$	-1	-1.557	0.045	12.39	-1.598	0.044	13.56	-0.996	0.048	0.079	-1.000	0.044	0.005
σ_μ	1	0.985	0.033	0.459	0.982	0.033	0.547	0.985	0.033	0.463	0.984	0.031	0.522
σ_α	1	0.988	0.036	0.326	0.987	0.034	0.397	0.986	0.035	0.403	0.989	0.035	0.316
σ_β	1	0.993	0.031	0.235	0.992	0.033	0.249	0.992	0.031	0.264	0.992	0.033	0.248
$\rho_{p\xi}$	0.5	-	-	-	0.646	0.009	16.33	0.509	0.012	0.757	0.507	0.005	1.365
σ_ξ	1	0.794	0.004	57.71	0.957	0.010	4.439	0.985	0.009	1.689	0.985	0.009	1.640

Table 6: Results of the Simulation Study Case 4: Random Coefficient Linear Panel Model

Note: $\sigma_\mu, \sigma_\alpha, \sigma_\beta$ are standard deviations of μ_i, a_i, b_i.

EMPIRICAL APPLICATION

In this section, we apply our method to a real marketing application. We illustrate the proposed method to address the price endogeneity issue using store-level sales data of toothpaste category in Chicago over 373 weeks from 1989 to 1997 [10]. To control for product size, we select toothpaste with the most common size, which is 6.4 oz. Retail price is usually considered endogenous. The endogeneity of retail price can come from unmeasured product characteristics or demand shocks that can influence both consumers' and retailers' decisions. Since these variables are unobserved by researchers, they are absorbed into the structural error, leading to the endogeneity problem. Prices of different stores are correlated and often used as an IV for each other. This allows us to test the performance of the proposed 2sCOPE method in an empirical setting where a good IV exists. Besides the endogenous price, two promotion-related variables, bonus promotion and direct price reduction, would also affect demand. Following Park and Gupta (2012), we treat the promotion variables as exogenous regressors. We focus on category sales in two large stores in Chicago (referred to as Stores

[10] We obtained the data from https://www.chicagobooth.edu/research/kilts/datasets/dominicks.

1 and 2). We convert retail price, in-store promotion and sales from UPC level to aggregate category level. They are computed as weekly market share-weighted averages of UPC-level variables. The correlation between log retail price and bonus promotion in Store 1 (Store 2) is -0.30 (-0.15), and the correlation between log retail price and price reduction promotion in Store 1 (Store 2) is -0.23 (-0.35). Both the correlations are significantly different from zero. The appreciable correlations between price and promotion variables actually provide a good setting for testing our method and examining the impact that our proposed method can make in the setting of correlated endogenous and exogenous regressors. The moderate sample size (T=373) also provides an opportunity to compare finite sample performance of different copula correction methods in the presence of potentially insufficient regressor non-normality in real data. Summary statistics of key variables are summarized in Table 7.

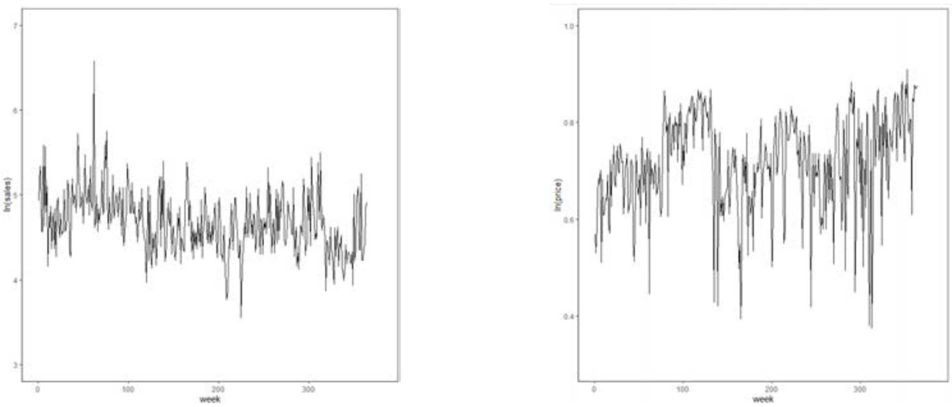
	Store 1				Store 2			
Variables	Mean	SD	Max	Min	Mean	SD	Max	Min
Sales (Unit)	115	52.8	720	35	165.7	93.7	1334	26
Price ($)	2.06	0.20	2.48	1.46	2.10	0.21	2.48	1.47
Bonus	0.18	0.20	0.80	0.00	0.16	0.19	0.79	0.00
PriceRedu	0.10	0.19	0.72	0.00	0.10	0.19	0.73	0.00

Table 7: Summary Statistics

We estimate the following linear regression model:

$$\log(\text{Sales}_t) = \beta_0 + \log(\text{Retail Price}_t) \cdot \beta_1 + W_t'\beta_2 + \xi_t,$$

where $t = 1, 2, ..., T$ indexes week. The vector W_t includes all exogenous regressors, which are two promotion variables, bonus promotion and price reduction, in this application. Figure 3 shows log sales and log retail prices of toothpaste at store 1 over time (store 2 is very similar). To control for the possible trend of retail price over time, we use detrended log retail prices (and for instrumental variables as well) for estimation below. Figure 4 shows the histograms of detrended log retail prices and the two promotion variables. All the



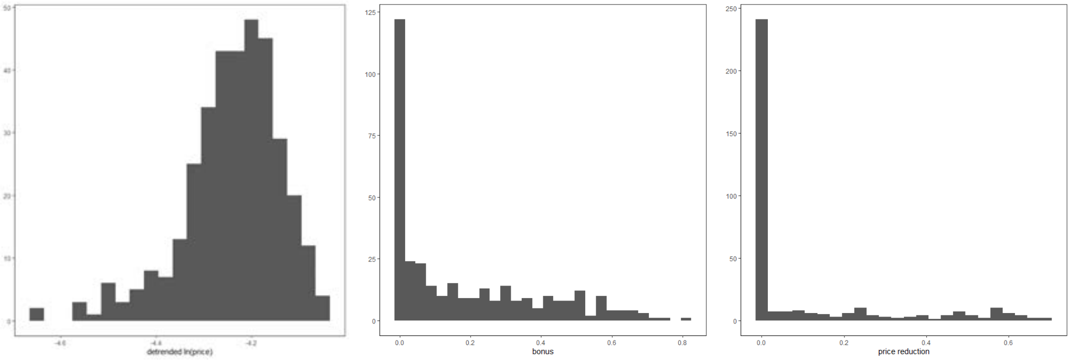
(a) Store 1 log sales

(b) Store 1 log retail price

Figure 3: Log Sales and Log Retail Price of Toothpaste in Store 1.

three variables are continuous variables. Moreover, except log retail price, which is a bit close to normal distribution, the other two regressors, bonus and price reduction, are both nonnormally distributed. Therefore, we expect that the proposed 2sCOPE method can exploit these additional features of exogenous regressors correlated with the endogenous regressor for model identification and estimation even if the endogenous regressor has a close-to-normal distribution. We estimate the model using the OLS, two-stage least-squares (TSLS), $\text{Copula}_{\text{Origin}}$, COPE, and our proposed 2sCOPE method.

We use the IV-based TSLS estimator as a benchmark to test the validity of our proposed method. Following Park and Gupta (2012), we use retail price at the other store as an instrument for price. This variable can be a valid instrument as it satisfies the two key requirements. First, retail prices across stores in a same market can be highly correlated because wholesale prices are usually offered the same (or very close). The Pearson correlation between the detrended log retail prices at Stores 1 and 2 is 0.79, providing strong explanatory power on the endogenous price. The correlation is comparable to that in Park and Gupta (2012). Second, some unmeasured product characteristics such as shelf-space allocation, shelf location and category location are determined by retailers and are usually not systematically related to wholesale prices (exclusion restriction). For the three copula-based methods, we make use of information from the existing endogenous and exogenous regressors and no



(a) detrended log price **(b)** bonus **(c)** price reduction

Figure 4: Histogram of Log Retail Price, Bonus and Price Reduction in Store 1

extra IVs are needed. In Copula$_{\text{Origin}}$, we add the copula transformation of the detrended log price, $\text{logP}^* = \Phi^{-1}(\widehat{H}(\text{logP}))$, as a "generated regressor" to the outcome regression. For the COPE method, we add another two "generated regressors", copula transformation of bonus and price reduction ($\text{Bonus}^* = \Phi^{-1}(\widehat{L}_1(\text{Bonus}))$, $\text{PriceRedu}^* = \Phi^{-1}(\widehat{L}_2(\text{PriceRedu}))$). For the 2sCOPE method, we first regress logP^* on Bonus^* and PriceRedu^*, and then add the residual as the only "generated regressor" to the outcome regression. $\widehat{H}(\cdot), \widehat{L}_1(\cdot), \widehat{L}_2(\cdot)$ are all estimated CDFs using the univariate empirical distribution for each regressor. Standard errors of parameter estimates are obtained using bootstrap.

Table 8 reports the estimation results. Beginning with the results from Store 1, OLS estimates are significantly different from TSLS estimates, indicating that the price endogeneity issue occurs. Instrumenting for retail price changes the price coefficient estimate from -0.767 to -1.797, implying that there is a positive correlation between unobserved product characteristics and the price. The estimates of ρ in the three IV-free copula-based methods, representing the correlation between the endogenous regressor P_t and the error term, are all significantly positive, further confirming our previous conclusion. This direction of correlation is consistent with previous empirical findings (e.g., Villas-Boas and Winer 1999, Chintagunta, Dubé, and Goh 2005). The price elasticity estimates from the Copula$_{\text{Origin}}$, the extension COPE and the proposed method 2sCOPE are -3.082, -3.111 and -2.014, respectively. Among the three estimates, the estimate of -2.014 from the proposed 2sCOPE is close

		OLS			TSLS			$\text{Copula}_{\text{Origin}}$			COPE			2sCOPE		
Store	Parameters	Est	SE	t-value	Est	SE	t-value	Est	SE	t-value	Est	SE	t-value	Est	SE	t-value
Store 1	Constant	1.301	1.197	0.25	-2.993	1.646	1.82	-8.526	2.619	3.26	-8.569	2.820	3.04	-3.908	2.314	1.69
	Price	-0.767	0.288	2.66	-1.797	0.396	4.54	-3.082	0.620	4.97	-3.111	0.664	4.69	-2.014	0.555	3.63
	Bonus	0.371	0.122	3.31	0.104	0.141	0.74	0.415	0.115	3.61	0.522	0.288	1.81	0.064	0.171	0.37
	PriceRedu	0.498	0.115	4.33	0.285	0.125	2.28	0.544	0.111	4.90	1.033	0.211	4.90	0.275	0.143	1.92
	ρ	-	-	-	-	-	-	0.521	0.098	5.32	0.662	0.117	5.66	0.297	0.089	3.34
Store 2	Constant	-3.898	1.246	3.13	0.763	1.943	0.39	1.107	3.404	0.33	1.324	3.430	0.39	0.001	2.702	0.00
	Price	-1.982	0.300	6.61	-0.864	0.467	1.85	-0.799	0.807	0.99	-0.783	0.811	0.96	-1.048	0.648	1.62
	Bonus	0.062	0.116	0.53	0.286	0.148	1.93	0.032	0.117	0.27	-0.819	0.426	1.92	0.239	0.151	1.58
	PriceRedu	0.283	0.111	2.55	0.540	0.137	3.94	0.275	0.110	2.5	0.540	0.194	2.78	0.467	0.152	3.07
	ρ	-	-	-	-	-	-	-0.319	0.177	1.80	-0.358	0.164	2.18	-0.188	0.109	1.72

Table 8: Estimation Results: Toothpaste Sales

to the estimate of -1.797 from the TSLS method, whereas the existing copula and the COPE yield substantially greater-sized price elasticity estimates. We confirm in the literature that the TSLS and 2sCOPE estimates are reasonable because the price elasticity of toothpaste category is around -2.0 (Hoch et al. 1995, Mackiewicz and Falkowski 2015). Comparing the estimates of ρ from the three IV-free copula-based methods, our proposed 2sCOPE provides a much smaller estimate of ρ (0.297 for 2sCOPE vs 0.521 for $\text{Copula}_{\text{Origin}}$ and 0.662 for COPE in Table 8), consistent with the over-correction in both $\text{Copula}_{\text{Origin}}$ and COPE.

Reasons for the substantial differences in the 2sCOPE estimates from the $\text{Copula}_{\text{Origin}}$ include (1) correlated endogenous and exogenous regressors and (2) the unimodal close-to-normality distribution for the logarithm of price variable, which can lead to inconsistent estimates/poor finite sample performance for $\text{Copula}_{\text{Origin}}$. In fact, the correlations between logP^* and the exogenous regressors are -0.44 for Bonus and -0.26 for PriceRedu, both of which are substantially larger than the corresponding correlations (-0.30 and -0.15, respectively) between logP and the exogenous regressors. The p-value for the null hypothesis of these correlations being zeros are significantly less than 0.05 (< 0.001), indicating a violation of Assumption 4 required for $\text{Copula}_{\text{Origin}}$ to yield consistent estimates.

Reasons for the substantial differences in the 2sCOPE estimates from the COPE method include (1) a uni-modal close-to-normality distribution for the price variable leading to potentially poor finite sample performance of COPE, and (2) loss of estimation precision manifested as larger standard errors of the COPE estimates as compared with those from 2sCOPE. As shown in previous sections, 2sCOPE can relax the non-normality assumption of the endogenous regressor, and yield consistent and efficient estimates even if the endogenous regressor follows a normal or nearly normal distribution. Moreover, 2sCOPE provides estimates with smaller standard error than COPE, which confirms Theorem 4 showing that using two-stage copula estimation reduces estimation variance.

Unlike Store 1, the results from Store 2 indicate that the retail price is not endogenous. First, the estimates of ρ (the correlation between price and the error term) are not significantly different from 0 for both $\text{Copula}_{\text{Origin}}$ and 2sCOPE (t-value $\leq$ 1.96 under columns "$\text{Copula}_{\text{Origin}}$" and "2sCOPE" for Store 2 in Table 8), and only slightly significantly different from 0 for COPE (a t-value of 2.18 under Column "COPE" in Table 8). The estimate of ρ for the COPE, however, is questionable because of the limitations of COPE mentioned in the paragraph above. Second, the estimated price coefficient of OLS is -1.982, which is very close to the estimates of TSLS and 2sCOPE in Store 1 and further confirming no endogeneity of price in Store 2. Overall, the price elasticity estimates from TSLS and the three IV-free copulas-based methods are close to each other for Store 2, and the observed differences between them and the OLS estimate can be attributed to estimation variability incurred from using more complicated models instead of the presence of endogeneity.

Evaluating Finite Sample Performance of Copula Correction Using Bootstrap

In the above, the convergence of results between TSLS and the proposed method 2sCOPE in both stores supports the validity of the proposed method in addressing the endogeneity issue. Moreover, the difference between the estimates in COPE and 2sCOPE in store 1 shows the advantages of 2sCOPE in terms of relaxing the non-normality assumption of the endogenous regressor and estimation efficiency gain by exploiting additional information from correlated exogenous regressors. To further validate the findings of performance difference among different copula correction methods, we perform the following bootstrap re-sampling experiment to evaluate finite sample performance of copula correction.

Bootstrap simulations can be used to evaluate the size of the bias in parameter estimates that may arise when sample size is small to moderate (Efron and Tibshirani 1994, Chap. 10; Hooker and Mentch 2018)[11], even if the estimation performs well when sample size is large. Our proposed bootstrap algorithm for evaluating finite sample performance of copula correction is described in Algorithm 1 below. Specifically, we randomly draw the same number of observations from the underlying copula model and the structural model estimated using the original sample and then perform the OLS, COPE and 2sCOPE estimation on the bootstrap sample as done with the original sample. We repeat this simulation B times, and obtain a distribution for each model coefficient estimate. We then compare the mean of each coefficient estimate's distribution with the corresponding coefficient estimate using the original data, which is the true parameter value in our model-based bootstrap re-sampling. The small-sample bias of a coefficient estimate is the difference between the average coefficient estimate from bootstrap samples and the coefficient estimate from the original sample.

Algorithm 1 A Bootstrap Algorithm for Evaluating Finite-sample Performance of 2sCOPE in Empirical Applications

```
Series Input: data $Y, X, W$, sample size N, $\widehat{\theta}(Y, X, W)-$ 2sCOPE estimates of the structural model parameters, $(\widehat{H}, \widehat{L})-$ empirical CDFs of $X$ and $W$, and $\widehat{\Sigma}-$ Gaussian copula correlation structure estimate.
for $b = 1$ to $B$ do
    Simulate $X_b^*$, $W_b^*$, $\xi_b^*$ from Gaussian Copula $\Psi_{\widehat{\Sigma}}(\Phi^{-1}(U_x), \Phi^{-1}(U_w), \Phi^{-1}(U_\xi))$, sample size=N;
    Obtain $X_b = \widehat{H}^{-1}(\Phi(X_b^*)), W_b = \widehat{L}^{-1}(\Phi(W_b^*))$ and $\xi_b = \widehat{\sigma}_\xi \cdot \xi_b^*$, where $\widehat{\sigma}_\xi$ is the 2sCOPE estimate of the standard deviation of structural error term;
    Obtain $Y_b = f(X_b, W_b, \xi_b, \widehat{\theta}(Y, X, W))$, where $f$ is the linear regression in this setting;
    Obtain the 2sCOPE estimate $\widehat{\theta}_b = \widehat{\theta}(Y_b, X_b, W_b)$ using the $b$th bootstrap sample.
end for
Calculate potential bias of the 2sCOPE estimator: $\frac{1}{B}\sum_{b=1}^{B} \widehat{\theta}_b - \widehat{\theta}(Y, X, W)$.
```

We apply the bootstrap algorithm to our empirical application with the true parameter values set to be the store 1's 2sCOPE estimates reported in Table 8 rounded to the

[11] Note that this bootstrap simulation is different from and should not be confused with the bootstrap method used to obtain standard error estimates of various copula methods in Table 8.

first non-zero number when generating bootstrap samples. The detailed steps to generate these bootstrap samples can be found in Web Appendix F. Table 9 summarizes means and standard deviations of parameter estimates for different estimation methods over the 1000 bootstrap samples, unlike the estimation result on one single observed data set reported in Table 8. The estimation results are broadly consistent with those in Table 8. First, OLS and $\text{Copula}_{\text{Origin}}$ are biased because of the endogeneity and the correlation between the endogenous and exogenous regressors. Second, there are still significant finite-sample biases in the COPE estimates, while the estimates of 2sCOPE are distributed closely to the true values, demonstrating that 2sCOPE performs well in our empirical application and eliminates almost all the finite sample bias of COPE caused by insufficient non-normality of the endogenous regressor (logarithm of price). Third, COPE estimates have substantially greater standard errors than 2sCOPE, demonstrating substantial efficiency gain of 2sCOPE. These differences in the performance of COPE and 2sCOPE confirm the advantage of 2sCOPE in dealing with close-to-normality endogenous regressors, and help explain why the COPE estimates are biased and differ significantly from the 2sCOPE estimates in Table 8.

		OLS			$\text{Copula}_{\text{Origin}}$			COPE			2sCOPE		
Parameters	True	Est	SE	t_{bias}	Est	SE	t_{bias}	Est	SE	t_{bias}	Est	SE	t_{bias}
Constant	-4	1.514	0.777	7.098	-2.377	2.737	0.593	-3.276	2.780	0.261	-3.782	1.619	0.135
Price	-2	-0.678	0.186	7.099	-1.594	0.646	0.628	-1.827	0.657	0.263	-1.946	0.388	0.139
Bonus	0.1	0.458	0.088	4.046	0.465	0.089	4.124	0.107	0.237	0.029	0.113	0.128	0.103
PriceRedu	0.3	0.571	0.089	3.058	0.576	0.089	3.111	0.292	0.138	0.060	0.309	0.112	0.079
ρ	0.3	-	-	-	0.316	0.194	0.080	0.229	0.205	0.345	0.284	0.071	0.222

Table 9: Finite-Sample Performance of Copula Correction. “Est” and “SE” denote the mean and standard deviation of the estimates over 1000 bootstrap samples of Store 1 Data.

CONCLUSION

Causal inference lies at the center of social science research, and observational studies often beg rigorous study designs and methodologies to overcome endogeneity concerns. In this paper, we focus on the instrument-free copula method to handle endogenous regressors. We propose a generalized two-stage copula endogeneity correction (2sCOPE) method that extends the existing copula correction methods (Park and Gupta 2012; Becker, Proksch, and Ringle 2021; Haschka 2022; Eckert and Hohberger 2022) to more general settings. Specifically, 2sCOPE allows exogenous regressors to be correlated with endogenous regressors and relaxes the non-normality assumption on the endogenous regressors. Similar to the original copula correction method (Copula$_{\text{Origin}}$), 2sCOPE corrects endogeneity by adding "generated regressors" derived from the existing regressors and is straightforward to use. However, unlike COPE that is a direct extension to Copula$_{\text{Origin}}$ by adding latent copula transformations of existing regressors, 2sCOPE has two stages. The first stage obtains the residuals from regressing latent copula data for the endogenous regressor on the latent copula data for the exogenous regressors. The second stage uses the first-stage residual as a "generated regressor" in the structural regression model. We theoretically prove that 2sCOPE can yield consistent cause-effect estimates when exogenous regressors are correlated with the endogenous regressors. Moreover, 2sCOPE can relax the non-normality assumption on endogenous regressors and substantially improve the finite sample performance of copula correction.

We evaluate the performance of 2sCOPE via simulation studies and demonstrate its use in an empirical application. The simulation results show that 2sCOPE yields consistent estimates under relaxed assumptions. Moreover, 2sCOPE method outperforms COPE in terms of dealing with normal endogenous regressors and improving estimation efficiency. Endogenous regressors are allowed to be normally distributed as long as one of the exogenous regressors is nonnormally distributed, which is a very weak assumption. The efficiency gain is substantial and can be up to ∼50%, implying that sample size can be reduced by ∼50% to achieve the same estimation efficiency as compared with COPE method that does not exploit the correlations between endogenous and exogenous regressors. Last but not least, our robustness checks show that the proposed method 2sCOPE is reasonably robust to the structural error distributional assumption and non-Gaussian copula correlation structure (Web

Appendix E.4 & E.5). We further apply 2sCOPE to a public dataset in marketing. When dealing with endogenous price, we find that the estimated price coefficient using our proposed 2sCOPE is very close to the TSLS estimate, while OLS and Copula$_{\text{Origin}}$ show large biases. Moreover, the difference between results of 2sCOPE and COPE demonstrates the advantage of 2sCOPE in dealing with (nearly) normal endogenous regressors and improving estimation efficiency. To validate the performance of 2sCOPE in the real application, we further provide a novel bootstrap simulation algorithm to evaluate the finite-sample performance of 2sCOPE, and demonstrate its use in our empirical application.

These findings have rich implications for guiding the practical use of the copula-based instrument-free methods to handle endogeneity. A known critical assumption for Copula$_{\text{Origin}}$ is the non-normality of endogenous regressors. The users of the method in the literature have all been practicing the check and verification of this assumption. However, our work shows that this is insufficient: one also needs to check Assumption 4 for the one-endogenous-regressor case, and Assumption 4(b) for the multiple-endogenous-regressors case. Note that neither assumption is the same as checking the pairwise correlations between the endogenous and exogenous regressors. Assumption 4 evaluates pairwise correlations involving copula transformation of the endogenous regressor, which, as shown in our empirical application, can be substantially different from the pairwise correlations using the original variables (Danaher and Smith 2011). Assumption 4(b) evaluates the correlations between exogenous regressors and the linear combination of generated regressors, which are even more different from checking pairwise correlations on the regressors themselves. When the above assumptions are satisfied, Copula$_{\text{Origin}}$ is preferred to our proposed 2sCOPE method, since the simpler and valid model outperforms more general but more complex models.

If any endogenous regressor has insufficient non-normality, or any exogenous regressor violates the Assumptions 4 or 4(b), our proposed 2sCOPE method can be used instead of Copula$_{\text{Origin}}$. The 2sCOPE is straightforward to extend to many other settings, and we have derived 2sCOPE for a range of commonly used marketing models, including linear regression models, linear panel models with mixed-effects, random coefficient logit models and slope endogeneity. The 2sCOPE method proposed here can be applied to these and many other cases not studied here, while accounting for correlations between exogenous and endogenous regressors and exploiting the correlations for model identification in the presence of insufficient non-normality of endogenous regressors. When endogenous regressors all have sufficient non-normality, our evaluation shows that 2sCOPE performs well. If any endogenous regressor has insufficient non-normality, 2sCOPE exploits exogenous regressors for model identification and requires non-normality of at least one exogenous regressor correlated with the endogenous one. One can check and test the correlations of endogenous and exogenous regressors using correlation tests or examining the first-stage regression coefficients. Non-normality of any exogenous regressor correlated with the endogenous one can be checked using visual plots and normality tests. We also propose a bootstrap algorithm to directly gauge and validate the finite sample performance of 2sCOPE in real applications, complementing the above indirect diagnosis methods using tests of normality and correlations.

Although 2sCOPE contributes to solving regressor endogeneity by relaxing key assumptions of the existing copula correction methods and extending them to more general settings, it is not without limitations. For 2sCOPE to work best, the distributions of the endogenous regressors need to contain adequate information. The condition is violated when the endogenous regressors follow Bernoulli distributions or discrete distributions with small support, as noted in Park and Gupta (2012). The proposed 2sCOPE method does not address this limitation. The simplicity of 2sCOPE hinges on the normal structural error and Gaussian copula dependence structure. Our evaluation shows 2sCOPE is robust to symmetric non-normal error distributions and certain non-Gaussian copula (Web Appendix E.4 & E.5). Such robustness may not hold for asymmetric nonnormal error distributions or all forms of dependence structure. Future research is needed for more flexible methods to test and relax these assumptions. Despite these limitations, we expect that 2sCOPE will provide a useful alternative to a broad range of empirical problems when instruments are not available.

REFERENCES

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt (2005), "Competition and innovation: An inverted-U relationship," *The Quarterly Journal of Economics*, 120 (2), 701–728.

Anderson, Eric T and Duncan I Simester (2004), "Long-run effects of promotion depth on new versus established customers: three field studies," *Marketing Science*, 23 (1), 4–20.

Angrist, Joshua D and Alan B Krueger (1991), "Does compulsory school attendance affect schooling and earnings?," *The Quarterly Journal of Economics*, 106 (4), 979–1014.

Angrist, Joshua D and Alan B Krueger (2001), "Instrumental variables and the search for identification: From supply and demand to natural experiments," *Journal of Economic perspectives*, 15 (4), 69–85.

Arora, Neeraj and Joel Huber (2001), "Improving parameter estimates and model prediction by aggregate customization in choice experiments," *Journal of Consumer Research*, 28 (2), 273–283.

Ataman, M Berk, Harald J Van Heerde, and Carl F Mela (2010), "The long-term effect of marketing strategy on brand sales," *Journal of Marketing Research*, 47 (5), 866–882.

Atefi, Yashar, Michael Ahearne, James G Maxham III, D Todd Donavan, and Brad D Carlson (2018), "Does selective sales force training work?," *Journal of Marketing Research*, 55 (5), 722–737.

Athey, Susan and Guido W Imbens (2006), "Identification and inference in nonlinear difference-in-differences models," *Econometrica*, 74 (2), 431–497.

Becker, Jan-Michael, Dorian Proksch, and Christian M Ringle (2021), "Revisiting Gaussian copulas to handle endogenous regressors," *Journal of the Academy of Marketing Science*, 50 (1), 1–21.

Berry, Steven T (1994), "Estimating discrete-choice models of product differentiation," *The RAND Journal of Economics*, pages 242–262.

Chintagunta, Pradeep, Jean-Pierre Dubé, and Khim Yong Goh (2005), "Beyond the endogeneity bias: The effect of unmeasured brand characteristics on household-level brand choice models," *Management Science*, 51 (5), 832–849.

Chintagunta, Pradeep, Tülin Erdem, Peter E Rossi, and Michel Wedel (2006), "Structural modeling in marketing: review and assessment," *Marketing Science*, 25 (6), 604–616.

Christopoulos, Dimitris, Peter McAdam, and Elias Tzavalis (2021), "Dealing with endogeneity in threshold models using copulas," *Journal of Business & Economic Statistics*, 39 (1), 166–178.

Danaher, Peter J (2007), "Modeling page views across multiple websites with an application to internet reach and frequency prediction," *Marketing Science*, 26 (3), 422–437.

Danaher, Peter J and Michael S Smith (2011), "Modeling multivariate distributions using copulas: Applications in marketing," *Marketing science*, 30 (1), 4–21.

Datta, Hannes, Bram Foubert, and Harald J Van Heerde (2015), "The challenge of retaining customers acquired with free trials," *Journal of Marketing Research*, 52 (2), 217–234.

Dotson, Jeffrey P and Greg M Allenby (2010), "Investigating the strategic influence of customer and employee satisfaction on firm financial performance," *Marketing Science*, 29 (5), 895–908.

Ebbes, Peter, Michel Wedel, and Ulf Böckenholt (2009), "Frugal IV alternatives to identify the parameter for an endogenous regressor," *Journal of Applied Econometrics*, 24 (3), 446–468.

Ebbes, Peter, Michel Wedel, Ulf Böckenholt, and Ton Steerneman (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, Christine and Jan Hohberger (2022), "Addressing Endogeneity Without Instrumental Variables: An Evaluation of the Gaussian Copula Approach for Management Research," *Journal of Management*, DOI: 10.1177/01492063221085913.

Efron, Bradley and Robert J Tibshirani (1994), *An introduction to the bootstrap* CRC press.

Elshiewy, Ossama and Yasemin Boztug (2018), "When back of pack meets front of pack: How salient and simplified nutrition labels affect food sales in supermarkets," *Journal of Public Policy & Marketing*, 37 (1), 55–67.

Godes, David and Dina Mayzlin (2009), "Firm-created word-of-mouth communication: Evidence from a field test," *Marketing science*, 28 (4), 721–739.

Greene, William H (2003), *Econometric analysis* Pearson Education India.

Hartmann, Wesley, Harikesh S Nair, and Sridhar Narayanan (2011), "Identifying causal marketing mix effects using a regression discontinuity design," *Marketing Science*, 30 (6), 1079–1097.

Haschka, Rouven E (2022), "Handling Endogenous Regressors using Copulas: A Generalization to Linear Panel Models with Fixed Effects and Correlated Regressors," *Journal of Marketing Research, https://doi.org/10.1177/00222437211070820.*

Heitmann, Mark, Jan R Landwehr, Thomas F Schreiner, and Harald J van Heerde (2020), "Leveraging brand equity for effective visual product design," *Journal of Marketing Research*, 57 (2), 257–277.

Hoch, Stephen J, Byung-Do Kim, Alan L Montgomery, and Peter E Rossi (1995), "Determinants of store-level price elasticity," *Journal of Marketing Research*, 32 (1), 17–29.

Hogan, Vincent and Roberto Rigobon (2003), "Using unobserved supply shocks to estimate the returns to education," *Unpublished manuscript.*

Hooker, Giles and Lucas Mentch (2018), "Bootstrap bias corrections for ensemble methods," *Statistics and Computing*, 28 (1), 77–86.

Johnson, Garrett A, Randall A Lewis, and Elmar I Nubbemeyer (2017), "Ghost ads: Improving the economics of measuring online ad effectiveness," *Journal of Marketing Research*, 54 (6), 867–884.

Keller, Wiebke IY, Barbara Deleersnyder, and Karen Gedenk (2019), "Price promotions and popular events," *Journal of Marketing*, 83 (1), 73–88.

Kim, Sungjin, Clarence Lee, and Sachin Gupta (2020), "Bayesian Synthetic Control Methods," *Journal of Marketing Research*, 57(5), 831–852.

Kleibergen, Frank and Eric Zivot (2003), "Bayesian and classical approaches to instrumental variable regression," *Journal of Econometrics*, 114 (1), 29–72.

Lewbel, Arthur (1997), "Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D," *Econometrica: journal of the econometric society*, pages 1201–1213.

Li, Yang and Asim Ansari (2014), "A Bayesian Semiparametric Approach for Endogeneity and Heterogeneity in Choice Models," *Management Science*, 60, 1161–1179.

Mackiewicz, Robert and Andrzej Falkowski (2015), "The Use of Weber Fraction as a Tool to Measure Price Sensitivity: a Gain and Loss Perspective.," *Advances in Consumer Research*, 43.

Mendelson, Haim (2000), "Organizational Architecture and Success in the Information Technology.," *Management Science*, 46, 513–529.

Narayanan, Sridhar and Kirthi Kalyanam (2015), "Position effects in search advertising and their moderators: A regression discontinuity approach," *Marketing Science*, 34 (3), 388–407.

Novak, Sharon and Scott Stern (2009), "Complementarity among vertical integration decisions: Evidence from automobile product development," *Management Science*, 55 (2), 311–332.

Otter, Thomas, Timothy J Gilbride, and Greg M Allenby (2011), "Testing models of strategic behavior characterized by conditional likelihoods," *Marketing Science*, 30 (4), 686–701.

Papies, Dominik, Peter Ebbes, and Harald J Van Heerde "Addressing endogeneity in marketing models," "Advanced methods for modeling markets," pages 581–627, Springer (2017).

Park, Sungho and Sachin Gupta (2012), "Handling endogenous regressors by joint estimation using copulas," *Marketing Science*, 31 (4), 567–586.

Petrin, Amil and Kenneth Train (2010), "A control function approach to endogeneity in consumer choice models," *Journal of marketing research*, 47 (1), 3–13.

Qian, Yi (2008), "Impacts of entry by counterfeiters," *Quarterly Journal of Economics*, 123, 1577–1609.

Qian, Yi and Hui Xie (2022), "Simplifying Bias Correction for Selective Sampling: A Unified Distribution-Free Approach to Handling Endogenously Selected Samples," *Marketing Science*, 41(2), 336–360.

Qian, Yi, Hui Xie, and Anthony Koschmann (2022), "Should Copula Endogeneity Correction

Include Generated Regressors for Higher-order Terms? No, It Hurts," *NBER Working Paper, https://www.nber.org/papers/w29978.*

Rigobon, Roberto (2003), "Identification through heteroskedasticity," *Review of Economics and Statistics*, 85 (4), 777–792.

Rossi, Peter E (2014), "Even the rich can make themselves poor: A critical examination of IV methods in marketing applications," *Marketing Science*, 33 (5), 655–672.

Rutz, Oliver J and George F Watson (2019), "Endogeneity and marketing strategy research: An overview," *Journal of the Academy of Marketing Science*, 47 (3), 479–498.

Shi, Huanhuan, Shrihari Sridhar, Rajdeep Grewal, and Gary Lilien (2017), "Sales representative departures and customer reassignment strategies in business-to-business markets," *Journal of Marketing*, 81 (2), 25–44.

Sklar, M (1959), "Fonctions de repartition an dimensions et leurs marges," *Publ. inst. statist. univ. Paris*, 8, 229–231.

Sorescu, Alina, Nooshin L. Warren, and Larisa Ertekin (2017), "Event Study Methodology in the Marketing Literature: An overview," *Journal of the Academy of Marketing Science*, 45, 186–207.

Sudhir, Karunakaran (2001), "Competitive pricing behavior in the auto market: A structural analysis," *Marketing Science*, 20 (1), 42–60.

Sun, Baohong (2005), "Promotion effect on endogenous consumption," *Marketing science*, 24 (3), 430–443.

Van Heerde, Harald J, Maarten J Gijsenberg, Marnik G Dekimpe, and Jan-Benedict EM Steenkamp (2013), "Price and advertising effectiveness over the business cycle," *Journal of Marketing Research*, 50 (2), 177–193.

Villas-Boas, J Miguel and Russell S Winer (1999), "Endogeneity in brand choice models," *Management science*, 45 (10), 1324–1338.

Wang, Yixin and David M Blei (2019), "The blessings of multiple causes," *Journal of the American Statistical Association*, 114 (528), 1574–1596.

Yang, Sha, Yuxin Chen, and Greg M Allenby (2003), "Bayesian analysis of simultaneous demand and supply," *Quantitative marketing and economics*, 1 (3), 251–275.

Addressing Endogeneity using a Two-stage Copula Generated Regressor Approach

WEB APPENDIX

TABLE OF CONTENTS

WEB APPENDIX A: PROOFS RELATED TO COPULA$_{\mathbf{ORIGIN}}$ AND COPE

Web Appendix A.1: Proof of Theorem 1

Under the Gaussian copula assumption for structural error ξ_t and the endogenous regressor P_t, and the normality assumption of ξ_t, the outcome regression becomes (Equation 6)

$$Y_t = \mu + P_t\alpha + W_t\beta + \sigma_\xi \cdot \rho \cdot P_t^* + \sigma_\xi \cdot \sqrt{1-\rho^2} \cdot \omega_t.$$

Because of the exogeneity assumption of W_t in linear model (Equation 1), $Cov(W_t, \xi_t) = 0$,

$$\begin{aligned} Cov(W_t, \xi_t) &= Cov(W_t, \sigma_\xi \cdot \rho \cdot P_t^* + \sigma_\xi \cdot \sqrt{1-\rho^2} \cdot \omega_t) \\ &= \sigma_\xi \cdot \rho \cdot Cov(W_t, P_t^*) + \sigma_\xi \cdot \sqrt{1-\rho^2} \cdot Cov(W_t, \omega_t) = 0. \end{aligned}$$

Thus, whenever W_t and P_t^* is correlated, the covariance between W_t and P_t^* is

$$Cov(W_t, \omega_t) = -\frac{\rho}{\sqrt{1-\rho^2}} Cov(W_t, P_t^*) \neq 0,$$

and W_t would be correlated with the new error term ω_t. **Theorem proved.**

Web Appendix A.2: Assumption 4(b) in Copula$_{\textbf{Origin}}$

According to Park and Gupta (2012), under a Gaussian copula model for $(P_{1,t}, P_{2,t}, \xi_t)$, the structural model in Equation (13) with two endogenous regressors can be re-expressed as

$$\begin{aligned} Y_t =& \mu + P_{1,t}\alpha_1 + P_{2,t}\alpha_2 + W_t\beta + \sigma_\xi \frac{\rho_{\xi 1} - \rho_{12}\rho_{\xi 2}}{1-\rho_{12}^2} \cdot P_{1,t}^* + \sigma_\xi \frac{\rho_{\xi 2} - \rho_{12}\rho_{\xi 1}}{1-\rho_{12}^2} \cdot P_{2,t}^* \\ &+ \sigma_\xi \cdot \sqrt{1-\rho_{\xi 1}^2 - \frac{(\rho_{\xi 2} - \rho_{12}\rho_{\xi 1})^2}{1-\rho_{12}^2}} \cdot \omega_t. \end{aligned} \tag{W1}$$

where $P_{1,t}^* = \Phi^{-1}(H_1(P_{1,t}))$, $P_{2,t}^* = \Phi^{-1}(H_2(P_{2,t}))$, and $H_1(\cdot)$ and $H_2(\cdot)$ are CDFs of $P_{1,t}$ and $P_{1,t}$, respectively, ρ_{12} is the correlation between $P_{1,t}^*$ and $P_{2,t}^*$, $\rho_{\xi 1}$ is the correlation between ξ and $P_{1,t}^*$, $\rho_{\xi 2}$ is the correlation between ξ and $P_{2,t}^*$, and ω_t is a standard normal random variable that is independent of $P_{1,t}^*$ and $P_{2,t}^*$. For the OLS estimation of Equation (W1) to yield consistent estimates, W_t need also be uncorrelated with ω_t, which requires that $Cov(W_t, \sigma_\xi \cdot \sqrt{1-\rho_{\xi 1}^2 - \frac{(\rho_{\xi 2}-\rho_{12}\rho_{\xi 1})^2}{1-\rho_{12}^2}} \cdot \omega_t) = -Cov(W_t, \frac{\rho_{\xi 1}-\rho_{12}\rho_{\xi 2}}{1-\rho_{12}^2} \cdot P_{1,t}^* + \frac{\rho_{\xi 2}-\rho_{12}\rho_{\xi 1}}{1-\rho_{12}^2} \cdot P_{2,t}^*) = 0$ (Assumption 4(b) in the main text) where $\frac{\rho_{\xi 1}-\rho_{12}\rho_{\xi 2}}{1-\rho_{12}^2} \cdot P_{1,t}^* + \frac{\rho_{\xi 2}-\rho_{12}\rho_{\xi 1}}{1-\rho_{12}^2} \cdot P_{2,t}^*$ is the CCF used to control for endogeneity in Copula$_{\text{Origin}}$.

Web Appendix A.3: COPE Method Development

Under the Gaussian copula model for the endogenous regressor, P_t, the correlated exogenous regressor, W_t, and the structural error term, ξ_t in Equation (10), the structural error in Equation (1) can be re-expressed as

$$\xi_t = \sigma_\xi \cdot \xi_t^* = \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* + \frac{-\sigma_\xi \rho_{pw} \rho_{p\xi}}{1-\rho_{pw}^2} W_t^* + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \omega_{3,t}. \quad \text{(W2)}$$

In this way, the structural error term ξ_t is split into two parts: one part as a function of P_t^* and W_t^* that captures the endogeneity of P_t and the association of W_t with $\xi_t|P_t$ [12], and the other part as an independent new error term. Then, we substitute Equation (W2) into the main model in Equation (1), and obtain the following regression equation:

$$Y_t = \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* + \frac{-\sigma_\xi \rho_{pw} \rho_{p\xi}}{1-\rho_{pw}^2} W_t^* + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}. \quad \text{(W3)}$$

Given P_t^* and W_t^* as additional regressors, $\omega_{3,t}$ is not correlated with all regressors on the right-hand side of Equation (W3) as proved in Theorem A1 below, and thus we can consistently estimate the model using the least squares estimator. The regressors P_t^* and W_t^* can be generated from the nonparametric distribution of P_t and W_t as $P_t^* = \Phi^{-1}(\widehat{H}(P_t))$ and $W_t^* = \Phi^{-1}(\widehat{L}(W_t))$, where $\widehat{H}(P_t)$ and $\widehat{L}(W_t)$ are the empirical CDFs of P_t and W_t, respectively.

Theorem A1. Estimation Consistency. Assuming (1) the error term is normal, (2) the endogenous regressor P_t and exogenous regressors W_t are non-normally distributed, and

[12] Although the exogenous regressor W_t and ξ_t are uncorrelated, W_t and $\xi_t|P_t$ (the error component in ξ_t remaining after removing the effect of the endogenous regressor P_t) can be correlated as seen by the correlation between W_t and ω_t in Figure 1 (b).

(3) a Gaussian Copula for the error term, P_t and W_t, $Cov(\omega_{3,t}, W_t) = Cov(\omega_{3,t}, P_t) = Cov(\omega_{3,t}, W_t^*) = Cov(\omega_{3,t}, P_t^*) = 0$ and thus the OLS estimation of Equation (W3) yields consistent estimates of model parameters.

Proof: See Web Appendix A.4, Proof of Theorem A1

As shown in Theorem A1, the proposed COPE method does not require the uncorrelatedness between P_t^* and W_t for consistent model estimation, an assumption needed for $\text{Copula}_{\text{Origin}}$. In fact, $\text{Copula}_{\text{Origin}}$ can be obtained as a special case of the COPE: when W_t is uncorrelated with P_t (i.e., $\rho_{pw} = 0$) and also uncorrelated with P_t^* under the joint copula model, $\frac{-\sigma_\xi \rho_{pw} \rho_{p\xi}}{1-\rho_{pw}^2} W_t^*$ in Equation (W3) vanishes and COPE based on Equation (W3) reduces to $\text{Copula}_{\text{Origin}}$ base on Equation (6). This broader applicability of COPE is a merit of COPE. However, similar to $\text{Copula}_{\text{Origin}}$, COPE requires the non-normality of the endogenous regressor P_t to fulfill the full-rank identification assumption. In addition, COPE requires the non-normality of the exogenous regressor W_t to fulfill the full-rank identification assumption. In the next subsection, we further extend the model to multiple endogenous regressors.

COPE in Multiple Endogenous Regressors Case

Under the Gaussian Copula assumption that $[P_{1,t}^*, P_{2,t}^*, W_t^*, \xi_t^*]$ follows a multivariate normal distribution:

$$\begin{pmatrix} P_{1,t}^* \\ P_{2,t}^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_p & \rho_{wp1} & \rho_{\xi p1} \\ \rho_p & 1 & \rho_{wp2} & \rho_{\xi p2} \\ \rho_{wp1} & \rho_{wp2} & 1 & 0 \\ \rho_{\xi p1} & \rho_{\xi p2} & 0 & 1 \end{bmatrix} \right),$$

we have:

$$\begin{pmatrix} P_{1,t}^* \\ P_{2,t}^* \\ W_t^* \\ \xi_t^* \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \rho_p & \sqrt{1-\rho_p^2} & 0 & 0 \\ \rho_{wp1} & \frac{\rho_{wp2}-\rho_p \rho wp1}{\sqrt{1-\rho_p^2}} & \sqrt{1-\rho_{wp1}^2 - \frac{(\rho_{wp2}-\rho_p\rho_{wp1})^2}{1-\rho_p^2}} & 0 \\ \rho_{\xi p1} & \frac{\rho_{\xi p2}-\rho_p\rho_{\xi p1}}{\sqrt{1-\rho_p^2}} & \frac{-\rho_{wp1}\rho_{\xi p1}-\frac{(\rho_{wp2}-\rho_p\rho_{wp1})(\rho_{\xi p2}-\rho_p\rho_{\xi p1})}{1-\rho_p^2}}{\sqrt{1-\rho_{wp1}^2-\frac{(\rho_{wp2}-\rho_p\rho_{wp1})^2}{1-\rho_p^2}}} & m \end{pmatrix} \cdot \begin{pmatrix} \omega_{1,t} \\ \omega_{2,t} \\ \omega_{3,t} \\ \omega_{4,t} \end{pmatrix},$$

$$\begin{pmatrix} \omega_{1,t} \\ \omega_{2,t} \\ \omega_{3,t} \\ \omega_{4,t} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right), \tag{W4}$$

where m is a function of all the ρs. Under the Gaussian Copula assumption above, we can derive ξ_t^* as a function of P_t and W_t. After simplification, the structural error in Equation (13) can be decomposed as

$$\xi_t = \sigma_\xi \xi_t^* = \eta_1 P_{1,t}^* + \eta_2 P_{2,t}^* - (\eta_1 \rho_{wp1} + \eta_2 \rho_{wp2}) W_t^* + \sigma_\xi \cdot m \cdot \omega_{4,t}. \tag{W5}$$

where

$$\eta_1 = \frac{\sigma_\xi \rho_{\xi p1}(1-\rho_{wp2}^2) - \sigma_\xi \rho_{\xi p2}(\rho_p - \rho_{wp1}\rho_{wp2})}{1-\rho_p^2-\rho_{wp1}^2-\rho_{wp2}^2+2\rho_p\rho_{wp1}\rho_{wp2}},$$
$$\eta_2 = \frac{\sigma_\xi(\rho_{wp1}\rho_{wp2}\rho_{\xi p1}+\rho_{\xi p2}-\rho_p\rho_{\xi p1}-\rho_{wp1}^2\rho_{\xi p2})}{1-\rho_p^2-\rho_{wp1}^2-\rho_{wp2}^2+2\rho_p\rho_{wp1}\rho_{wp2}}. \tag{W6}$$

The COPE method with one endogenous regressor in Equation (W3) is then extended to

$$Y_t = \mu + P_{1,t}\alpha_1 + P_{2,t}\alpha_2 + W_t\beta + \eta_1 P_{1,t}^* + \eta_2 P_{2,t}^* - (\eta_1\rho_{wp1}+\eta_2\rho_{wp2})W_t^* + \sigma_\xi \cdot m \cdot \omega_{4,t}. \tag{W7}$$

In Equation (W7), the new error term $\omega_{4,t}$ is uncorrelated with all the regressors on the right-hand side of Equation (W7). Thus, the OLS estimation of Equation (W7) provides consistent estimates of structural regression model parameters $(\mu, \alpha_1, \alpha_2, \beta)$.

Web Appendix A.4: Proof of Theorem A1

Under the Gaussian copula model for (P_t, ξ_t) and the normality assumption of the error term ξ_t, we can divide ξ_t into an endogenous and an exogenous part, and COPE is based on the OLS estimation of the regression below by adding P_t^* and W_t^* as generated regressors.

$$Y_t = \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* + \frac{-\sigma_\xi \rho_{pw}\rho_{p\xi}}{1-\rho_{pw}^2} W_t^* + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}$$

We aim to prove that the new error $\omega_{3,t}$ is uncorrelated with all terms of the right-hand side. Since $\omega_{1,t}$, $\omega_{2,t}$ and $\omega_{3,t}$ follow a standard multivariate Gaussian distribution (Equation 10), they are independent. According to the same equation, W_t^* and P_t^* are linear functions of $\omega_{1,t}$ and $\omega_{2,t}$. Thus, P_t^* and W_t^* are normally distributed and are independent of $\omega_{3,t}$. Since functions of independent variables are still independent, P_t (W_t), as a function of P_t^* (W_t^*), would be uncorrelated with $\omega_{3,t}$ and thus $\omega_{3,t}$ is not correlated with P_t, P_t^*, W_t and W_t^* on the right-hand side of Equation (W3). Since P_t and W_t are nonnormal distributed, the full rank assumption is satisfied and thus COPE yields consistent estimates. **Theorem proved**.

Next, we show that this result can be readily extended to the multi-dimension W_t case. We first derive the regression of the COPE method. Here we take 2-dimension W_t as an example. When there are one endogenous regressor P_t and two exogenous regressors W_t, the linear regression is:

$$Y_t = \beta_0 + \beta_1 P_t + \beta_2 W_{1,t} + \beta_3 W_{2,t} + \xi_t \qquad \text{(W8)}$$

Under the Gaussian Copula assumption,

$$\begin{pmatrix} P_t^* \\ W_{1,t}^* \\ W_{2,t}^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_\xi \\ \rho_1 & 1 & \rho_w & 0 \\ \rho_2 & \rho_w & 1 & 0 \\ \rho_\xi & 0 & 0 & 1 \end{bmatrix} \right) \tag{W9}$$

The multivariate normal distribution can be written as follows:

$$\begin{pmatrix} P_t^* \\ W_{1,t}^* \\ W_{2,t}^* \\ \xi_t^* \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \rho_1 & \sqrt{1-\rho_1^2} & 0 & 0 \\ \rho_2 & \frac{\rho_w-\rho_1\rho_2}{\sqrt{1-\rho_1^2}} & \sqrt{1-\rho_2^2-\frac{(\rho_w-\rho_1\rho_1)^2}{1-\rho_1^2}} & 0 \\ \rho_\xi & \frac{-\rho_1\rho_\xi}{\sqrt{1-\rho_1^2}} & \frac{\frac{(\rho_w-\rho_1\rho_2)\rho_1\rho_\xi}{1-\rho_1^2}-\rho_2\rho_\xi}{\sqrt{1-\rho_2^2-\frac{(\rho_w-\rho_1\rho_2)^2}{1-\rho_1^2}}} & \gamma \end{pmatrix} \cdot \begin{pmatrix} \omega_{1,t} \\ \omega_{2,t} \\ \omega_{3,t} \\ \omega_{4,t} \end{pmatrix},$$

where $\omega_{k,t} \sim N(0,1), k = 1,2,3,4,$ $\gamma = \sqrt{1-\rho_\xi^2-\frac{\rho_1^2\rho_\xi^2}{1-\rho_1^2}-\left(\frac{\frac{(\rho_w-\rho_1\rho_2)\rho_1\rho_\xi}{1-\rho_1^2}-\rho_2\rho_\xi}{\sqrt{1-\rho_2^2-\frac{(\rho_w-\rho_1\rho_2)^2}{1-\rho_1^2}}}\right)^2}$. Structural error ξ_t can then be written as a function of P_t^* and W_t^*,

$$\xi_t = \sigma_\xi\xi_t^* = \frac{\sigma_\xi\rho_\xi(1-\rho_w^2)}{1-\rho_1^2-\rho_2^2+2\rho_1\rho_2\rho_w+\rho_w^2}\left(P_t^*-\frac{\rho_1-\rho_2\rho_w}{1-\rho_w^2}W_{1,t}^*-\frac{\rho_2-\rho_1\rho_w}{1-\rho_w^2}W_{2,t}^*\right)+\sigma_\xi\gamma\cdot\omega_{4,t}. \tag{W10}$$

Thus, the COPE method in 2-W case becomes:

$$Y_t = \beta_0+\beta_1P_t+\beta_2W_{1,t}+\beta_3W_{2,t}+\beta_4P_t^*+\beta_5W_{1,t}^*+\beta_6W_{2,t}^*+\sigma_\xi\gamma\cdot\omega_{4,t}, \tag{W11}$$

$$\text{where} \quad \beta_4 = \frac{\sigma_\xi \rho_\xi (1-\rho_w^2)}{1-\rho_1^2-\rho_2^2+2\rho_1\rho_2\rho_w+\rho_w^2}$$
$$\beta_5 = \frac{-\sigma_\xi \rho_\xi (1-\rho_w^2)}{1-\rho_1^2-\rho_2^2+2\rho_1\rho_2\rho_w+\rho_w^2} \cdot \frac{\rho_1-\rho_2\rho_w}{1-\rho_w^2}$$
$$\beta_6 = \frac{-\sigma_\xi \rho_\xi (1-\rho_w^2)}{1-\rho_1^2-\rho_2^2+2\rho_1\rho_2\rho_w+\rho_w^2} \cdot \frac{\rho_2-\rho_1\rho_w}{1-\rho_w^2}.$$

Since $\omega_{4,t}$ is independent of P_t^*, $W_{1,t}^*$ and $W_{2,t}^*$, it would also be uncorrelated with any functional form of P_t^*, $W_{1,t}^*$ and $W_{2,t}^*$, and thus $\omega_{4,t}$ is uncorrelated with any other terms in Equation (W11). The COPE method can easily be extended to the case with multiple endogenous regressors by adding copula transformation of each regressor as generated regressors into the outcome regression, and the proof of estimation consistency is similar.

WEB APPENDIX B: PROOFS FOR 2SCOPE

Web Appendix B.1: Proof of Theorem 2 Consistency of 2sCOPE

We have shown the derivation of 2sCOPE method in the main text. The system of equations used in the 2sCOPE method (Equations 8, 9) leads to the following equations

$$
\begin{aligned}
Y_t &= \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2}\epsilon_t + \sigma_\xi\sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}, \\
P_t^* &= \rho_{pw} W_t^* + \epsilon_t.
\end{aligned}
$$

Since $\omega_{3,t}$ is independent of P_t^* and W_t^*, it would also be uncorrelated with any functional form of P_t^* and W_t^*, and thus $\omega_{3,t}$ is uncorrelated with P_t, W_t and ϵ_t. Once P_t or W_t is nonnormal, ϵ_t is not a linear function of P_t and W_t, satisfying the full rank condition required for model identification using the 2sCOPE method. **Theorem proved**.

Next, we show that this result can be easily extended to the multi-dimension W_t case. We first derive the system of equations of the 2sCOPE method. Here we take 2-dimension W_t as an example. Because of the Gaussian relationship among P_t^* and W_t^* we assumed in Equation (W9), the first stage regression becomes

$$
\begin{aligned}
P_t^* &= \frac{\rho_1 - \rho_2\rho_w}{1-\rho_w^2} W_{1,t}^* + \frac{\rho_2 - \rho_1\rho_w}{1-\rho_w^2} W_{2,t}^* + \sqrt{1-\rho_1^2 - \frac{(\rho_2-\rho_1\rho_w)^2}{1-\rho_w^2}}\omega_{3,t} \\
&= \frac{\rho_1 - \rho_2\rho_w}{1-\rho_w^2} W_{1,t}^* + \frac{\rho_2 - \rho_1\rho_w}{1-\rho_w^2} W_{2,t}^* + \epsilon_{2,t} \\
&= \gamma_1 W_{1,t}^* + \gamma_2 W_{2,t}^* + \epsilon_{2,t}. \qquad \text{(W12)}
\end{aligned}
$$

The structural error ξ_t in Equation (W8) and the first-stage error term $\epsilon_{2,t}$ are linear transformations of the Gaussian data $(\xi_t, P_t^*, W_{1,t}^*, W_{2_t}^*)$ and thus follow a bivariate normal distribution. Thus, ξ_t can be decomposed to a sum of one term containing $\epsilon_{2,t}$ and an independent new error term, resulting in the following regression equation:

$$Y_t = \beta_0 + \beta_1 P_t + \beta_2 W_{1,t} + \beta_3 W_{2,t} + \beta_4 \epsilon_{2,t} + \sigma_\xi \gamma \cdot \omega_{4,t}. \tag{W13}$$

where

$$\beta_4 = \frac{\sigma_\xi \rho_\xi (1 - \rho_w^2)}{1 - \rho_1^2 - \rho_2^2 + 2\rho_1 \rho_2 \rho_w + \rho_w^2}.$$

Since $\omega_{4,t}$ is independent of P_t^*, $W_{1,t}^*$ and $W_{2,t}^*$, it is uncorrelated with any functional form of P_t^*, $W_{1,t}^*$ and $W_{2,t}^*$, and thus $\omega_{4,t}$ is uncorrelated with P_t, $W_{1,t}$, $W_{2,t}$ and $\epsilon_{2,t}$ in Equation (W13). Thus, 2sCOPE that performs OLS regression of Equation (W13) yields consistent model estimates. Without loss of generality, the result can be extended to cases with any dimension of W_t.

Web Appendix B.2: Proof of Theorem 3 Non-normality Assumption Relaxed

In this section, we prove that our proposed 2sCOPE method can relax the non-normality assumption on the endogenous regressors imposed in $\text{Copula}_{\text{Origin}}$, while COPE does not.

We first examine the COPE method in Equation (W3),

$$Y_t = \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* + \frac{-\sigma_\xi \rho_{pw}\rho_{p\xi}}{1-\rho_{pw}^2} W_t^* + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}.$$

If the endogenous regressor P_t is normally distributed, $P_t = \Phi_{\sigma_p}^{-1}(\Phi(P_t^*)) = \sigma_p P_t^*$ and thus P_t^* and P_t would be fully collinear, violating the full rank assumption and making the model unidentified.

We then examine the 2sCOPE method in Equation (12).

$$\begin{aligned}
Y_t &= \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} \epsilon_t + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}, \\
\epsilon_t &= P_t^* - \rho_{pw} W_t^*.
\end{aligned}$$

When the endogenous regressor P_t is normally distributed, $P_t = \Phi_{\sigma_p}^{-1}(\Phi(P_t^*)) = \sigma_p P_t^*$. Since we add the residual ϵ_t from the first stage to the outcome regression instead of adding each P_t^* and W_t^*, ϵ_t would not be perfectly collinear with P_t and W_t as long as one of the Ws correlated with P_t is not normally distributed. **Theorem proved.**

Web Appendix B.3: Proof of Theorem 4 Variance Reduction

According to the COPE method in Equation (W3),

$$Y_t = \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* + \frac{-\sigma_\xi \rho_{pw}\rho_{p\xi}}{1-\rho_{pw}^2} W_t^* + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}.$$

The coefficients of P_t^* and W_t^* follows a linear relationship. Denote δ_3 and δ_4 the coefficients of P_t^* and W_t^* respectively. Then,

$$\delta_4 + \rho_{pw}\delta_3 = 0.$$

With the two-stage estimation in 2sCOPE (Equation 12), ρ_{pw} is estimated in the first stage and is thus treated as a known parameter in the main regression. That is, 2sCOPE can be viewed as the COPE method with a linear restriction. The linear restriction is,

$$\delta_4 + \hat{\rho}_{pw}\delta_3 = 0. \tag{W14}$$

In this case, the two-stage copula method (2sCOPE) can be viewed as one kind of restricted least squares estimation based on COPE. We next prove that restricted least squares can achieve reductions in standard errors. Suppose we simplify the regression expression in Equation (W3) as

$$y = X\theta + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2 I)$, $X \equiv (1, P_t, W_t, P_t^*,\ W_t^*)$, and $\theta = (\mu, \alpha, \beta, \delta_3, \delta_4)$. The restriction in Equation (W14) becomes

$$R\theta = 0, \text{where} R = (0, 0, 0, \hat{\rho}_{pw}, 1).$$

Thus, the 2sCOPE yields the least squares estimates $\hat{\theta}_2$ of Equation (W3) subject to the above restriction, whereas COPE yields the unrestricted least squares estimates, $\hat{\theta}_1$, as follows.

$$\hat{\theta}_1 \sim N(\theta, \sigma^2(X'X)^{-1}),$$
$$\hat{\theta}_2 \sim N(\theta, \sigma^2 M(X'X)^{-1}M').$$

where according to restricted least squares theory, $M = I - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}R$. Let us compare the variance of $\hat{\theta}_1$ and $\hat{\theta}_2$. Note that,

$$\begin{aligned} & M(X'X)^{-1}M' \\ =& (I - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}R)(X'X)^{-1}(I - R'(R(X'X)^{-1}R')^{-1}R(X'X)^{-1}) \\ =& (X'X)^{-1} - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}R(X'X)^{-1}. \end{aligned}$$

Therefore,

$$\begin{aligned} Var(\hat{\theta}_1) - Var(\hat{\theta}_2) &= \sigma^2\{(X'X)^{-1} - M(X'X)^{-1}M'\} \\ &= \sigma^2(X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}R(X'X)^{-1} \geq 0. \end{aligned}$$

Since the matrix $Var(\hat{\theta}_1) - Var(\hat{\theta}_2)$ is positive semi-definite, all the diagonal elements should be greater than or equal to zero. Thus, the imposition of the linear restriction brings about a variance reduction. **Theorem proved.**

We have proved that there would be variance reduction when there exist restriction of parameters. When the exogenous regressor W_t is a scalar, the linear restriction is shown in Equation (W14). We next show that when W_t is extended to a multi-dimension vector, there are still linear restrictions and variance reduction of 2sCOPE. We take a 2-dimension W_t as an example below. According to the 2sCOPE method with 2-dimension W_t in Equations (W12, W13), 2sCOPE is equivalent to adding two restrictions to COPE in Equation (W11). The two restrictions are:

$$\beta_5 + \hat{\gamma}_1 \beta_4 = 0$$

$$\beta_6 + \hat{\gamma}_2 \beta_4 = 0$$

where $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are estimated and obtained in the first stage in Equation (W12). Thus, compared with COPE, we still have variance reduction using 2sCOPE in the 2-W case. Without loss of generality, this result can be extended to cases with any dimension of W_t.

WEB APPENDIX C: 2SCOPE FOR SLOPE ENDOGENEITY

In this section, we describe the 2sCOPE approach to addressing slope endogeneity with correlated regressors in the following model:

$$Y_t \;=\; \mu + P_t\alpha_t + W_t'\beta_t + \eta_t, \qquad \text{where } \alpha_t = \bar{\alpha} + \xi_t, \tag{W15}$$

α_t, β_t are individual-specific regression coefficients and $\bar{\alpha}$ is the mean of α_i, $\xi_t \sim N(0, \sigma_\xi^2)$. The normal error term η_i is uncorrelated with the regressors P_t and W_t and thus causes no endogeneity concern. However, the random coefficient ξ_t can be correlated with the regressor P_t, causing the problem of "slope endogeneity". P_t and W_t can be correlated. Assuming that (P_t, W_t, α_t) follows a Gaussian copula model, the COPE approach to addressing the slope endogeneity problem is derived as follows.

$$\begin{aligned} Y_t &= \mu + P_t\Big(\bar{\alpha} + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* + \frac{-\sigma_\xi \rho_{pw}\rho_{p\xi}}{1-\rho_{pw}^2} W_t^* + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}}\,\omega_{3,t}\Big) + W_t'\beta_t + \eta_t \\ &= \mu + P_t\bar{\alpha} + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t \times P_t^* + \frac{-\sigma_\xi \rho_{pw}\rho_{p\xi}}{1-\rho_{pw}^2} P_t \times W_t^* + W_t'\beta_t + \\ &\quad \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}}\, P_t \times \omega_{3,t} + \eta_t. \end{aligned} \tag{W16}$$

Given both $P_t \times P_t^*$ and $P_t \times W_t^*$ in Equation (W16), the unobserved variable $w_{3,t}$ is independent of all regressors (P_t, W_t, P_t^*, W_t^*) and uncorrelated with functions of these regressors. Thus, Equation (W16) can be estimated using standard methods for random-effects models with $\omega_{3,t}$ as the random effect and $(P_t \times P_t^*,\ P_t \times W_t^*)$ as generated regressors. The method of Park and Gupta (2012) adds only $P_t \times P_t^*$ as a generated regressor, and may fail to yield consistent estimates when P_t and W_t are correlated, resulting in the correlation between the random effect in their method and the regressor W_t.

The 2sCOPE for addressing the slope endogeneity problem with correlated regressors is derived as follows

$$\begin{aligned} Y_t &= \mu + P_t(\bar{\alpha} + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2}\epsilon_t + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}) + W_t'\beta_t + \eta_t \\ &= \mu + P_t\bar{\alpha} + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* \times \epsilon_t + W_t'\beta_t + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} P_t \times \omega_{3,t} + \eta_t, \end{aligned} \tag{W17}$$

where only one generated regressor, $P_t^* \times \epsilon_t$, is needed, given which the random effect $\omega_{3,t}$ is independent of all regressors in Equation (W17).

Both COPE and 2sCOPE can be implemented using the standard methods for random effects models by simply adding generated regressors to control for endogenous regressors. By contrast, the maximum likelihood approach requires constructing a complicated joint likelihood of $(\xi_t, \eta_t, P_t^*, W_t^*)$, which is not what the standard random effects method uses and thus requires separate development and significantly more computation involving numerical integration.

WEB APPENDIX D: 2SCOPE FOR RANDOM COEFFICIENT LOGIT MODEL

We next consider endogeneity bias in the following random utility model with correlated endogenous and exogenous regressors:

$$\begin{aligned}
u_{hjt} &= \psi_{hj} + P'_{jt}\alpha_h + W'_{jt}\beta_h + \xi_{jt} + \epsilon_{hjt}, \qquad j = 1, \cdots, J, \\
u_{h0t} &= \epsilon_{h0t}, \qquad j = 0 \text{ if no purchase,}
\end{aligned}$$

where u_{hjt} denotes the utility for household $h = 1, \cdots, n_h$ at occasion $t = 1, \cdots, T$ with $j = 1, \cdots, J$ alternatives and $j = 0$ denotes the option of no purchase. In the utility function, ψ_{hj} is the individual-specific preference for choice j with ψ_{hJ} normalized to be zero for identification purpose, (P_{jt}, W_{jt}) include the choice characteristics, and (α_h, β_h) denote the individual-specific random coefficients. These individual-specific coefficients $(\psi_{hj}, \alpha_h, \beta_h)$ permit heterogeneity in both intercepts and regressor effects across cross-sectional units, such as consumers or households. In this model, the association between regressors in P_{jt} and the unobserved common shock ξ_{jt} causes endogeneity bias. We further allow P_{jt} and W_{jt} to be correlated. The term ϵ_{hjt} is the idiosyncratic error uncorrelated with all regressors. An individual at any occasion chose the alternative with the largest utility, i.e., $Y_{hjt} = 1 \text{ iff } u_{hjt} > u_{hj't} \, \forall j' \neq j$. When ϵ_{hjt} follows an *i.i.d* Type I extreme value distribution, the choice probability follows the random-coefficient multinomial logit model.

The 2sCOPE approach can be used to address the endogeneity issue using the following two-step procedure. In the first step, we estimate the model

$$u_{hjt} = \delta_{jt} + \widetilde{\psi}_{hj} + P'_{jt}a_h + W'_{jt}b_h + \epsilon_{hjt},$$

where $\delta_{jt} = \mu_j + P'_{jt}\bar{\alpha} + W'_{jt}\bar{\beta} + \xi_{jt}$, $(\mu_j, \bar{\alpha}, \bar{\beta})$ is the mean of random effects $(\psi_{hj}, \alpha_h, \beta_h)$, $\widetilde{\psi}_{hj} = \psi_{hj} - \mu_j, a_h = \alpha_h - \bar{\alpha}$ and $b_h = \beta_h - \bar{\beta}$. δ_{jt} is treated as occasion- and choice-specific fixed-effect parameters in this model. Since the regressors are uncorrelated with the error term ϵ_{hij}, there is no endogeneity bias in the model. In the second step, we estimate the equation below.

$$\widehat{\delta}_{jt} = \mu_j + P'_{jt}\bar{\alpha} + W'_{jt}\bar{\beta} + \xi_{jt} + \eta_{jt}, \tag{W18}$$

where $\widehat{\delta}_{jt}$ denotes the estimate of the fix-effect δ_{jt}; η_{jt} denotes the estimation error of $\widehat{\delta}_{ij}$ and is approximately normally distributed. In the second-step model, the structural error is correlated with P_{jt}, leading to endogenous bias. We then apply 2sCOPE to correct for the endogenous bias, which can avoid the potential bias of $\text{Copula}_{\text{Origin}}$ due to the potential correlations between P and W, as well as make use of this correlation to relax the non-normality assumption of P_{it}, improve model identification and sharpen model estimates. The above development is for individual-level data. Park and Gupta (2012) also derived their copula method for addressing endogeneity bias in random coefficient logit models using aggregate-level data. It is straightforward to extend the 2sCOPE to the setting with correlated regressors and (nearly) normal regressor distributions.

WEB APPENDIX E: ADDITIONAL SIMULATION RESULTS

Web Appendix E.1: Additional Results for Smaller Sample Size for Case 1

In the simulation study case 1, we use sample size T=1000. Here we further check the robustness of results with respect to smaller sample size. We simulate 1000 data sets, each of which has sample size T=200, and use the same DGP as described in Case 1. Table W1 shows that 2sCOPE has unbiased estimates for small sample size T=200. Hence, our proposed method is robust and can be applied to small sample size.

			OLS			Copula$_{\text{Origin}}$			COPE			2sCOPE		
ρ_{pw}	Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
0.5	μ	1	0.683	0.097	3.264	1.228	0.191	1.194	1.020	0.223	0.091	0.999	0.137	0.005
	α	1	1.583	0.079	7.388	1.048	0.178	0.271	0.990	0.184	0.056	0.996	0.175	0.023
	β	-1	-1.265	0.068	3.902	-1.291	0.068	4.293	-1.019	0.166	0.116	-1.004	0.101	0.044
	$\rho_{p\xi}$	0.5	-	-	-	0.559	0.122	0.489	0.493	0.139	0.048	0.489	0.097	0.109
	σ_ξ	1	0.857	0.044	3.224	1.016	0.107	0.148	1.018	0.100	0.176	1.001	0.094	0.013
	D-error			-			-			0.016598			0.009069	
0.7	μ	1	0.723	0.091	3.050	1.304	0.175	1.740	1.006	0.197	0.031	0.983	0.114	0.153
	α	1	1.817	0.095	8.583	1.255	0.161	1.584	1.032	0.182	0.175	1.044	0.174	0.253
	β	-1	-1.539	0.084	6.388	-1.574	0.086	6.686	-1.045	0.180	0.250	-1.033	0.131	0.251
	$\rho_{p\xi}$	0.5	-	-	-	0.624	0.103	1.200	0.490	0.135	0.077	0.480	0.067	0.297
	σ_ξ	1	0.796	0.039	5.156	0.988	0.105	0.116	0.999	0.096	0.011	0.982	0.090	0.205
	D-error			-			-			0.016245			0.008867	

Table W1: Results of the Simulation Study for Case 1 with Sample Size of 200

Web Appendix E.2: Multiple Endogenous Regressors

In this case, we examine the performance of our proposed 2sCOPE when the model has multiple endogenous regressors. Specifically, we use the DGP with two endogenous regressors and one exogenous regressor that is correlated with the endogenous regressors below:

$$\begin{pmatrix} P_{1,t}^* \\ P_{2,t}^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 & 0.4 & 0.5 \\ 0.3 & 1 & 0.4 & 0.5 \\ 0.4 & 0.4 & 1 & 0 \\ 0.5 & 0.5 & 0 & 1 \end{bmatrix} \right), \tag{W19}$$

$$\xi_t = G^{-1}(U_{\xi,t}) = G^{-1}(\Phi(\xi_t^*)) = \Phi^{-1}(\Phi(\xi^*)) = 1 \cdot \xi_t^*, \tag{W20}$$

$$P_{1,t} = H_1^{-1}(U_{p1}) = H_1^{-1}(\Phi(P_{1,t}^*)), \;\; P_{2,t} = H_2^{-1}(\Phi(P_{2,t}^*)), \tag{W21}$$

$$W_t = L^{-1}(U_{W,t}) = L^{-1}(\Phi(W_t^*)), \tag{W22}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + \xi_t = 1 + 1 \cdot P_{1,t} + 1 \cdot P_{2,t} + (-1) \cdot W_t + \xi_t, \tag{W23}$$

where $H_1^{-1}(\cdot)$ ($H_2^{-1}(\cdot)$) and $L^{-1}(\cdot)$ are the inverse distribution functions of the gamma and exponential distributions used to generate these regressors. We generate 1000 data sets, each of which has a sample size T=1000. Table W2 shows the estimation results. Both the OLS and Copula$_{\text{Origin}}$ estimates are biased, while our proposed method provides unbiased estimates for all parameters, indicating that 2sCOPE performs well with multiple endogenous regressors.

		OLS			Copula$_{\text{Origin}}$			COPE			2sCOPE		
Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
μ	1	0.419	0.045	13.02	1.267	0.090	2.949	1.012	0.097	0.125	1.008	0.069	0.120
α_1	1	1.450	0.029	15.46	1.040	0.060	0.665	0.990	0.060	0.166	0.991	0.059	0.153
α_2	1	1.450	0.031	14.72	1.040	0.059	0.673	0.990	0.058	0.177	0.991	0.056	0.167
β	-1	-1.320	0.029	11.04	-1.353	0.028	12.56	-0.997	0.057	0.061	-0.995	0.040	0.134
$\rho_{\xi p1}$	0.5	-	-	-	0.567	0.043	1.545	0.503	0.049	0.052	0.502	0.040	0.048
$\rho_{\xi p2}$	0.5	-	-	-	0.568	0.042	1.625	0.503	0.047	0.073	0.503	0.038	0.075
σ_ξ	1	0.772	0.018	12.58	1.019	0.048	0.402	1.012	0.044	0.283	1.010	0.042	0.233

Table W2: Results of the Simulation Study: Multiple Endogenous Regressors

Web Appendix E.3: Multiple Exogenous Control Covariates

We investigate the performance of our proposed method when there exist multiple exogenous regressors consisting of both continuous and discrete variables. We generate the data using the following DGP:

$$\begin{pmatrix} P_t^* \\ W_{1,t}^* \\ W_{2,t}^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.3 & 0 \\ 0.5 & 0.3 & 1 & 0 \\ 0.5 & 0 & 0 & 1 \end{bmatrix}\right), \tag{W24}$$

$$\xi_t = G^{-1}(\Phi(\xi_t^*)) = \Phi^{-1}(\Phi(\xi^*)) = 1 \cdot \xi_t^*, \tag{W25}$$

$$P_t = H^{-1}(\Phi(P_t^*)), \quad W_{1,t} = L^{-1}(\Phi(W_{1,t}^*)), \tag{W26}$$

$$W_{2,t} = \begin{cases} 1, & \text{if} \quad \Phi(W_{2,t}^*) \geq 0.5 \\ 0. & \text{if} \quad \Phi(W_{2,t}^*) < 0.5 \end{cases}, \tag{W27}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta_1 \cdot W_{1,t} + \beta_2 \cdot W_{2,t} + \xi_t = 1 + P_t + (-1) \cdot W_{1,t} + (-1) \cdot W_{2,t} + \xi_t, \tag{W28}$$

where $H^{-1}(\cdot)$ and $L^{-1}(\cdot)$ are the inverse distribution functions of the gamma and exponential distributions. $W_{2,t}$ is a binary variable that follows a Bernoulli distribution. We set sample size $T = 1000$ and generate 1000 data sets to estimate parameters using OLS and copula methods. We follow the approach of Park and Gupta (2012) to generate latent copula data for discrete variables. Specifically, for a discrete regressor W_t, such as the binary exogenous regressor $W_{2,t}$, we define $U_{W,t}$, uniformly distributed on [0,1], as the CDF for a latent variable W_t^* that determines the discrete value of W_t. We then relate $U_{W,t}$ to W_t through the following inequality: $K(W_t - 1) < U_{W,t} < K(W_t)$, where $K(\cdot)$ is the CDF of W_t

and can be directly estimated from the frequencies of the observed data. The above inequality implies the following relationship between $W_t^* = \Phi^{-1}(U_{W,t})$ and $K_{W,t}$: $\Phi^{-1}(K(W_t - 1)) < W_t^* < \Phi^{-1}(K(W_t))$.

The estimation results for the multiple-exogenous-regressor case with both discrete and continuous ones are summarized in Table W3. The OLS and Copula$_{\text{Origin}}$ estimates are biased because of endogeneity and correlated exogenous regressors, respectively. The proposed 2sCOPE method performs well and provides consistent estimates for all parameters. This indicates that our proposed method performs well with multiple exogenous correlated regressors. Moreover, correcting for endogeneity using our proposed method does not require every exogenous correlated regressor to be informative (i.e., continuously distributed).

		OLS			Copula$_{\text{Origin}}$			COPE			2sCOPE		
Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
μ	1	0.701	0.046	6.452	1.281	0.083	3.394	1.007	0.115	0.057	1.005	0.061	0.085
α	1	1.573	0.038	15.10	1.037	0.071	0.532	0.985	0.073	0.208	0.987	0.072	0.180
β_1	-1	-1.225	0.041	5.523	-1.220	0.039	5.584	-0.990	0.069	0.140	-0.992	0.048	0.161
β_2	-1	-1.096	0.075	1.273	-1.202	0.073	2.758	-1.006	0.115	0.051	-1.003	0.080	0.042
$\rho_{p\xi}$	0.5	-	-	-	0.589	0.045	1.976	0.503	0.061	0.053	0.504	0.038	0.097
σ_ξ	1	0.862	0.020	7.066	1.023	0.044	0.532	1.011	0.040	0.264	1.006	0.040	0.115

Table W3: Results of the Simulation Study: Multiple Exogenous Control Covariates

Web Appendix E.4: Misspecification of ξ_t

Similar to Park and Gupta (2012), we assume the structural error ξ_t to be normally distributed, a reasonable and commonly used assumption in marketing and economics literature. However, the true distribution of ξ_t is often unknown. Thus, in this simulation study, we examine the robustness of 2sCOPE to the departures from the normality of ξ_t. We generate 1,000 data sets using the same multivariate normal distribution as in Equation (20). The rest of DGP is:

$$\xi_t = G^{-1}(U_{\xi,t}) \quad = \quad G^{-1}(\Phi(\xi_t^*)), \tag{W29}$$

$$P_t = H^{-1}(U_{p,t}) = H^{-1}(\Phi(P_t^*)), \qquad W_t = L^{-1}(U_{w,t}) = L^{-1}(\Phi(W_t^*)), \tag{W30}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t \quad + \quad \xi_t = 1 + 1 \cdot P_t + (-1) \cdot W_t + \xi_t, \tag{W31}$$

where we set $P_t \sim \mathit{Gamma}(1, 1)$ and $W_t \sim \mathit{Exp}(1)$ in the simulation. We check the robustness of the structural error ξ_t using different distributions (e.g., a uniform distribution, beta distribution and t distribution) instead of a normal distribution. For estimation, we assume normality of ξ_t and use the OLS estimator, COPE and the proposed method.

Table W4 reports estimation results. As shown in Table W4, 2sCOPE can recover the true parameter values despite the misspecification of ξ_t, demonstrating the robustness of the proposed 2sCOPE method to the normal error assumption.

			OLS			COPE			2sCOPE		
Distribution of ξ_t	Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
U[-0.5,0.5]	μ	1	0.912	0.013	6.808	1.004	0.025	0.144	1.002	0.017	0.105
	α	1	1.160	0.010	16.41	0.996	0.017	0.263	0.996	0.017	0.233
	β	-1	-1.072	0.009	8.033	-1.000	0.019	0.023	-0.998	0.011	0.147
	$\rho_{p\xi}$	0.5	-	-	-	0.497	0.049	0.054	0.495	0.035	0.155
	σ_ξ	0.289	0.251	0.004	9.018	0.291	0.009	0.287	0.290	0.008	0.197
Beta(0.5,0.5)	μ	1	0.896	0.016	6.461	1.005	0.031	0.166	1.003	0.020	0.145
	α	1	1.190	0.012	15.72	0.994	0.019	0.343	0.994	0.018	0.318
	β	-1	-1.086	0.011	7.763	-0.999	0.022	0.043	-0.998	0.014	0.183
	$\rho_{p\xi}$	0.5	-	-	-	0.483	0.050	0.339	0.481	0.033	0.593
	σ_ξ	0.354	0.311	0.005	9.046	0.357	0.009	0.355	0.356	0.009	0.258
Beta(4,4)	μ	1	0.948	0.008	6.928	1.000	0.014	0.022	1.000	0.010	0.009
	α	1	1.095	0.006	16.61	0.999	0.011	0.057	1.000	0.010	0.044
	β	-1	-1.043	0.005	8.149	-1.000	0.011	0.004	-1.000	0.007	0.030
	$\rho_{p\xi}$	0.5	-	-	-	0.499	0.053	0.010	0.499	0.037	0.025
	σ_ξ	0.167	0.144	0.003	7.969	0.167	0.006	0.077	0.167	0.006	0.011
t (df=3)	μ	1	0.504	0.082	6.071	0.972	0.198	0.142	0.983	0.127	0.135
	α	1	1.903	0.089	10.13	1.026	0.227	0.113	1.024	0.217	0.110
	β	-1	-1.410	0.064	6.448	-1.003	0.129	0.020	-1.012	0.109	0.111
	$\rho_{p\xi}$	0.5	-	-	-	0.449	0.088	0.577	0.454	0.069	0.676
	σ_ξ	1.732	1.503	0.231	0.992	1.701	0.246	0.124	1.698	0.244	0.141
t (df=5)	μ	1	0.603	0.059	6.723	0.992	0.126	0.066	0.997	0.080	0.039
	α	1	1.727	0.053	13.65	1.006	0.118	0.053	1.006	0.113	0.057
	β	-1	-1.328	0.043	7.642	-0.997	0.090	0.030	-1.002	0.067	0.037
	$\rho_{p\xi}$	0.5	-	-	-	0.483	0.065	0.254	0.486	0.047	0.292
	σ_ξ	1.291	1.118	0.049	3.506	1.292	0.071	0.008	1.289	0.070	0.032

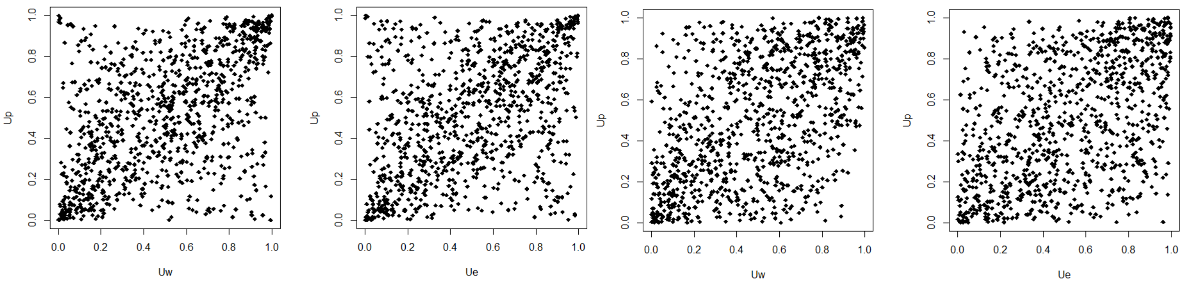
Table W4: Results of the Simulation Study: Misspecification of ξ_t

Web Appendix E.5: Misspecification of Copula

In the proposed method, we use the Gaussian copula to capture the dependence structure among the regressors and error term (U_p, U_w and U_ξ). In practice, the dependence might come from an economic mechanism (such as marketing strategic decisions) and thus might be different from what the Gaussian copula generates. In this section, we examine the robustness of the Gaussian copula assumption in capturing the dependence among the endogenous regressors, exogenous regressors and the error term using simulated data. Specifically, we generate the dependence among U_p, U_w and U_ξ using copula models other than the Gaussian copula. Our simulation setting requires the availability of a random number generation routine from a tri-variate copula model other than Gaussian copula with non-homogeneous correlations among the three variables. Among copula models other than Gaussian copula, we find only T copula has this flexibility of providing flexible random number generation from arbitrary and heterogeneous correlation structures among more than two variables. We thus consider using the following T copula models in which

$$C(U_p, U_w, U_\xi) = \int_{-\infty}^{t_\nu^{-1}(U_p)} \int_{-\infty}^{t_\nu^{-1}(U_w)} \int_{-\infty}^{t_\nu^{-1}(U_\xi)} \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\pi\nu)^d|\Sigma|}} \left(1 + \frac{x'\Sigma^{-1}x}{\nu}\right) dx, \tag{W32}$$

where t_ν^{-1} denotes the quantile function of a standard univariate t_ν distribution. We set the degree of freedom ν=2, and the dimension of the copula d=3 in this example. Σ is covariance matrix capturing correlations among variables. The data-generating process (DGP) of t



(a) T Copula ($\rho = 0.5$)

(b) Gaussian Copula ($\rho = 0.5$)

Figure W1: Scatter plots of Randomly Generated Pairs U_p, U_w (U_p,U_ξ) for Considered Copulas.

copula is summarized below:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim t_\nu^d \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{p\xi} \\ \rho_{pw} & 1 & 0 \\ \rho_{p\xi} & 0 & 1 \end{bmatrix} \right) = t_\nu^d \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right). \quad \text{(W33)}$$

Figure W1 shows the scatter plots of randomly generated (U_p, U_w, U_ξ) pairs from the above copulas, as well as the Gaussian copula with the same correlation of 0.5. The figure shows disparate dependence structures between U_p and ξ_t (U_p and U_w) for these two copulas.

We then use the following process to generate P_t, W_t and ξ_t:

$$\xi_t = G^{-1}(U_\xi) = \Phi^{-1}(U_\xi), \quad \text{(W34)}$$

$$P_t = H^{-1}(U_p), W_t = L^{-1}(U_w), \quad \text{(W35)}$$

$$Y_t = 1 + 1 \cdot P_t + (-1) \cdot W_t + \xi_t. \quad \text{(W36)}$$

where $H(\cdot)$ is a gamma distribution and $L(\cdot)$ is an exponential distribution. We set $T = 1000$, generate 1000 data sets and estimate the parameters using the OLS estimator and the proposed 2sCOPE method.

Table W5 summarizes the estimation results. OLS estimates are still biased for all parameters. By contrast, estimates from the proposed COPE and 2sCOPE methods are centered closely around the true values. Therefore, the proposed methods based on the Gaussian copula are reasonably robust to the mis-specifications of the copula dependence structure among the regressors and the structural error.

		OLS			COPE			2sCOPE		
Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
μ	1	0.710	0.530	5.463	1.002	0.127	0.016	0.988	0.077	0.156
α	1	1.580	0.044	13.13	1.030	0.115	0.257	1.029	0.116	0.250
β	-1	-1.289	0.047	6.142	-1.033	0.127	0.262	-1.017	0.070	0.248
$\rho_{p\xi}$	0.5	-	-	-	0.463	0.085	0.435	0.458	0.067	0.622
σ_{ξ}	1	0.864	0.026	5.236	0.993	0.054	0.133	0.988	0.054	0.230

Table W5: Results of the Simulation Study: Misspecification of Copula

WEB APPENDIX F: IMPLEMENTING THE BOOTSTRAP RE-SAMPLING METHOD IN EMPIRICAL APPLICATION

To gauge and validate the finite sample performance of 2sCOPE, we apply the bootstrap algorithm described in Algorithm 1 to our empirical application and conduct a bootstrap re-sampling study by drawing repeated samples of the same size as the observed data from the underlying copula model and the structural model estimated from the original sample using data from store 1 in the application, and perform estimation on each bootstrap sample. Specifically, we generate data using the following DGP:

$$\begin{pmatrix} \text{Price}^* \\ \text{Bonus}^* \\ \text{PriceRedu}^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 & -0.3 & 0.3 \\ -0.5 & 1 & -0.3 & 0 \\ -0.3 & -0.3 & 1 & 0 \\ 0.3 & 0 & 0 & 1 \end{bmatrix} \right), \tag{W37}$$

$$\xi_t = G^{-1}(\Phi(\xi_t^*)) = \Phi_\sigma^{-1}(\Phi(\xi^*)) = 0.4 \cdot \xi_t^*, \tag{W38}$$

$$\text{Price} = \hat{H}^{-1}(\Phi(\text{Price}^*)), \quad \text{Bonus} = \hat{L}_1^{-1}(\Phi(\text{Bonus}^*)), \tag{W39}$$

$$\text{PriceRedu} = \hat{L}_2^{-1}(\Phi(\text{PriceRedu}^*)), \tag{W40}$$

$$Y_t = -4 + (-2) \cdot \text{Price} + 0.1 \cdot \text{Bonus} + 0.3 \cdot \text{PriceRedu} + \xi_t, \tag{W41}$$

where $\hat{H}(\cdot), \hat{L}_1(\cdot), \hat{L}_2(\cdot)$ are all estimated CDFs using the univariate empirical distribution in the application for regressors Price, Bonus and PriceRedu, respectively. The correlation matrix of the copula transformation of variables (i.e., Price*, Bonus*, PriceRedu*, ξ^*) in Equation (W37) and the standard deviation of the error term (i.e., σ_ξ) are set according to the estimated parameter values using real data. After generating the regressors and the structural error, we set the coefficients using the 2sCOPE estimates of original data to generate Y in Equation (W41). We set the sample size $T = 373$, which is the same as the sample size in the application, and generate $B = 1000$ bootstrap data sets in each of which we estimate the structural model parameters using OLS and copula methods.