NBER WORKING PAPER SERIES

A PRACTICAL GUIDE TO ENDOGENEITY CORRECTION USING COPULAS

Yi Qian
Anthony Koschmann
Hui Xie

Working Paper 32231
http://www.nber.org/papers/w32231

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2024, Revised June 2025

We acknowledge the support by Social Sciences and Humanities Research Council of Canada [grants 435-2018-0519 and 435-2023-0306], Natural Sciences and Engineering Research Council of Canada [grant RGPIN-2018-04313 and 2023-04348] and US National Institute of Health [grant 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. Procedures used were in compliance with British Columbia's Freedom in Information and Privacy Protection Act. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

A Practical Guide to Endogeneity Correction Using Copulas
Yi Qian, Anthony Koschmann, and Hui Xie
NBER Working Paper No. 32231
March 2024, Revised June 2025
JEL No. C01, C10, C5

ABSTRACT

Causal inference is of central interest in many empirical applications, yet often challenging because of the presence of endogenous regressors. The classical approach to the problem requires using instrumental variables that must satisfy the stringent condition of exclusion restriction. In recent research, instrument-free copula methods have been increasingly used to handle endogenous regressors. This article aims to provide a practical guide for how to handle endogeneity using copulas. The authors give an overview of copula endogeneity correction, outlining its theoretical rationales and advantages for empirical research. They also discuss recent advances that enhance the understanding, applicability, and robustness of copula correction, and address implementation aspects of copula correction such as constructing copula control functions and handling higher-order terms of endogenous regressors. To facilitate the appropriate usage of copula correction in order to realize its full potential, the authors detail a process of checking data requirements and identification assumptions to determine when and how to use copula correction methods, and illustrate its usage using empirical examples.

Yi Qian
University of British Columbia
Department of Marketing
and NBER
yi.qian@sauder.ubc.ca

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

Anthony Koschmann
Eastern Michigan University
akoschma@emich.edu

Many research questions in marketing, management, economics, and health sciences are interested in matters of causality rather than simply questions of association. Frequently, these questions are tackled by using relevant data to estimate structural regression models representing causal relationships. A pervasive issue in these empirical investigations is the presence of endogenous regressors, which can arise when the regressors representing the causes (e.g., an economic program to be evaluated, marketing mix variables, etc.) are not randomly assigned in the data; the regressors thus correlate with unobservables (e.g., unobserved product characteristics or common market shocks) in the structural error term (Villas-Boas and Winer 1999). Estimation methods that ignore potential regressor-error dependence can lead to severely biased estimates of structural model parameters (i.e., endogeneity bias).

Given the ubiquity of endogenous regressors and the importance of addressing endogeneity bias, a large body of literature is devoted to developing appropriate methods to solve or mitigate the endogeneity issue. The instrumental variable (IV) method is the classical econometric approach for endogeneity bias correction. This method relies on valid and strong IVs that satisfy the stringent requirement of exclusion restriction (ER), which makes IVs difficult to find and justify in practice (Ebbes et al. 2005). Amid the concerns about the availability and quality of IVs, there has been growing interest in IV-free endogeneity correction methods (Ebbes, Wedel, and Böckenholt 2009; Papies, Ebbes, and Van Heerde 2017; Rutz and Watson 2019; Papies, Ebbes, and Feit 2023). These methods exploit higher moments (HM, Lewbel 1997), identification via heteroscedastic error structures (IH, Rigobon 2003), latent IVs (LIV, Ebbes et al. 2005), semiparametric odds ratio endogeneity models (SORE, Qian and Xie 2024), and copulas[1], starting from the seminal work of Park and Gupta (2012).[2]

Copula correction methods provide substantial advantages for addressing the prevalent and thorny issue of endogenous regressors. These methods directly address the regressor-

[1] "Copula" was introduced by Sklar (1959) from the Latin "to link", as a function linking two variables. Copulas encompass different forms, but we use 'copulas' here to speak synonymously with Gaussian copulas (GC).

[2] See also Christopoulos, McAdam, and Tzavalis (2021); Tran and Tsionas (2021); Becker, Proksch, and Ringle (2022); Haschka (2022); Eckert and Hohberger (2023); Yang, Qian, and Xie (2024a,b); Liengaard et al. (2024); Breitung, Mayer, and Wied (2024); Park and Gupta (2024); Hu, Qian, and Xie (2025).

error dependence using copulas, a widely used multivariate dependence model applicable in many practical applications (Danaher and Smith 2011). Unlike the IV approach and other IV-free methods, copula correction does not require the endogenous regressor to contain an exogenous component (either observed or latent) that must satisfy the stringent ER condition. Thus, copula correction is feasible in many situations when appropriate conditions are met. Moreover, it can be implemented by incorporating copula-based control functions—constructed from existing regressors—into the structural model as additional regressors to address endogeneity. Thus, copula correction using control functions is straightforward to apply in a wide array of settings, including both linear and nonlinear models (e.g., discrete choice models), multiple endogenous regressors, endogenous interaction and higher-order terms, and the slope endogeneity problem.

Largely due to these advantages, copula correction has gained increasing popularity in empirical research. Although the focus here is on marketing, copula correction has been increasingly used in diverse fields outside marketing such as in economics, management, and information systems (e.g., Christopoulos, McAdam, and Tzavalis 2021; Becerra and Markarian 2021; Ananthakrishnan et al. 2025). The pie chart in Figure 1 breaks down by discipline book chapters and journal publications (n=615) using copula endogeneity correction, according to Google Scholar. Each slice in the pie chart matches journals and journal fields as defined by the Australian Business Dean's Council. Strategy and information systems are the two most common business disciplines outside marketing to use copula endogeneity correction. Focusing on the marketing field, the bar chart in Figure 1 breaks down publications (n=100) by substantive area using copula correction in leading marketing journals [3] from 2013 to 2025 (see the complete list of these publications in Web Appendix A).

Like other causal inference methods for nonexperimental data, copula correction requires underlying conditions and data requirements to be met. Prior work reviewed and evaluated

[3]This list includes *Journal of Marketing, Journal of Marketing Research, Marketing Science, Journal of Consumer Research, Journal of the Academy of Marketing Science, Journal of Retailing, International Journal of Research in Marketing,* and *Journal of Consumer Psychology.*

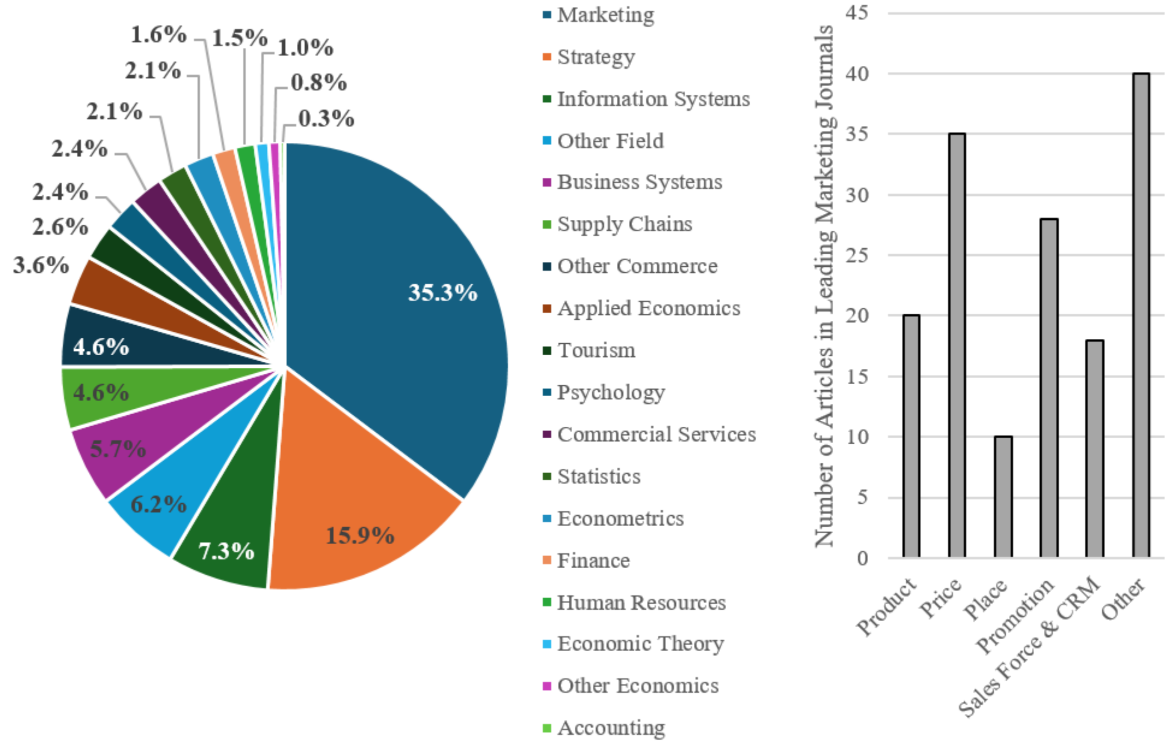

Figure 1: Pie Chart (left): Publications (n=615) using Copula Correction by Disciplines. Bar Chart (right): Publications (n=100) using Copula Correction in Leading Marketing Journals by Substantive Areas. "Other" includes word-of-mouth, warranty claims, etc.

the assumptions and limitations of the original Park and Gupta (2012) method (Papies, Ebbes, and Van Heerde 2017; Becker, Proksch, and Ringle 2022; Eckert and Hohberger 2023). Since then, substantial advances have relaxed these assumptions and requirements, allowing copula correction to function under much weaker conditions. We demonstrate that copula correction using control functions does not depend on normally distributed structural errors or specific copula regressor-error dependence structures, making the method more broadly applicable and robust than previously thought. Although copula correction originally required endogenous regressors to be uncorrelated with exogenous regressors and have sufficient nonnormality, limiting its applicability, the recent two-stage copula endogeneity correction (2sCOPE) approach by Yang, Qian, and Xie (2024a) simultaneously relaxes these restrictions and provides a general framework for further development (e.g., Liengaard et al. 2024; Yang, Qian, and Xie 2024b). Haschka (2022) and Yang, Qian, and Xie (2024b) generalize copula correction to panel data. Hu, Qian, and Xie (2025) introduce nonparametric copula control functions that generalize and unify existing copula correction methods.

Qian and Xie (2024) and Hu, Qian, and Xie (2025) develop IV-free methods for handling noncontinuous endogenous regressors (e.g, binary) that current copula control function methods cannot accommodate. Given significant advances since Park and Gupta (2012)'s study, guidelines for using the expanded copula correction toolbox are clearly needed.

Focusing on assisting potential users of copula correction, the objectives of this article are: (a) raise awareness of the importance of addressing endogenous regressors in empirical studies and demystify theoretical rationales for copula correction; (b) provide a synthesis of recent developments that advance the understanding as well as broaden the applicability and robustness of copula correction; (c) provide updated guidance and delineate a process of checking data requirements and identification assumptions to aid proper usage of copula correction; and (d) demonstrate use of copula correction in practical applications.

With these objectives in mind, the rest proceeds as follows. First, we overview the theoretical rationale for endogeneity correction using copulas. This addresses how, when, and why copulas work, including identification assumptions, data requirements, and boundary conditions. Second, we discuss variations in implementing copula correction. Third, we present the methodological background: how copula correction assumptions might be relaxed, its usage for panel data, how it is constructed, optimal estimation (for moderators and nonlinear effects), and obtaining standard errors. Fourth, we provide guidance for practical usage, including a flowchart 'cookbook' to check data requirements and assumptions at key steps. Fifth, we present two examples that follow the flowchart, using real world sales data. Finally, we conclude with discussions and future research directions.

THEORETICAL RATIONALE FOR ENDOGENEITY CORRECTION USING COPULAS

As an entry point, we first provide an overview for how copulas address endogeneity correction: why and when should they be used? How do they work? We examine what assumptions and data requirements are actually needed for model identification and discuss

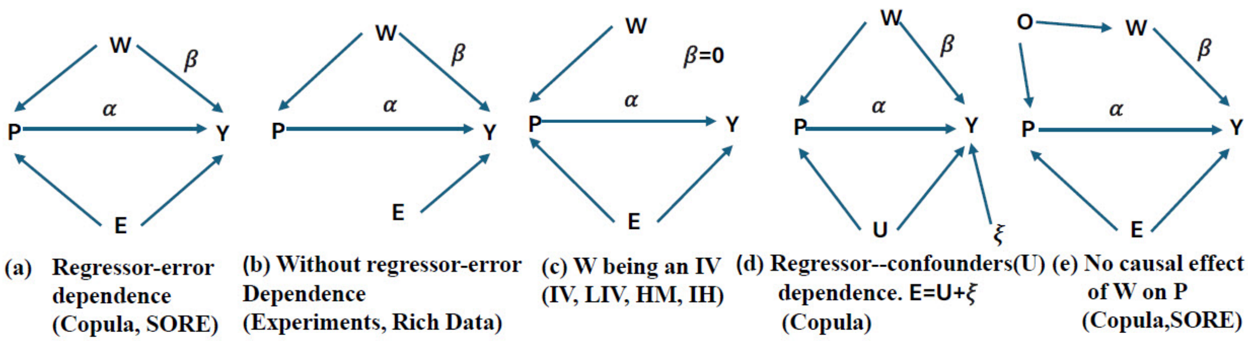

Figure 2: Directed Acyclic Graph (DAG) for Endogeneity

boundary conditions to guide appropriate use. This leads to the impact of copula correction.

Why and When Use Copula Correction?

Empirical examples of endogenous regressors abound. For concreteness, consider a running example of estimating the following linear structural model using nonexperimental data:

$$Y_i \quad = \quad \mu + \alpha P_i + \beta' W_i + E_i, \tag{1}$$

where $i = 1, \cdots, I$ indexes cross-sectional units or markets across spatial regions or over time; Y_i is a scalar response variable (e.g., log-transformed volume of ice cream sold in market i); P_i contains the endogenous regressor (log-transformed price), and W_i contains a vector of exogenous control variables affecting both the endogenous regressor P_i and the response Y_i (i.e., the two arrows from W to P and Y in Figure 2.a). The model parameters are (μ, α, β), among which α captures the causal or independent effect of P_i and is of primary interest. The exogenous control variables in W are determined outside the system (e.g., weather) or under control by researchers such that no dependence between W_i and E_i exists (i.e., no arrow between W and E in Figure 2.a) and thus $\texttt{Cov}(W_i, E_i) = 0$. Unlike W_i, the endogenous regressor P_i, however, can be affected by unobservables, such as unobserved common market shocks or product attributes contained in E_i, leading to the dependence between P_i and E_i (i.e., the arrow connecting E to P in Figure 2.a).

Copula endogeneity correction's advantages contributing to its wide usage include broad applicability and high feasibility, as compared with alternative methods (Table 1). The directed acyclic graph (DAG) in Figure 2.a explicitly includes the unobserved error term E and highlights the important role of P-E dependence. In this case, the distribution of the

Table 1: An Overview of Copula Correction with Alternative Approaches

Approach	Estimator	Description	Data Requirements	Main Assumptions	When to use
Experiment (Lab or field)	Between-group or within-group before-after comparisons for treatment effect	Random group assignment to avoid spurious association.	Random assignment of treatment. Categorical focal variable (not continuous).	No treatment-error dependence Randomization balances all confounders, good compliance to assigned treatment.	When feasible to manipulate focal variables without concerns of ethics and external validity; treatment is categorical.
Natural Experiment	Regression discontinuity, Interrupted time-series, DiD[a]	Leverage random event/threshold to determine causal effects.	Availability of random and exogenous event/threshold.	No concomitant occurrence of other confounding events around the focal event/threshold.	When natural event is available and special design and data requirements are fulfilled, and all confounders are accounted for.
Rich data	Regression adjustment, Matching, or Weighting[b]	Use a rich set of control variables or panel data to control for observed and unobserved effects.	All potential confounders are accurately measured or proxied for by the control variables or by panel data.	No regressor-error dependence given observed control variables and unobserved panel fixed effects; all control variables are exogenous.	When researchers are confident that the set of control variables or panel data modeling captures all potential confounders such that no regressor-error dependence exists.
Instrumental variables (IV)	Two-stage least squares (2SLS), Generalized method of moments (GMM), Control function[c].	Use observed IVs to address unobserved confounders.	Few control variables. Require IVs satisfying exclusion restriction (ER) and relevance.	No direct effects of IVs on the outcome (ER); IVs affect the endogenous regressor; all control variables are exogenous.	When the endogeneity concern exists in data at hand, and strong and valid IVs are identified and supported by institutional knowledge.
Latent IV	Likelihood-based estimation	Use latent discrete IVs to address unobserved confounders.	Few control variables. No observed IVs required. Endogenous regressors are required to be continuous and non-normal with a normal error distribution.	Latent IVs are discrete and satisfy ER and relevance conditions; all control variables are exogenous.	When the endogeneity concern exists in data at hand. When the latent discrete IVs can be justified by institutional knowledge. When endogenous regressors are continuous.
SORE	Likelihood-based estimation	Address regressor-error dependence via distribution-free odds ratio multivariate models that nests copulas as special cases.	Few control variables. No requirement of ER. Endogenous regressors are required to be nonnormal with a normal error distribution.	Distribution-free odds ratio multivariate models adequately captures regressor-error dependence; all control variables are exogenous.	When the endogeneity concern exists in data at hand. When valid or strong IVs are unavailable for all endogenous regressors. Can handle both continuous and noncontinuous endogenous regressors.
Copula	Control function (P&G[d], 2sCOPE[e], 2sCOPE-np[f]); Likelihood-based estimation (Haschka 2022). See flowchart in Fig. 5.	Address the regressor-error dependence via copulas. Copula control functions permit both copula and non-copula regressor-error dependence.	Few control variables. No requirement of ER. Either the model error or the endogenous part of error is normal. See flowchart Fig. 5 for specific data requirements.	GC adequately captures the dependence between regressor and the error (or the endogenous part of the error) (i.e., double robustness). A GC regressor-error dependence is sufficient but not necessary. All control variables are exogenous.	When the endogeneity concern exists in data at hand. When valid or strong IVs are unavailable for all endogenous regressors. Can handle normally distributed endogenous regressors. See Table 2 for common use cases. See Table 3 for data requirements.

Note: a: DiD: Difference in Difference. b: Regression adjustment includes methods such as OLS, random-effects and fixed-effects for panel data with unobserved effects. Matching (via propensity score, Mahalanobis matching, synthetic control, etc) and weighting (via inverse probability weighting) control for confounding effects by balancing the distributions of a rich set of control variables. c: See Petrin and Train (2010) for control function using IVs. d: Park and Gupta (2012). e: two-stage copula endogeneity correction (Yang, Qian, Xie 2024a); f: nonparametric 2sCOPE (Hu, Qian, Xie 2025). Methods in the table can be combined as a multi-methods approach to improving the applicability, robustness and quality of causal inference.

endogenous regressor P provides information about model parameters through its association with E. Thus, estimation methods ignoring the regressor-error dependence, such as ordinary least squares (OLS), assume the incorrect DAG in Figure 2.b and can yield severely biased model parameter estimates (i.e., endogeneity bias). Efforts to make P independent of E—such as through random assignment of P in experiments or by measuring and controlling for all confounders in W—are often impractical (Germann, Ebbes, and Grewal 2015). By contrast, copula correction does not impose the exogeneity assumption on all regressors as OLS does; it considers the general DAG in Figure 2.a that includes the DAG in Figure 2.b as a special case and requires neither experiments nor measuring all confounders (Table 1).

As the classical approach to addressing endogeneity bias, the IV method assumes the DAG in Figure 2.c, another special case of the DAG in Figure 2.a. The IV, W, needs to not only affect P (relevance), but also be exogenous (no arrow between W and E) and have no direct effects on Y (i.e., $\beta = 0$ in Figure 2.c, the untestable ER condition assuming no arrow between W and Y). The conditions of relevance and ER are often in conflict: although the IV approach has a strong theoretical basis, finding good and valid IVs can be very challenging in practice, which demands more flexible ways to handle regressor endogeneity. Other IV-free methods (LIV, IH, HM) decompose P into endogenous and exogenous parts, with the exogenous part (W in Figure 2.c) satisfying the ER condition (Park and Gupta 2012). Unlike IV and these other IV-free methods, copula correction needs not to argue for any exogenous variable in W to satisfy the ER condition ($\beta = 0$ in Figure 2.c) or to causally affect P.[4] Thus, copula correction substantially increases the feasibility of endogeneity correction.

To summarize, Table 2 lists some common use cases for copula correction.

How Does Copula Correction Work? — A Primer on Copula Correction

Copula correction, first proposed by Park and Gupta (2012) (P&G), is based on the idea that adequately capturing regressor-error dependence can resolve endogeneity issues and yield unbiased causal estimates. To address the endogeneity of P in Equation 1, P&G

[4] Empirical association between W and P (Figure 2.e) is sufficient (Yang, Qian, and Xie 2024a).

Table 2: Common Use Cases for Copula Correction

- When experiments are infeasible or cannot balance all confounders*; rich data are expensive, impossible to collect, or fail to completely/accurately measure all relevant confounders; or valid and strong IVs are unavailable.
- When one wants to conduct multi-methods causal inference as robustness checking to cross-validate each other (Germann, Ebbes, and Grewal 2015; Papies, Ebbes, and Van Heerde 2017; Qian and Xie 2024). Examples are when IVs exist but are imperfect with questionable validity or weak relevance; control variables included in rich data methods have questionable comprehensiveness, accuracy, or validity of exogeneity.
- When a combination of multiple methods is required to address endogeneity. For example, an IV for the treatment variable is available but potential moderators are endogenous and have no IVs available. In this case, copula correction can be used together with the IV to handle multiple endogenous regressors. Similarly, copula can be combined with other methods to address remaining endogeneity (e.g., after regression adjustment for a rich set of control variables) or used together with SORE (Qian and Xie 2024) to handle a mixture of continuous and discrete endogenous regressors.

*: Examples are (1) randomization of price levels conducted in a focal firm experiment may not balance competitors' responses to these price levels (Rutz and Watson 2019); (2) events or thresholds in natural experiments may be nonrandom and have concomitant events.

propose two estimation methods based on a Gaussian copula (GC) dependence model for the joint distribution of (P_i, E_i) and a normal structural error. The first maximizes the likelihood derived from the joint distribution of (P_i, E_i). The second uses a simpler control function approach that rewrites the maximum likelihood estimation as a regression augmented with copula generated regressors. Specifically, to address the endogeneity of P in Equation 1, the P&G control function approach adds a regressor $P_i^* = \Phi^{-1}(\widehat{F}(P_i))$ to the model, where Φ^{-1} is the inverse standard normal cumulative distribution function (CDF) and $\widehat{F}(P_i)$ is the empirical CDF of P. Both estimators were derived under completely specified model likelihood, leading to the belief that copula correction is likelihood-based (e.g., Table 1 in Eckert and Hohberger 2023). Consequently, estimation and inference may be sensitive to possible likelihood misspecifications and depend on a set of strict identification assumptions and data requirements (TAs 1 to 5 in Table 3).

Contrary to the belief, we show copula control function methods require neither a normal error distribution nor GC regressor-error dependence, and can work under substantially weaker conditions (Table 3). To demonstrate how copula correction works under these weaker conditions (and derive general copula control functions), consider the DAG in Figure 2.d which decomposes the structural error as $E_i = U_i + \xi_i$, where U_i is the error's endogenous part as the combined effect of unobserved confounders, ξ_i is an exogenous disturbance term satisfying $\mathbb{E}(\xi_i|P_i, W_i) = 0$, and $\mathbb{E}$ is the expectation operator. We rewrite Equation 1 as:

$$Y_i \;=\; \mathbb{E}(Y_i|P_i, W_i) + \epsilon_i = \mu + \alpha P_i + \beta' W_i \;+ \overbrace{\underbrace{\mathbb{E}(U_i|P_i, W_i)}_{Endogenous} \;+\; \underbrace{\epsilon_i}_{Exogenous}}^{\text{Structural error } E}. \tag{2}$$

Equation 2 decomposes the error E_i into two parts: (1) $\mathbb{E}(U_i|P_i, W_i)$: the expected omitted effect U_i given regressors (P_i, W_i), which is the error's part correlated with regressors and (2) the exogenous part ϵ_i uncorrelated with all regressors and $\mathbb{E}(U_i|P_i, W_i)$[5]. Important from Equation 2 is that one needs not to know unobserved U_i to control for endogeneity: $\mathbb{E}(U_i|P_i, W_i)$ sufficiently captures the relevant endogenous part of the structural error and completely controls the confounding effects of omitted variables. If $\mathbb{E}(U_i|P_i, W_i)$ is known and added to the outcome model as an offset term, all regressors are exogenous and the resulting regression yields consistent estimates of the structural model parameters, as the new error term ϵ_i is uncorrelated with all other right-hand side random variables $(P_i, W_i, \mathbb{E}(U_i|P_i, W_i))$. This approach clearly hinges on whether the control function $\mathbb{E}(U_i|P_i, W_i)$ can be recovered.

Copula correction based on Equation 2 proceeds by noting that the dependence between the endogenous regressor P_i and the omitted effect U_i unexplained by control variables in W_i can be captured by copula models. This copula dependence structure and economic theory[6] enable the derivation and recover of control functions that break the dependence between endogenous regressors and the structural error. As discussed in the next subsection, the copula dependence model is chosen for a number of merits, including its flexibility, wide applicability, and the ability to faithfully maintain regressor distributional features

[5]Note $\epsilon_i = \epsilon_i^U + \xi_i$ and both $\epsilon_i^U = U_i - \mathbb{E}(U_i|P_i, W_i)$ and ξ_i are uncorrelated with each of $(U_i, W_i, \mathbb{E}(U_i|P_i, W_i))$.

[6]While the first-stage model for endogenous regressors can be made nonparametric/assumption-lean in copula correction, economic theory can guide the choices of exogenous control variables in W which play an important role in model identification.

Table 3: An Overview of Identification Assumptions, Data Requirements, and Methodological Aspects of Copula Correction

Points to Consider	Traditional Beliefs	New Understandings and Recent Advances
Identification Assumptions & Data Requirements	**TA1.** ***The structural error distribution is normal***.[a]	***The structural error distribution can be nonnormal and left unspecified.*** (See Assumption 1 in Table 4, Feature 3 in Table 5).
	TA2. ***The GC model describes the joint distribution of regressors and the error term***[a].	***Permit both GC and non-GC regressor-error joint distributions.*** (See Assumption 2 in Table 4, Features 4 and 5 in Table 5).
	TA3. ***Endogenous regressors and exogenous regressors are independent.*** The copula term is $Cp_k=P_k^*=\Phi^{-1}(F(P_k))$[a], where $F(P_k)$ is the marginal cumulative distribution function (CDF) of the *k*th endogenous regressor P_k.	***Endogenous regressors and exogenous regressors can be correlated.*** With correlated endogenous and exogenous regressors, the general copula correction term is $C_{pk\|w}=\Phi^{-1}(F(P_k\|W))$[b], where $F(P_k\|W)$ is the conditional CDF of P_k given the exogenous regressors in W and can be replaced with model-free nonparametric estimates (2sCOPE/2sCOPE-np, See Features 6 and 9 in Table 5).
	TA4. ***Endogenous regressor is nonnormal***[a].	***Endogenous regressors can be normal.*** Relevant exogenous regressors can be leveraged to handle normally distributed endogenous regressors (2sCOPE/2sCOPE-np. See Feature 7 in Table 5).
	TA5. ***Endogenous regressors have sufficient support***[a].	***Can handle noncontinuous endogenous regressors*** (binary, discrete regressors with few levels, or semicontinuous regressors) via the likelihood-based SORE method, or via the copula control function (2sCOPE-np) (Feature 8 in Table 5).
Methodological Aspects		
Panel Data	***Copula correction for panel data can be handled in the same way as the cross-section data.***	***Copula correction for panel data requires special handling,*** such as fixed-effect transformation of panel outcome models (Haschka 2022) or proper calculation of copula correction terms based on time demeaned regressors (Table 7).
Estimation Approach	***Estimation is likelihood-based.*** Even if control functions are used, one needs to specify the joint distribution of the error term and regressors (Eckert and Hohberger 2023).	***Estimation can be likelihood-based, or alternatively via control functions without specifying the error and regressor distributions, the joint distribution of the structural error term and regressors, or the associated likelihood function*** (see Features 1 and 2 in Table 5). Thus, copula correction can be rendered less sensitive to distributional assumptions, and applicable to broader applications.
Models with intercept	***Sample size needs to be very large to avoid estimation bias if the structural model includes the intercept term*** (Becker, Proksch, and Ringle 2022).	***Judicious handling of copula transformation of regressors removes the bias described in Becker, Proksch and Ringle (2022)***. See the Methodological Background section and Equation 11 therein.
Endogenous interaction terms	***Adding copula correction terms for endogenous interaction terms is thought to help control their endogeneity,*** despite that it is sufficient to add copula correction terms for first-order endogenous terms only (Papies, Ebbes, and Van Heerde 2017).	Theoretical proof and empirical evaluation described in the Methodological Background section (1) extends the results of Papies, Ebbes, and Van Heerde (2017) to more general copula correction methods and (2) ***demonstrates a stronger result that adding the unnecessary high-order copula correction terms is suboptimal and has significant adverse effects***.

Note: **TA:** traditional assumption. [a]: See Park and Gupta (2012, 2024), Papies, Ebbes, and Van Heerde (2017), Becker, Proksch, and Ringle (2022), Haschka (2022), Eckert and Hohberger (2023), Papies, Ebbes, and Feit (2023), Qian and Xie (2024), Breitung et al. (2024), and Liengaard et al. (2024) for description of these traditional assumptions. [b]: Under a GC model for (P,W), this copula correction term reduces to $C_{pk|w}=P^*-\delta W^*$, which is the residual from a first stage model that regresses P* on W* and removes the effects of exogenous regressors from P* (Yang, Qian, and Xie 2024a).

and multivariate dependence crucial for model identification. Yang, Qian, and Xie (2024a) propose a two-stage copula endogeneity correction procedure (2sCOPE) that employs a joint GC distribution for all regressors and the error: the residual from a first-stage GC model for the endogenous regressor P_i breaks the regressor-error dependence and is the control function up to a constant. The two-stage approach using the first-stage residuals has been adapted to various settings (see the later Methodological Background section and Table 5). Hu, Qian, and Xie (2025) further develop a two-stage nonparametric copula control function (2sCOPE-np) procedure that generalizes and unifies these residual-based two-stage procedures; one can estimate the control function noparametrically without postulating specific first-stage auxiliary models for regressors. Under a normal distribution $N(0, \sigma_u^2)$ for U_i and a GC model for (U_i, P_i) conditional on W_i, Hu, Qian, and Xie (2025) show that Equation 2 becomes:

$$Y_i \;=\; \mu + \alpha P_i + \beta' W_i + \gamma C_{i,p|w} + \epsilon_i, \quad \text{where} \quad C_{i,p|w} = \Phi^{-1}(\widehat{F}(P_i|W_i)), \tag{3}$$

where $\gamma C_{i,p|w}$ is the control function $\mathbb{E}(U_i|P_i, W_i)$; $\gamma = \sigma_u \cdot \rho$ and ρ is the correlation coefficient in the GC distribution capturing the remaining dependence between P_i and U_i unexplained by exogenous regressors in W; $C_{i,p|w} = \Phi^{-1}(\widehat{F}(P_i|W_i))$ is the copula correction term, and $F(P_i|W_i)$ denotes the conditional CDF of P given W. Note $F(P_i|W_i)$ can be consistently estimated using a nonparametric conditional CDF estimator, $\widehat{F}(P_i|W_i)$, where W_i can contain both continuous and noncontinuous regressors. Consequently, one can add the copula term $C_{i,p|w} = \Phi^{-1}(\widehat{F}(P_i|W_i))$ as an additional control variable to correct for endogeneity: the standard estimation (e.g., OLS) of Equation 3 yields consistent estimates with $\hat{\gamma} C_{i,p|w}$ being the estimated control function. When $\rho = 0$ (i.e., no endogeneity), the true coefficient for $C_{i,p|w}$ is zero. Thus, with the copula control function approach, one can test the presence of endogeneity by statistically testing whether the coefficient γ for $C_{i,p|w}$ is zero or not.

Copula control function is an alternative to the control function of Petrin and Train (2010). Unlike Petrin and Train (2010), copula correction requires no IVs that must satisfy the strict ER condition, a stronger requirement than exogeneity. No arguments for the nature, direction, or forms of relationships between W and P are needed: empirical association between P and W is sufficient and 2sCOPE-np employs first-stage model-free control functions. Thus, copula correction greatly increase the feasibility of endogeneity correction.

For K continuous endogenous regressors $(P_1, \cdots, P_K)$, the 2sCOPE-np control function approach estimates the following augmented regression model:

$$Y_i \;=\; \mu + \sum_{k=1}^{K} \alpha_k P_{i,k} + +\beta' W_i + \sum_{k=1}^{K} \gamma_k C_{i,p_k|w} + \epsilon_i, \quad \text{where} \quad C_{i,p_k|w} = \Phi^{-1}(\widehat{F}(P_{i,k}|W_i)); \tag{4}$$

γ_k is the coefficient parameter for $C_{i,p_k|w}$, and $\sum_{k=1}^{K} \gamma_k C_{i,p_k|w}$ is the linear combination of the K copula terms $\{C_{i,p_k|w}\}$ used to control for endogenous regressors and is the copula control function (CCF) with multiple endogenous regressors.

Identification Assumptions and Data Requirements for Copula Correction

As shown above, copula correction can be achieved using control functions via the method of moments[7] estimation of the augmented regression in Equation 4 without specifying distributions for the error and individual regressors, the regressor-error joint distribution, or the associated likelihood (Table 3). This increases the robustness of copula correction and allows us to focus on the most essential assumptions and data requirements for copula correction. We now contrast the assumptions (Table 4) used to derive the above general 2sCOPE-np procedure with traditional assumptions (TAs) listed in Table 3.

Nonnormal error distribution and non-copula regressor-error dependence

Assumptions 1 and 2 (Table 4) mean that the copula control function methods require neither a normal error distribution nor GC regressor-error dependence (TAs 1 and 2 in Table 3) and can work under substantially weaker conditions than previously believed. In fact, the same 2sCOPE-np procedure can be derived under both Figure 2.a and Figure 2.d and hence possesses a desirable property of *double robustness* (Web Appendix B): when a GC model adequately captures either the regressor-error dependence or regressor-confounder dependence unexplained by exogenous regressors, the copula corrects endogeneity bias. The

[7] For instance, the OLS estimation is a spacial case of the method of moments by equating the sample covariances between regressors and the error term to the population covariance values (Wooldridge 2010).

Table 4: Assumptions Used to Derive 2sCOPE/2sCOPE-np

Assumption 1. *Either the error E_i or its endogenous component U_i is normally distributed (DR1).*
Assumption 2. *Either $(P, E)|W$ or $(P, U)|W$ follows a Gaussian copula*[a] *(DR2).*
Assumption 3. *Full rank of all regressors and* $\texttt{Cov}(W, E) = 0$. [b]
Assumption 4. *P is continuous and the copula term $C_{p|w}$ is linearly independent of* $(1, P, W)$. [c]

Assess Assumptions

a. Assumptions 1 and 2 are unverifiable from data alone. To assess their plausibility, consider the source of endogeneity and if needed revise the model (e.g., transform outcomes/regressors if U is suspected to be highly skewed) or copula correction strategies (Web Appendix C).
Checking Boundary Condition 1 below helps detect violations of the two assumptions.

b. To satisfy Assumption 3, researchers should take care to properly specify the structural model, such as avoiding mistakenly including redundant regressors. To ensure $\texttt{Cov}(W, E) = 0$, include only exogenous control variables in W based on institutional knowledge.

c. To verify Assumption 4, check Boundary Condition 1 for 2sCOPE-np. Assumption 4 for 2sCOPE is that the continuous P or one correlated and continuous regressor in W is nonnormal, which can be verified by Boundary Condition 2 below.

Boundary Conditions

1. Inspect the inflation of standard errors of copula-corrected estimates relative to those of uncorrected estimates; inflation > 6 suggests potential model identification or misspecification issues.
2. For 2sCOPE, the continuous P is nonnormal (normality test $p < 0.05$) or one exogenous W is continuous, sufficiently nonnormal (normality test $p < 0.001$), and sufficiently relevant (first-stage F statistics > 10).
3. A minimum sample size of 300 is recommended for satisfactory performance of 2sCOPE-np[d].

Note: DR: Double Robustness. Assumptions 1 and 2 are not strictly required and copula correction demonstrates robustness to a range of departures from them (Web Appendix B). [a]: Assumption 2 means that the GC model describes the $P - E$ (or $P - U$) dependence unexplained by exogenous control variables in W. 2sCOPE additionally assumes a joint GC model for all regressors (P, W). [b]: Assumption 3 is standard and required for other econometric methods, including OLS, IV regression, and P&G. Full rank means $\text{rank}(X^{'}X) = Q$, in which $X = (1, P, W)$ with column dimension of Q. [c]: Assumption 4 means that the endogenous regressor P can be normally distributed by leveraging exogenous regressors. 2sCOPE-np can also leverage exogenous regressors to handle noncontinuous P. [d]: See Web Appendix Figure W4.

reason is intuitive: the exogenous part of E_i, ξ_i, simply adds noise but does not affect endogeneity correction. Because ξ_i does not need to follow a normal distribution (or any GC assumption) for the augmented regression to correct bias, copula control functions do not require the error E_i to be normally distributed or follow a specific copula structure jointly with P_i. Consequently, the normal error distribution and the GC regressor-error dependence are only sufficient but not necessary conditions for copula control functions to work.

In many settings, it is plausible to assume that E_i is normally distributed (Yang, Chen, and Allenby 2003; Ebbes et al. 2005) or U_i is normally distributed as a sum of many confounders' effects (Qian and Xie 2024; Breitung, Mayer, and Wied 2024; Yang, Qian, and

Xie 2024a)[8], satisfying Assumption 1. Furthermore, in many settings, the GC model can adequately capture the dependence between U_i (or E_i) and P_i unexplained by exogenous regressors, satisfying Assumption 2. The GC model has desirable properties, making it widely used and applicable in empirical research to robustly capture multivariate dependence that traditional models, such as linear additive dependence models, often fail to capture (Danaher and Smith 2011; Qian and Xie 2024). GC permits the full (-1,1) range of correlation coefficient and is readily extensible and computationally scalable to more than two variables. Moreover, GC separates modeling dependence from modeling individual variables' distributions. Thus, distribution-free GC models can capture regressor-error dependence irrespective of (potentially complex) regressors' distributional features, while nonparametrically preserving these distributional features essential for model identification. Copula correction also demonstrates robustness to a range of departures from the GC assumption. Consequently, copula correction has broad applicability and become a valuable resource in the toolkit for handling regressor endogeneity in various fields (Figure 1). In many applications, including those in marketing (Web Appendix A), copula correction yields credible findings that are consistent with theoretical predictions, attesting to its effectiveness and applicability.

Correlated endogenous and exogenous regressors

The copula correction term $C_{i,p|w} = \Phi^{-1}(\widehat{F}(P_i|W_i))$ derived from 2sCOPE-np has high face validity and sheds light on prior copula control function methods. It is a statistically general and natural way of extending the P&G method to account for correlated exogenous and endogenous regressors. When P and W are independent, the conditional CDF $F(P_i|W_i)$ reduces to the marginal CDF $F(P_i)$ and the copula term reduces to $P_i^* = C_{i,p} = \Phi^{-1}(\widehat{F}(P_i))$ used in the P&G method (TA 3 in Table 3). But when P and W are dependent, using P^* as the copula correction term will confound the control function with effects of exogenous

[8] For example, U_i may be salesperson ability which combines many genetic and environmental factors, product quality which sums many unmeasured product attributes, or a category attribute which sums many UPC products' attributes. Normality of U_i here requires an enough number of composite confounders rather than a large sample size (Billingsley 1995, Section 27, The Central Limit Theorem).

regressors, biasing both control function and model parameter estimates.

The general copula correction term $C_{i,p|w} = \Phi^{-1}(\widehat{F}(P_i|W_i))$ removes the influences of exogenous regressors on the entire distribution of P rather than just on some aspects (e.g. the mean) of this distribution. To account for correlated exogenous regressors, some methods propose using a mean regression model for P given W (e.g., Breitung, Mayer, and Wied 2024). Such methods implicitly rely on the assumption that W only affects the mean of P, which is known to be violated for bounded, truncated, or discrete endogenous regressors[9]; this assumption can also be questionable for unbounded continuous regressors (Chen 2007; Danaher and Smith 2011) and can introduce bias to control function estimates and model estimates. The 2sCOPE procedures are more flexible, permitting W to affect not only the mean but also higher moments of P and providing a model-free nonparametric way to estimate $F(P_i|W_i)$ and account for exogenous regressors. Although correlated exogenous regressors complicate copula correction and need to be carefully dealt with, they also provide opportunities to relax key constraints of copula correction, as seen below.

Normally distributed endogenous regressors

Under P&G, the source of identification comes from distributional features of the endogenous regressor P (nonnormally distributed). If P is normally distributed, then P&G fails by violating the full rank condition of the regressor matrix (TA 4 in Table 3). Although distributional shapes of endogenous regressors are observed and regressor normality can be tested, this requirement limits the applicability of copula correction because many important endogenous regressors have close to normal distributions (Eckert and Hohberger 2023).

Recent 2sCOPE procedures relax this restriction by leveraging relevant exogenous control regressors to identify the effects of endogenous regressors with insufficient nonnormality.

[9]Examples include the percentage of trained salespeople that takes on continuous values in [0, 1] (Atefi et al. 2018), or brand price that takes on values between minimum and maximum prices (Qian and Xie 2011). One should not confuse the first-stage models in copula correction with those in two-stage least squares (2SLS) using IVs. The simple linear additive equation commonly used in the first stage mean regression for endogenous regressors in 2SLS is not a dependence model or DGP, but merely a projection of endogenous regressors to the space of exogenous regressors (Wooldridge 2010). 2SLS achieves identification through exclusion restriction rather than dependence modeling.

Under 2sCOPE procedures, the source of identification can come either from features of the conditional distribution $F(P|W)$ (nonnormal conditional distribution) or from nonlinear relationships between P and W, in order to ensure the full rank of the copula augmented model in Equation 4 (Assumption 4 in Table 4). Even if P is normally distributed, Assumption 4 is satisfied and the copula model is identified when a continuous exogenous regressor in W is nonnormally distributed and nonlinearly related to P (Yang, Qian, and Xie 2024a).

Noncontinuous endogenous regressors

Noncontinuous endogenous regressors, such as binary, count with small means, or semicontinuous, have many ties (observations with the same value). Such endogenous regressors have insufficient support and can cause identification problems in copula correction (TA 5 in Table 3) because nonuniqueness of the inverse discrete CDF can bias copula dependence estimation in control functions (Qian and Xie 2024).[10] Recent research shows the 2sCOPE-np control function can leverage variations in exogenous regressors to overcome plateaus in discrete CDFs and the resulting identification challenges (Hu, Qian, and Xie 2025): 2sCOPE-np treats noncontinuous endogenous regressors as discretized forms of conceptually continuous variables, replacing these endogenous regressors' discrete *marginal* CDFs with smoothed, continuous *conditional* CDFs given exogenous regressors. Unlike latent variable models, 2sCOPE-np imposes no distributional assumptions on the conceptually continuous variables.

Boundary Conditions for Applying Copula Correction

The considerations above show that copula control function methods work under substantially weaker conditions and thus are more robust and applicable than previously believed. When properly applied with the underlying assumptions and data requirements being met, copula correction can be a powerful tool for addressing endogeneity bias using nonexperimental data. However, it is important to also check boundary conditions to minimize the potential pitfalls and inappropriate use of copula correction.

[10] Their proposed likelihood-based SORE approach can bypass such inverse mapping and thus can handle noncontinuous endogenous regressors.

Copula correction can be ill-behaved when copula terms exhibit severe collinearity with existing regressors.[11] Even for theoretically identified models, strong collinearity inflates standard errors of copula corrected estimates, reduces statistical power of hypothesis testing, and inflicts finite-sample estimation bias. While 2sCOPE enables identification of nearly normal endogenous regressors, it relies on availability of sufficiently relevant exogenous regressors that also exhibit enough nonnormality to prevent severe collinearity. Yang, Qian, and Xie (2024a) provide guidelines (Boundary Condition 2 in Table 4) on the threshold values of the first-stage F statistic (for relevance) and the p-value (for nonnormality) of candidate exogenous regressors, assuming a joint GC model for all regressors and error. When the more general 2sCOPE-np is used, one can check the inflation of standard errors of copula corrected estimates relative to the uncorrected estimates for signs of severe collinearity and weak nonidentification (Boundary Condition 1 in Table 4).

Assumptions 1 and 2 are not strictly required. Copula correction demonstrates robustness to a range of departures from these two assumption such as symmetric nonnormal distributions for U and certain non-Gaussian copula structures (Web Appendix B Table W5; Haschka 2022; Yang, Qian, and Xie 2024a), but can be adversely affected by gross violations of these two assumptions. Because neither E nor U is observed, these two assumptions are difficult to verify. To assess their plausibility, one can consider the source of endogeneity in the focal application. Theoretical considerations can help assess the distribution of U (see Footnote 8) and the dependence structure between U and endogenous regressors. Because copula control function permits non-normal error distributions, nonnormal residual distributions do not mean copula control function would fail.

However, if the distribution of U is suspected to be highly skewed, it is advisable to revise model specifications (e.g, transform variables or add more control variables). Misspecifying the regressor-error dependence can induce weak identification and large standard errors of

[11] The copula terms capture the error's endogenous part and thus are expected to correlate with endogenous regressors. However, copula augmented models become weakly identified or non-identified when copula terms are severely or perfectly collinear with existing regressors (i.e., failure of Assumption 4 in Table 4). This can occur when regressors and the error jointly follow a multivariate normal distribution (Web Appendix C Table W6).

coefficient estimates of endogenous regressors (Web Appendix Table W7). We develop a new diagnostic statistic, named ICON, that checks the ratios of standard errors of copula-corrected estimates to those of the uncorrected estimates for signs of GC misspecifications (i.e., Boundary Condition 1 in Table 4). When GC dependence misspecification is suspected, one can revise model specifications by including more relevant control variables, refining copula correction approaches (Web Appendix Table W8), or use other endogeneity correction methods. Assumption 2 only requires unexplained dependence between P and U given W to be captured by GC, which implies adding relevant control variables to the model may make Assumption 2 plausible even if unconditionally P and U do not follow a GC model.

Impacts of Copula Correction

In many cases, copula correction provides a feasible approach to controlling for the thorny regressor endogeneity issue and offers opportunities for optimal managerial decision making, as illustrated in the following running example.

Example 1: Price Sensitivity Estimation. Store managers and policy-makers are often interested in learning price sensitivity for category demand growth. This example estimates price sensitivity for the diaper category using store scanner purchase data from the IRI Academic data set for one focal store in the Buffalo, NY market from 2002-2006 (261 weeks). Here, price may be endogenous because of unobserved variables (e.g., product characteristics, retailer pricing decisions, number of shelf facings) that, when omitted from a model, become part of the structural error. It is expected that these unobserved characteristics induce positive correlation between price and the error term, thereby causing the OLS estimate of price sensitivity to bias toward zero (i.e., less negative). As shown in a later section, the OLS price elasticity estimate here is -1.367, which is significantly less than the copula corrected price elasticity estimate of -2.205 (Figure 3), a 61% difference reflecting a large impact of a "wrong" estimate. The manager will underestimate consumer price sensitivity using OLS, and mistakenly set the price too high, resulting in lost revenue and profit. The analysis in

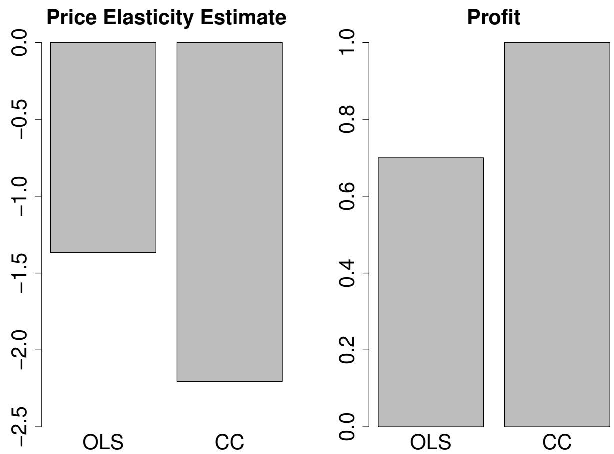

Figure 3: Example 1: Impact of copula correction on price sensitivity estimation. OLS: ordinary least squares; CC: copula correction.

the later section shows that using the OLS price estimate will yield 30% less profit compared to using the copula corrected price sensitivity estimate (Figure 3).

Meta-analyses of studies that compare estimates after endogeneity correction to uncorrected estimates also find similar differences. Bijmolt, Van Heerde, and Pieters (2005) found price elasticity was -2.47 without endogeneity correction, but -3.74 when corrected. Sethuraman, Tellis, and Briesch (2011) found "Advertising elasticity is lower when endogeneity in advertising is not incorporated in the model" (p.470).[12] With personal selling (i.e., salesforce), models that account for endogeneity have lower elasticity (.282) than models without endogeneity correction (.373), a significant difference of 0.091 that importantly represents an over-estimation of 32% (Albers, Mantrala, and Sridhar 2010). The importance of endogeneity correction is apparent: Without correction, managers and academics likely experience underestimated effects of pricing and advertising and overestimated effects of salesforce.

VARIATIONS IN IMPLEMENTING COPULA CORRECTION

Given the advances in copula correction, appreciable variation in copula endogeneity correction usage exists in practice. These variations can greatly affect the performance of copula correction, which call for clear guidelines in optimal copula correction given the importance

[12] Sethuraman, Tellis, and Briesch (2011) note that the bias when ignoring endogeneity will depend on the relationship between the omitted variable (e.g., price, product, or promotions), the endogenous variable (advertising), and the dependent variable (sales). For instance, price, when omitted, should bias advertising's effect downward: price has (-) relationship to sales, but (+) with advertising (i.e., high price brands advertise; low price brands let their price do the 'selling').

of endogeneity correction and copula correction's growing popularity. Becker, Proksch, and Ringle (2022) discovered substantial bias of P&G's copula corrected parameter estimates if the structural model contains the intercept, and cautioned using copula correction in such models with small to moderate sample sizes. We study this issue and evaluate an alternative implementation of copula transformation that has strong theoretical justification and avoids such bias. Recent work also shows that failure to account for exogenous regressors correlated with endogenous regressors can adversely affect copula correction effectiveness in eliminating endogeneity bias (Haschka 2022; Qian and Xie 2024; Yang, Qian, and Xie 2024a).

Another important issue arises regarding the best way to address endogeneity bias for models containing higher-order terms of endogenous regressors (such as interactions or moderators). Many applications in different fields are often interested in estimating structural models with higher-order terms of endogenous regressors, studying moderators of causal relationships to determine optimal policy and managerial intervention. Considerable inconsistencies exist regarding how to handle higher-order terms of endogenous regressors in copula correction (Web Appendix Table W3). While some studies exclude copula generated regressors for endogenous higher-order terms (often without stating the reason), others argued for including these generated regressors to control for endogeneity. To illustrate the impact of variations in using copula correction, consider the following running example.

Example 2: Moderator of Price Sensitivity Price may work together with a retail store's feature advertising to achieve synergistic effects on sales. This can be tested by estimating the interaction term between price and feature advertisement in a sales model, with feature advertisement as a potential moderator of price. Blattberg and Neslin (1990) note that feature advertising "may interact with price discounts. If the consumer is not informed that a price discount is offered, the price elasticity is likely to be small" (p.347). This suggests a negative sign for the interaction term between price and feature advertisement.

Figure 4 shows mean price sensitivity estimates per quartile of feature intensity in the peanut butter category, predicted from a sales demand model with an interaction term

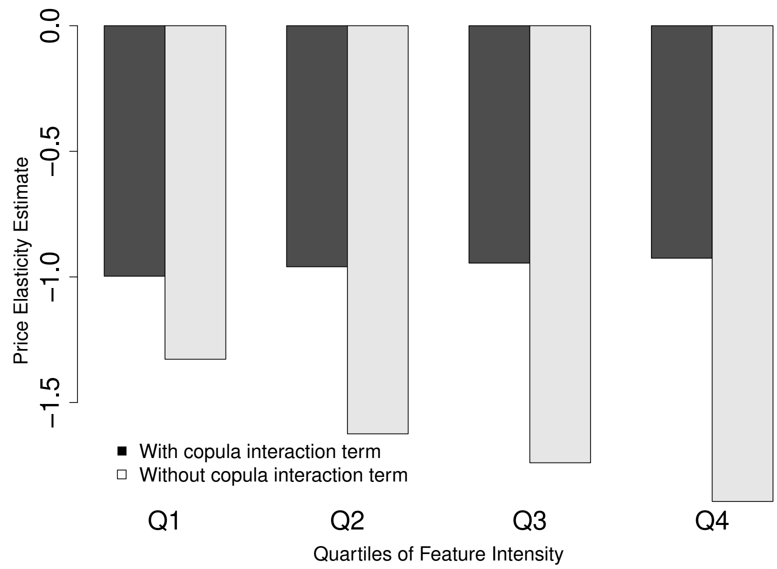

Figure 4: Mean price sensitivity estimates by quartile of feature intensity.

between price and feature, also using the IRI academic data for a New York City store. The black (gray) bars are price sensitivity estimates estimated with (without) a copula term for the interaction term. Including the copula term for the interaction yields similar price sensitivity estimates across different feature intensity (i.e., lack of interactive effect); excluding the copula term yields a greater magnitude of price sensitivity, and estimated price sensitivity increases with greater feature advertisement. As shown later, adding the copula term for the interaction term can induce bias and increase parameter estimate variability.

METHODOLOGICAL BACKGROUND

In this section, we discuss methodological aspects of copula endogeneity correction. We first acquaint readers with recent advances in copula correction, then speak to copula correction in panel data, show proper construction of the copula, address inconsistencies in copula correction (including for higher-order terms), and lastly how to obtain standard errors.

Methods to Relax Assumptions of Copula Correction

Recent methodological developments relax key assumptions and data requirements of the P&G method, which considerably widens the applicability of copula correction. These methods differ in their features (Table 5) and are broadly grouped into two classes: moment-

Table 5: Copula Correction Methods with Enhanced Capabilities.

Features	Methods
1. Control function approach without specifying model likelihood	2sCOPE procedures[a]
2. Likelihood-based joint estimation	Haschka (2022); SORE
3. Permit nonnormal structural error	2sCOPE procedures[a]
4. Permit both GC and non-GC regressor-error dependence	SORE; 2sCOPE-np Yang, Qian, and Xie (2024a,b); Liengaard et al. (2024)
5. Permit both GC and non-GC regressor-confounder dependence[c]	SORE; 2sCOPE-np
6. First-stage model-free control function	2sCOPE-np
7. Handle normal endogenous regressors	Yang, Qian, and Xie (2024a,b); Liengaard et al. (2024) 2sCOPE-np
8. Handle discrete or semicontinuous endogenous regressors	SORE; 2sCOPE-np
9. Handle noncontinuous and continuous exogenous regressors correlated with endogenous regressors & handle nonlinear terms such as interactions.	Haschka (2022); SORE 2sCOPE procedures[a]
10. Handle heterogeneous copula structure over levels of discrete exogenous regressors	Liengaard et al. (2024); Yang, Qian, and Xie (2024b) 2sCOPE-np

Note: SORE: Semiparametric odds ratio endogeneity model (Qian and Xie 2024). [a]: 2sCOPE procedures are defined broadly here to include Yang, Qian, and Xie (2024a,b); Liengaard et al. (2024); Breitung, Mayer, and Wied (2024); Mayer and Wied (2025); Hu, Qian, and Xie (2025). Hu, Qian, and Xie (2025) (2sCOPE-np) unifies the 2sCOPE procedures by employing first-stage model-free nonparametric copula control functions.

based two-stage control function methods and likelihood-based methods.

Two-stage control function methods

Yang, Qian, and Xie (2024a) propose a flexible and feasible two-stage copula endogeneity correction (2sCOPE) framework using control functions. The 2sCOPE framework extends the P&G method in three important aspects. First, unlike P&G, 2sCOPE adds the first-stage residual terms as the control function instead of $P^* = \Phi^{-1}(F(P))$. As a result, the control function in 2sCOPE accounts for correlated exogenous regressors (Feature 9 in Table 5). Second, 2sCOPE relaxes the P&G's data requirement that endogenous regressors must be nonnormally distributed (Feature 7 in Table 5). Third, while exogenous regressors are not used for generating copula terms in P&G, 2sCOPE can leverage these exogenous regressors to sharpen the structural model estimates. If a powerful exogenous regressor is available and included in the model to generate control functions, 2sCOPE can eliminate P&G's finite sample bias caused by insufficient nonnormality of endogenous regressors, and increase the accuracy of parameter estimates. As shown in Yang, Qian, and Xie (2024a), 2sCOPE increases modeling robustness and reduces dependence on model assumptions as compared with the P&G method. As a result, 2sCOPE has increased robustness to small sample size, normality of endogenous regressors, and violations of Gaussian copula dependence.

As noted previously, the copula term $P^* = \Phi^{-1}(F(P))$ used in P&G does not use exogenous regressors in W, which can cause incomplete correction of endogeneity bias when P^* is correlated with these exogenous regressors. 2sCOPE corrects this issue by introducing a two-stage process. The idea of 2sCOPE is to remove P^*'s component that is correlated with exogenous regressors, using the remaining cleaned part of P^* to control for endogeneity. Yang, Qian, and Xie (2024a) show that this can be achieved by using properly constructed residuals from a first-stage auxiliary model for regressors. Under the assumption of Gaussian copula for (P, W, E) or (P, W, U) — where U is the endogenous part of E in Figure 2.d — for K continuous endogenous regressors $(P_1, \cdots, P_K)$, the 2sCOPE procedure described in Table 6 estimates the following augmented regression model:

$$Y_i \quad = \quad \mu + \sum_{k=1}^{K} \alpha_k P_{i,k} + \beta' W_i + \sum_{k=1}^{K} \gamma_k C_{i,p_k|w} + \epsilon_i, \tag{5}$$

$$\text{where } C_{i,p_k|w} \quad = \quad V_{i,k} = P^*_{i,k} - \widehat{\delta}'_k W^*_i, \tag{6}$$

$P^*_{i,k} = \Phi^{-1}(\widehat{F}_{P_k}(P_{i,k}))$ and $W^*_{i,l} = \Phi^{-1}(\widehat{F}_{W_l}(W_{i,l}))$ for the lth $(l = 1, \cdots, L)$ variable in W are copula transformations of the model regressors.[13] 2sCOPE performs well with multiple continuous endogenous regressors and multiple exogenous regressors (both continuous and discrete control covariates: Web Appendices E2 and E3 in Yang, Qian, and Xie 2024a).[14]

[13] The algorithm in the later Equation 11 is used for the copula transformation of these regressors including noncontinuous exogenous regressors in W.

[14] One could also eliminate discrete control covariates from the model before applying 2sCOPE by using within group demeaning of the outcome and continuous regressors with groups formed by combinations of discrete covariates, in a similar way to the fixed-effect transformation of panel data to remove fixed-effects. Alternatively, one can condition on discrete endogenous regressors and apply Stage 1 of 2sCOPE to only group-demeaned continuous regressors and include residuals as generated regressors (Table 7), while leaving outcomes unchanged.

Table 6: Summary of the 2sCOPE and 2sCOPE-np Estimation Procedures

Stage 1:

- Obtain empirical CDFs for each regressor in P_i and W_i, $\widehat{F}_{P_k}(\cdot)$ and $\widehat{F}_{W_l}(\cdot)$;
- Compute $P^*_{i,k} = \Phi^{-1}(\widehat{F}_{P_k}(P_{i,k}))$ and $W^*_{i,l} = \Phi^{-1}(\widehat{F}_{W_l}(W_{i,l}))$ using copula transformation algorithm defined in Equation 11;
- Regress $P^*_{i,k}$ on W^*_i and obtain residual $C_{i,p_k|w} = P^*_{i,k} - \delta'_k W^*_i$ (Equation 6), which removes the component related to exogenous regressors.

Stage 2:

- Add $C_{i,p_k|w}$ to the outcome structural regression model as a generated regressor to control for endogeneity of P_k. The augmented regression model takes the form of Equation 5 (or Equation 14 when the model contains higher-order or interaction terms of regressors).

Note: Table presents the 2sCOPE procedure. 2sCOPE-np replaces Stage 1 with the following: for the kth endogenous regressor P_k, perform nonparametric kernel conditional CDF estimation for P_k given W (Web Appendix Equation W35 or Equation W37) and obtain the copula term $C_{i,p_k|w} = \Phi^{-1}(\widehat{F}(P_{i,k}|W_i))$, which involves no copula transformations of individual regressors. For both 2sCOPE and 2sCOPE-np, W can contain a mixture of continuous and noncontinuous variables (See Footnote 14 for alternative 2sCOPE implementations that bypass copula transformations of noncontinuous variables in W).

The two-step 2sCOPE procedure first regresses each $P^*_{i,k}$ on W^*_i and then adds these first-stage residual terms $V_{i,k} = P^*_{i,k} - \widehat{\delta}'_k W^*_i$ to control for endogeneity, where $\widehat{\delta}_k$ denotes first-stage coefficient parameter estimates.[15] By conditioning on the first-stage residual $V_{i,k}$ (the component in P_k that causes endogeneity but is uncorrelated with exogenous regressors), the new error ϵ_i becomes independent of all regressors (P_i, W_i, V_i), thereby ensuring the consistency of standard estimation methods.

The 2sCOPE is derived from the assumptions in Table 4. Assumptions 1 and 2 mean that 2sCOPE also has the double robustness property: the error term does not need to be normally distributed and regressor-error dependence does not need to follow a GC relationship as long as GC adequately captures the dependence between regressors and U_i. However, the pairwise dependence between the endogenous regressor P_i and U_i unconditioned on W is restricted to a GC relationship, which is less general than 2sCOPE-np that permits both GC and non-GC pairwise dependence between P_i and U_i (Feature 5 in Table 5). Assumption 4 is less stringent than P&G's requirements that P has a nonnormal distribution and is uncorrelated with W. Even if the endogenous regressor is normally distributed, 2sCOPE can identify the model

[15] Although $P^*_i = \delta' W^*_i + V_i$ includes no intercept, the implementation of 2sCOPE includes the intercept, which is more general and performs well in simulation studies.

as long as one correlated W is continuous[16] and nonnormally distributed, which is feasible in many empirical applications. The 2sCOPE assumes the regressor-error GC correlation structure is constant and does not vary in the population. Recent studies (Liengaard et al. 2024; Yang, Qian, and Xie 2024b) relax this assumption through a robustness check by permitting the GC dependence structure and 2sCOPE copula terms to vary by the levels of discrete exogenous regressors (Web Appendix Table W19)[17].

Breitung, Mayer, and Wied (2024) and Mayer and Wied (2025) propose using a first-stage mean regression model for endogenous regressors to account for correlated exogenous regressors. Like 2sCOPE, their approach is relatively easy to apply and allows nonnormal error. Interestingly, although it originates from copula correction and permits non-GC regressor-error dependence, their approach does not permit GC regressor-error dependence in general (Table 5) and can yield biased estimates when regressor-error follows GC dependence (Hu, Qian, and Xie 2025). Implicitly, their approach assumes (1) a degenerated GC dependence[18] between U_i and the unobserved error parts in the first-stage models for endogenous regressors and (2) W affects only the mean but not higher-moments of the conditional distribution for $P|W$. Copula procedures using more flexible multivariate dependence models (i.e., methods with Feature 4 in Table 5) can better account for correlated exogenous regressors and also permit both GC and non-GC types of regressor-error dependence. In particular, the nonparametric 2sCOPE-np procedure fully nests their approach as a special case.

As described earlier, Hu, Qian, and Xie (2025)'s 2sCOPE-np procedure unifies all existing copula correction methods. The 2sCOPE-np employs nonparametric copula control functions (Table 6) that generalize and make robust the existing copula correction methods using model-based first-stage residuals. It possesses a number of substantial merits, including first-stage model-free without postulating specific dependence models among regressors, ability to

[16]Discrete exogenous regressors with few levels have high multicollinearity with their copula transformed values and thus are uninformative to help identify models with normally distributed endogenous regressors.

[17]These methods require a sufficient sample size and data requirements (shown later in the Flowchart in Figure 5) being met within each level of combinations of discrete exogenous regressors.

[18]Specifically, the correlation coefficient in the GC model is fixed at 1 or -1 (i.e., a deterministic relationship) such that U_i is a linear function of the copula transformed error term for the endogenous regressor.

permit both GC and non-GC types of regressor-error (or regressor-confounder) dependence, and to better handle normal or discrete endogenous regressors (Features 4,5,6,7,8 in Table 5). These merits significantly increase the applicability and robustness of copula correction. The nonparametric feature of 2sCOPE-np does require a larger sample size (a minimum sample size of 300, see Table 4) for satisfactory performance and greater computation cost for associated kernel conditional CDF estimation (Hu, Qian, and Xie 2025).

Likelihood-based copula correction procedures

Haschka (2022) and Qian and Xie (2024) develop likelihood-based methods that generalize P&G (Table 5). Here we will describe Qian and Xie (2024) and then Haschka (2022), which was developed in the context of panel data to be described next. Qian and Xie (2024) propose a bias correction procedure that accounts for regressor-error dependence using a flexible semiparametric odds ratio endogeneity (SORE) model. The semiparametric odds ratio model is often used in marketing and other fields as a flexible multivariate model to measure dependence (Chen 2007), model multivariate missing data and selective sampling (Qian and Xie 2011, 2022), and combine data with sensitive elements (Qian and Xie 2014, 2015; Feit and Bradlow 2021). The SORE model encompasses a number of existing dependence models (including copulas), capable of capturing both GC and non-copula dependence structures. SORE requires a special estimation algorithm that eliminates potentially high-dimensional nuisance parameters in the nonparametric baseline distribution function, and maximizes the profile likelihood concentrating on the parameter of interest. Likelihood-based model selection measures (such as AIC and BIC) help select proper odds ratio dependence functions, encoding regressor endogeneity and identification strategies.

Distinct from other IV-free methods except 2sCOPE-np[19], SORE can handle noncontinuous endogenous regressors (Feature 8 in Table 5). Unlike copula control function methods, SORE can bypass the inverse CDF mapping of discrete endogenous regressors entirely and thus can achieve identification for noncontinuous endogenous regressors without any help

[19] Unlike 2sCOPE-np, SORE does not require the availability of powerful exogenous control covariates.

from exogenous regressors. In this aspect, SORE nests likelihood-based GC models for non-continuous endogenous regressors and offers more identification strategies for noncontinuous endogenous regressors. Furthermore, handling noncontinuous endogenous regressors with SORE is straightforward because SORE conditions on exogenous regressors and employs simple-to-evaluate likelihood involving no integrals with respect to latent variables. Thus SORE is applicable to many applications involving noncontinuous endogenous regressors. By contrast, more methods can handle noncontinuous exogenous control regressors (Feature 9 in Table 5 and Footnote 14).

Control Function vs Likelihood-based Correction Methods

Comparing methods, SORE is the likelihood-based method most similar in features to the moment-based 2sCOPE-np (Table 5). Generally, SORE employs one-step estimation, potentially offering greater efficiency (smaller standard errors) and allows well-established likelihood-based tests and model comparisons. In contrast, the moment-based 2sCOPE-np requires fewer assumptions and is computationally simpler. While the SORE model nests the copula dependence model and is capable of explicitly capturing both GC and non-copula dependence structures, the 2sCOPE-np also permits departures from GC regressor-error dependence by modeling regressor-confounder dependence. Thus, SORE and 2sCOPE-np are complementary to each other and non-hierarchical (one not nesting the other).

Copula Correction in Panel/Clustered Data

Copula correction can also address various sources of bias in panel data (Park and Gupta 2012; Haschka 2022; Yang, Qian, and Xie 2024a,b). Haschka (2022) generalizes copula endogeneity correction to the following fixed-effects (FE) panel data model:

$$y_{it} = \mu_i + P_{it}'\alpha + W_{it}'\beta + e_{it}, \tag{7}$$

where y_{it} denotes the dependent variable (e.g., store sales) for cross-sectional unit $i = 1, \cdots, N$ at occasions $t = 1, \cdots, T$; the fixed effect parameter μ_i captures the effects of time-constant (unobserved) variables (e.g., store size and market characteristics that do not change over time); P_{it} denotes endogenous regressors (e.g., price) such that $\texttt{Cov}(P_{it}, e_{it}) \neq 0$ due to time-varying unobservables (e.g., unmeasured consumer tastes or brand attributes varying over time), where the error $e_{it} \sim N(0, \sigma_e^2)$; W_{it} denotes exogenous control variables (e.g., prearranged promotions, quarter time periods). The parameters α and β capture the effect of P_{it} and W_{it}, respectively. Given fixed-effects μ_i, all regressors in (P_{it}, W_{it}) must be time-varying. Since fixed-effect parameters μ_i can be correlated with the regressors P_{it} and W_{it}, the fixed-effects transformation (Wooldridge 2010, p.302-303) is often used to eliminate these incidental intercept parameters. Because fixed-effects transformation changes the panel error structure to be nonspherical (nondiagonal covariance matrix), the GLS transformation is applied to handle nonspherical errors and collapses panel data to pooled observations with spherical errors $\widetilde{\xi}_{it} \overset{iid}{\sim} N(0, \sigma_\xi^2)$. After eliminating the nonspherical error problem, Haschka (2022) developed an efficient MLE estimation procedure that maximizes the likelihood of a GC model for the error and all explanatory variables to address regressor endogeneity.

Panel studies often need to consider slope heterogeneity. As shown in extant marketing studies, consumers' heterogeneous responses to marketing mix variables (e.g., price slope coefficients) are ubiquitous and substantial bias can arise when ignoring such slope heterogeneity. Thus, it is important to allow for individual-specific slope coefficients in marketing studies, by employing panel data models with random coefficients or mixed-effects (i.e., both fixed-effects and random coefficients). Extending the copula MLE method to these more general models with endogenous regressors can be challenging, because the model likelihood contains new intractable integrals of complex functions that involve products of copula density functions (Yang, Qian, and Xie 2024a). Copula correction implementation in these general panel data models is not well developed and understood.

For greater generality and computational tractability, Yang, Qian, and Xie (2024a,b) propose copula control function approaches for the following more general panel data model:

$$y_{it} = \mu_i + \alpha_i P_{it} + \beta_i' W_{it} + e_{it}, \tag{8}$$

where individual-specific parameters $(\mu_i, \alpha_i, \beta_i)$ can be treated as fixed-effects, random-

Table 7: Copula Control Function Estimation Procedures for Panel/Clustered Data

Stage 1:

- Do time demeaning of $P_{it,k}$ and W_{it} within each panel unit, and obtain the demeaned regressors $(\widetilde{P}_{it,k}, \widetilde{W}_{it})$.
- Apply 2sCOPE/2sCOPE-np in Table 6 to the demeaned regressors $(\widetilde{P}_{it,k}, \widetilde{W}_{it})$ and obtain $C_{it,k} = (\widetilde{P}_{it,k})^* - \hat{\delta}_k'(\widetilde{W}_{it})^*$ for 2sCOPE or $C_{it,k} = \Phi^{-1}(\widehat{F}(\widetilde{P}_{it,k}|\widetilde{W}_{it}))$ for 2sCOPE-np.

Stage 2:

- Add copula term $C_{it,k}$ as the generated regressor to control for endogeneity of $P_{it,k}$. The augmented panel regression model takes the form of Equation 10.

Note: Time demeaning subtracts each unit's averages over time of $P_{it,k}$ and W_{it} from the original values of $P_{it,k}$ and W_{it}. Thus, time demeaning removes the effects of all time-constant confounders correlated with μ_i and is needed for handling endogenous regressors that vary over both i and t. Endogenous regressors that vary only over t or only over i do not need time-demeaning. The algorithm can also be applied to grouped/clustered data formed by discrete exogenous regressors, in which i indexes the ith group/cluster.

effects, or a mixture of fixed-effects and random-effects. The model includes the FE panel model in Equation 7 as a special case. Their copula control functions involve no numerical integrals and can be implemented straightforwardly using standard software programs, assuming all regressors are exogenous.

In principle, the general copula term can extend to the panel data setting as $\Phi^{-1}(F(P_{it,k}|W_{it}, D_i)$, where D_i is the dummy variable for unit i to account for panel data structure. This may involve a high-dimensional conditional CDF estimation, which can be computationally intensive. To balance robustness and computational ease, we propose using the following general location GC model (Yang, Qian, and Xie 2024b) to calculate proper copula terms in multilevel data[20] when $\texttt{cov}(P_{it}, e_{it}) \neq 0$, $\texttt{cov}(P_{it}, \mu_i) \neq 0$, and/or $\texttt{cov}(P_{it}, W_{it}) \neq 0$:

$$p_{it} = \alpha_{i,p} + e_{it,p}, \quad \text{and} \quad w_{it} = \alpha_{i,w} + e_{it,w}, \tag{9}$$

where the regressors p_{it} and w_{it} are allowed to depend on unit-specific mean levels $\alpha_{i,p}$ and $\alpha_{i,w}$. The fixed-effects $\alpha_{i,p}$ and $\alpha_{i,w}$ capture the dependence between μ_i and regressors ($\texttt{cov}(P_{it}, \mu_i) \neq 0$ and $\texttt{cov}(W_{it}, \mu_i) \neq 0$). The error terms in (8) and (9) then follow the GC model, capturing the regressor endogeneity of p_{it} ($\texttt{cov}(P_{it}, e_{it}) \neq 0$) and the dependence among endogenous and exogenous regressors ($\texttt{cov}(P_{it}, W_{it}) \neq 0$). Assuming a homogeneous GC model, a two-stage copula control function approach estimates the following augmented

[20] This general-location GC model can also be applied to grouped/clustered data formed by discrete exogenous regressors.

panel regression model:

$$y_{it} = \mu_i + P'_{it}\alpha_i + W'_{it}\beta_i + \sum_{k=1}^{K} \gamma_k C_{it,k} + \omega_{it}, \tag{10}$$

where $C_{it,k}$ is the copula term in Table 7; $\widetilde{P}_{it,k}$ and $\widetilde{W}_{it}$ in the copula term are the time demeaned value of $P_{it,k}$ and W_{it} (i.e., fixed-effect transformation). Thus, the procedure is to apply the 2sCOPE/2sCOPE-np in Table 6 to the time-demeaned regressors. The new error term ω_{it} is shown to be uncorrelated with all regressors and the fixed-effects μ_i in the augmented panel regression model in Equation 10, thereby eliminating the regressor-error dependence. Standard panel regression estimators assuming all regressors are exogenous can be applied to Equation 10 and yield consistent estimates. Yang, Qian, and Xie (2024b) formally demonstrate the consistency, asymptotic normality, and standard error formula for the model parameter estimates. Copula correction assuming homogeneity is found to be robust to heterogeneous endogeneity across panel units (Haschka 2022; Yang, Qian, and Xie 2024b). When the panel is sufficiently long, Yang, Qian, and Xie (2024b) explicitly permit the copula dependence to vary across panel units and recover estimates of panel-specific endogeneity. Yang, Qian, and Xie (2024b) further employ the mean group estimator to estimate the augmented panel regression model in Equation 10 with slope endogeneity ($\texttt{cov}(P_{it}, \alpha_i) \neq 0$).

Copula correction can also be applied to address regressor endogeneity in random coefficients logit (RCL) models for panel discrete choice outcomes (Park and Gupta 2012; Yang, Qian, and Xie 2024a). In RCL models, the endogeneity of price is modeled as the dependence between product price and unobserved time-varying product characteristics. One can then map an RCL model specified at the consumer level to an aggregate linear model for the product utility averaged across all consumers, for which copula correction for linear models can be directly applied to address regressor endogeneity.

Proper Construction of Nonparametric Rank-Based Copula Transformation

Applications of copula endogeneity correction mostly employ the nonparametric rank-based copula transformation based on the empirical marginal distributions of regressors (e.g., Equation 6). Although convenient and immune to misspecifications of these nuisance marginal distributions, the empirical copula transformation requires special handling of mapping from ranks to latent copula data. The empirical rank-based copula transformation follows two steps. First, the observations are ordered and mapped to a ranked percentile. The second step computes the inverse normal CDF of that ranked percentile. Web Appendix Table W9 presents a toy example of the two-step copula transformation.

During the copula transformation, the observation with the largest rank is technically the 100th percentile; however, the inverse normal CDF of the 100th percentile is undefined. To avoid generating undefined latent copula data, one can adjust the copula transformation of the maximum value of the regressor (e.g. P) for a sample size n as follows:

$$P_i^* = \Phi^{-1}(F_P(P_i)) \;=\; \begin{cases} \Phi^{-1}(\texttt{Rank}(P_i)/n) & \text{if } P_i < \max(P) \\ \Phi^{-1}(n/(n+1)) & \text{if } P_i = \max(P). \end{cases} \tag{11}$$

The above percentile adjustment for the maximum value yields a theoretically valid maximum value of the underlying copula data. A justification for this adjustment is that the expected value of the maximum of a standard normal sample of size n can be approximated by $\Phi^{-1}(\frac{n-\alpha}{n+1-2\alpha})$ with a recommended value for α as $\alpha = 0.375$ (Royston 1982). The use of $\Phi^{-1}(\frac{n}{n+1})$ can be viewed as setting $\alpha = 0$ in the formula, which is simpler to use and leads to an almost identical result as setting $\alpha = 0.375$ for a typical sample size (i.e., $n >> \alpha$).

To demonstrate the importance of the empirical copula transformation, consider an alternative empirical copula construction as used in Becker, Proksch, and Ringle (2022) to set the percentile for the last observation to a fixed value of 0.9999999:

$$P_{i,Fix}^* = \Phi^{-1}(F_P(P_i)) \;=\; \begin{cases} \Phi^{-1}(\texttt{Rank}(P_i)/n) & \text{if } P_i < \max(P) \\ \Phi^{-1}(0.9999999) = 5.1999 & \text{if } P_i = \max(P), \end{cases} \tag{12}$$

where P_{Fix}^* means a fixed percentile value is used for the largest rank. The fixed value is chosen to be 0.9999999 (close to 1) in order to maintain the same rank order after copula transformation unless sample size is extremely large (i.e, n >1,000,000). However, when sample size is small or moderate, copula transformation of the maximum can differ substantially from the theoretically predicted value; this becomes an outlier with its covariate values distant from the centroid of covariate distributions. Such an outlier has high leverage and is expected to have outsized influence on coefficient estimates in the copula augmented regression, adversely impacting the performance of copula correction.

To assess the impact of empirical copula construction on the performance of copula correction, we compare the algorithms in Equations 11 and 12 using simulation studies[21] in which the true parameter values are known. The simulation studies consider the settings of without correlated exogenous regressors as described in Becker, Proksch, and Ringle (2022) and with correlated exogenous regressors, as described in Web Appendix D.

The results in Web Appendix D[22] reveal that judicious handling of copula transformation is crucial for the performance of copula correction. Importantly, including an intercept in the model does not cause concern as long as the last-ranked value of the empirical CDF is properly handled by using the recommended copula transformation algorithm. A key finding of this study is that the substantial bias of the P&G method for models with intercept, discovered in Becker, Proksch, and Ringle (2022), is largely solved by adjusting the largest rank using Equation 11.

Altogether, our analysis provides theoretical and empirical justifications for optimal copula transformation. The new insights help demystify misinterpretations about copula correction and promote optimal copula transformation that greatly affects the effectiveness of copula correction. We recommend against assigning a fixed percentile value for the largest rank, instead favoring the algorithm in Equation 11 to produce valid empirical copula construction regardless of sample size.

[21]The R codes for simulation studies and empirical examples are available at https://osf.io/by2ge/?view_only=27cc862a9c02446abbafd3a745722603.

[22]We also provide an interactive applet interfaced supplement accessible at https://unknown8866.github.io/histogram-webpage/ for readers to visually explore the results of the simulation study.

Optimal Copula Estimation of Endogenous Moderating and Nonlinear Effects

Many applications in different fields are interested in estimating structural models with higher-order terms (e.g., interactions) of endogenous regressors to gain deeper understanding of causal mechanisms. Copula correction methods can handle these nonlinear terms (Table 5), but considerable confusion and variation exist in addressing such higher-order endogenous regressors (Web Appendix Table W3). In this section we look at the best approach to handle these higher-order terms via both theoretical proof and empirical evaluations.

Consider the following general model containing higher-order terms of regressors:

$$Y_i = \mu + \alpha_1' P_i + \alpha_2' f_1(P_i) + \alpha_3' f_2(P_i, W_i) + \beta' W_i + \eta f_3(W_i) + E_i, \quad (13)$$

where P_i is a vector of K continuous and endogenous regressors, and W_i is a vector of exogenous regressors. The structural model in Equation 13 expands the model in Equation 1 to include higher-order endogenous terms, namely $f_1(P_i)$ and $f_2(P_i, W_i)$, and higher-order exogenous terms, $f_3(W_i)$. Below are examples of these higher-order terms:

- Polynomial functions of a scalar P_i: $\alpha_2' f_1(P_i) = \alpha_2 P_i^2$
- Interaction of two endogenous regressors $P_i = (P_{1i}, P_{2i})$: $\alpha_2' f_1(P_i) = \alpha_2 P_{1i} P_{2i}$
- Interaction of endogenous and exogenous regressors: $\alpha_3'(P_i, W_i) = \alpha_3 P_i W_i$.

Because higher-order terms of endogenous regressors, $f_1(P_i)$ and $f_2(P_i, W_i)$, are also endogenous, it is tempting to control their endogeneity by adding separate copula correction terms for them. However, the point of not needing these copula correction terms for these higher-order terms is clearly shown in the following copula augmented regression, including only copula correction terms for the first-order endogenous terms (i.e., main effects):

$$Y_i = \mu + \alpha_1' P_i + \alpha_2' f_1(P_i) + \alpha_3' f_2(P_i, W_i) + \beta' W_i + \eta f_3(W_i) + \gamma' C_{i,main} + \epsilon_i, \quad (14)$$

where $C_{i,main} = (C_{i,1}, \cdots, C_{i,K})$ contains copula correction terms for main terms P_i only (Table 6). Because the new error term ϵ is independent of P and W, ϵ is also independent of $f_1(P)$, $f_2(P, W)$ and $f_3(W)$, all of which are deterministic functions of P and W. Thus, once the copula correction terms for main effects C_{main} are included as control variables in Equation 14, the new error term ϵ is already independent of (and uncorrelated with) these high-order terms, so extra correction terms for $f_1(P)$ and $f_2(P, W)$ are not needed. This simplicity of handling higher-order endogenous regressors is a merit of copula correction.[23]

Although it is unnecessary to add the copula correction terms for higher-order terms,[24] a further question is what will happen if the additional copula generated regressors for the higher-order terms are included. Will doing this lead to better or worse performance of copula correction? The issue with adding unnecessary regressors $C_{f_1(P_i)}$ and $C_{f_2(P_i,W_i)}$ is the significant collinearity between these higher-order copula terms and their co-varying constituents (P, $f_1(P)$, $f_2(P, W)$, and C_{main}). This substantially decreases precision of coefficient estimates, and makes copula correction methods perform worse than otherwise, shown formally by Theorem 1 in Web Appendix E. Consistent with the theoretical results, simulation studies (Web Appendix F) demonstrate substantial harmful effects if copula terms for higher-order regressors are added to control for their endogeneity. These effects include large finite sample bias and inflated variability of structural model parameter estimates.

Obtaining Standard Errors

For methods performing joint estimation in one step (Qian and Xie 2024), standard errors by inverting the Hessian matrix can be straightforwardly obtained as a byproduct of the estimation process. For two-step copula methods, bootstrapping is applied to obtain proper standard errors in order to account for additional uncertainty associated with obtaining generated regressors in the first step. Starting with the original dataset consisting of n independent observations, bootstrapping resamples the data and randomly draws n observations from the original dataset with replacement, and then calculates the copula corrected estimates on the bootstrap sample. This simulation process is repeated many times to obtain a distribution for each parameter estimate. The standard deviation of this bootstrap distri-

[23]IV control function approach also shares this merit (Petrin and Train 2010), but not the 2SLS (Wooldridge 2010 Chapter 6.2); this may sow the confusion, as the 2SLS approach advocates for including instruments of the endogenous higher-order terms.

[24]Papies, Ebbes, and Van Heerde (2017) (p. 615) noted this point for the P&G method. Our analysis (1) extends this result to more general methods and (2) demonstrates a stronger result that adding the unnecessary high-order copula terms is suboptimal and has significant adverse effects using both theoretical proof and empirical evaluation.

bution then estimates the standard error of the estimate. For panel data, cluster bootstrap should be used to resample independent cross-sectional units instead of individual observations (Haschka 2022). That is, only the cross-sectional units (clusters) are resampled, while all the observations within the sampled clusters are retained and unchanged. This ensures the bootstrap samples retain dependence structures among panel observations existing in the original data. Simulation studies have shown the bootstrap produces reliable standard error estimates with single or multiple endogenous regressors, with or without correlated exogenous regressors (Park and Gupta 2012; Haschka 2022; Yang, Qian, and Xie 2024a).

GUIDANCE FOR PRACTICAL USE

Based on recent findings and advances, this section describes a procedure guiding practical usage of copula correction methods. Figure 5 presents a step-by-step flowchart[25] for the steps and checkpoints in using copula correction. Before entering the flowchart, one should ensure the model is appropriately specified and theoretically supported, with pertinent control variables included in W and the regressor matrix being full rank. To ensure exogeneity of W, include only necessary exogenous control variables. Control variables believed to be endogenous should be treated as endogenous regressors or removed from the model.

When the need to use copula correction is affirmed using Table 2 at the start of the flowchart, assess the plausibility of the underlying assumptions in the focal application (Item a in Table 4). The double robustness property of copula correction using control functions means that copula correction can be used with departures from GC regressor-error dependence, as long as GC adequately captures unexplained dependence between endogenous regressors and U (the combined effects of all unobserved confounders) given exogenous regressors. Copula correction also works with a nonnormal error distribution. However, out of an abundance of caution and for optimal robustness, consider revising model specifications (e.g., transform variables or add more control variables) if the error distribution is suspected to be highly skewed. If copula correction is chosen, follow the rest of the flowchart to

[25] A web selector tool is available at `https://unknown8866.github.io/flowchart-webpage/`

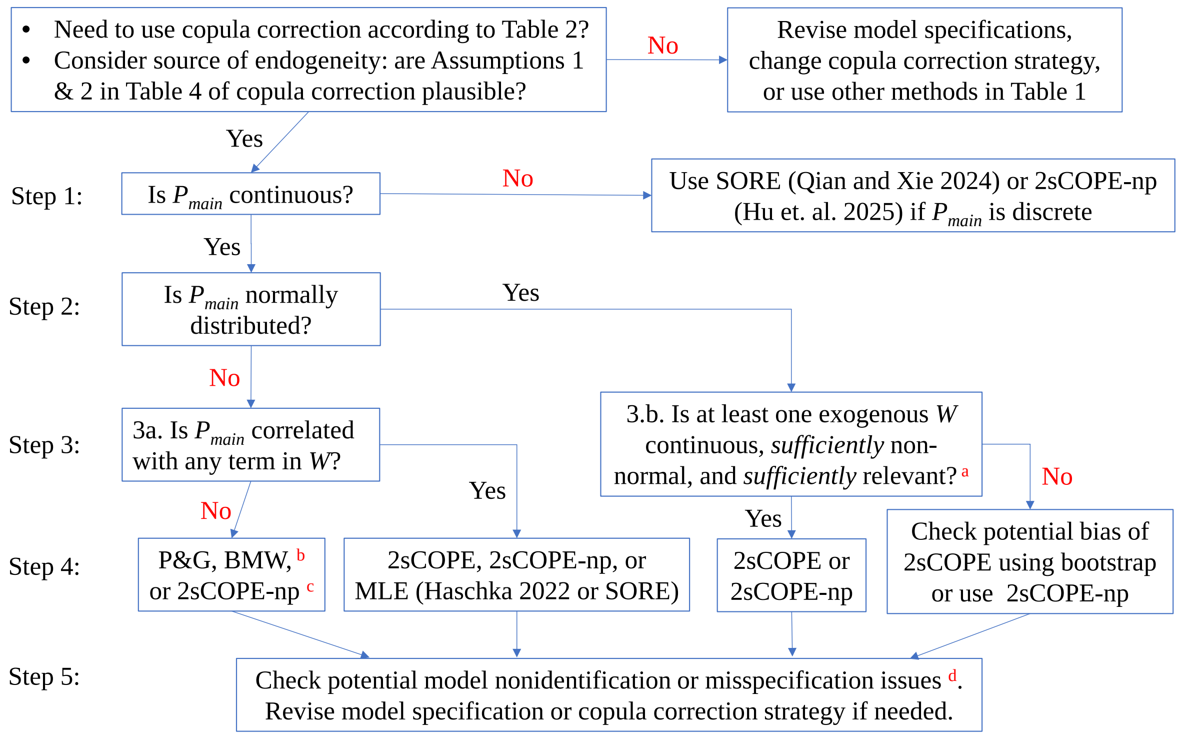

Figure 5: Flowchart for Copula Procedure.

Note: P_{main} denotes the first-order terms of endogenous regressors. W denotes exogenous control variables and $\texttt{Cov}(W, E) = 0$.

For panel data, use the MLE method of Haschka (2022) or the copula control function method (Table 7) that checks Steps 2-3 and obtains the copula correction terms using time-demeaned regressors.

a: W is sufficiently nonnormal if normality test $p < .001$ & sufficiently relevant to P_{main} if F statistics >10.

b: The BMW method (Breitung, Mayer, and Wied 2024) was suggested as a robust check of the P&G method (Park and Gupta 2024). This method and Mayer and Wied (2025) rely on some implicit assumptions (e.g., W affects only the mean of P) and do not permit GC regressor-error dependence in general (so can yield biased results when regressor-error follows GC dependence). Thus, using this method as a robustness check of the GC regressor-error dependence can yield ambiguous results. By contrast, 2sCOPE/2sCOPE-np permit GC and non-GC regressor-error dependence. In particular, the nonparametric 2sCOPE-np nests the BMW method as a special case and is better suited for robustness checking purposes.

c: A sample size of > 300 is recommended for 2sCOPE-np.

d: Check the inflation of standard errors of copula corrected estimates relative to those of uncorrected estimates. An inflation of > 6 times flags potential model identification and misspecification issues. Consider a robustness check using the methods of Liengaard et al. (2024) and Yang, Qian, and Xie (2024b) when sample size is sufficient and boundary conditions are met within each level of combinations of discrete exogenous regressors.

determine appropriate copula correction methods. As shown previously, copula correction only needs to include CCFs corresponding to the first-order terms of endogenous regressors, P_{main}, even when the structural model contains higher-order terms of endogenous regressors. Thus, the flowchart only needs to consider P_{main}. Furthermore, when the structural model includes an intercept, the copula transformation should use the algorithm in Equation 11 to avoid the estimation bias discovered in Becker, Proksch, and Ringle (2022). When conditions are met, the P&G method can be followed, but more recent research relaxes these conditions and presents the path to perform copula correction when these conditions are not met.

Step 1. This step checks whether the endogenous regressor P_{main} has sufficient support. If P_{main} is noncontinuous (binary, discrete with only a few levels, or semicontinuous), use likelihood-based SORE (Qian and Xie 2024) or moment-based 2sCOPE-np (Hu, Qian, and Xie 2025). Otherwise, continue to Step 2 below.

Step 2. This step checks whether P_{main} is normally distributed or not. If P_{main} is normally distributed, the P&G method cannot be used because the model is unidentified. However, a normally distributed P_{main} can still be a candidate for copula correction through 2sCOPE. Yet, this route follows a different path, as seen in Figure 5 and discussed more below in Step 3.b. The literature notes that more powerful tests for normality, such as the Shapiro-Wilk test or Anderson-Darling test, might not fully rule out nonidentification, because these tests can detect small departures from normality that are insufficient for copula correction (Becker, Proksch, and Ringle 2022; Eckert and Hohberger 2023). Yet, the Kolmogorov-Smirnov (KS) test is relatively conservative among the most commonly used normality tests; a p-value less than 0.05 from the KS normality test has been shown to perform well for ruling out finite sample bias due to insufficient regressor nonnormality (Yang, Qian, and Xie 2024a).

Step 3. This step marks one of the biggest shifts in copula usage since P&G, consisting of two disjoint steps (3.a and 3.b), depending on the outcome of Step 2. The data requirements in this step are established using comprehensive factorial design simulation experiments to assure satisfactory performance of copula correction, across a wide range of conditions in finite samples (Web Appendix E.8 in Yang, Qian, and Xie 2024a).

3.a. If the endogenous regressor P_{main} has sufficient nonnormality (KS p-value < 0.05) in Step 2 above, Step 3 will check an additional condition of no correlation between P_{main} and all exogenous regressors (using Fisher's Z test for correlation) to determine if the P&G method can be used.[26] When this condition is met and sample size is small, the P&G

[26]A less stringent condition for using P&G with K endogenous regressors is no correlation between $\sum_{k=1}^{K} P^*_{main,k}\gamma_k$ with all exogenous regressors (Yang, Qian, and Xie 2024a).

method may be preferred because a simpler and valid model is more efficient than a more general method. As sample size increases, 2sCOPE has negligible efficiency loss relative to P&G and is the preferred method. In this case, one should use 2sCOPE/2sCOPE-np to handle correlated exogenous regressors. Alternatively, an MLE copula procedure (either the one-step SORE or the two-step procedure of Haschka 2022) can be used.

3.b. If the endogenous regressor P_{main} is found to have insufficient nonnormality (KS p-value > 0.05) in Step 2, then one cannot use the P&G method, but can use 2sCOPE or 2sCOPE-np to leverage correlated exogenous regressors to achieve model identification. In order to compensate for the lack of nonnormality of endogenous regressor P in 2sCOPE, at least one exogenous and continuous regressor W needs to satisfy the following two conditions: (1) sufficient nonnormality, and (2) sufficient association with the endogenous regressor P. A conservative rule of thumb for such a W is the p-value from the KS test on W being < 0.001 and a strong association with P (F statistic for the effect of W^* on $P^*_{main} > 10$ in the first-stage regression). When these conditions are met, 2sCOPE is expected to yield estimates with negligible bias even if P_{main} is normally distributed. When these conditions are not met, Yang, Qian, and Xie (2024a) suggest gauging potential bias of 2sCOPE for data at hand via a bootstrap procedure described there, and using 2sCOPE only if the potential bias is small. Alternatively, the more general 2sCOPE-np can be used to leverage exogenous regressors for model identification (see Step 4 below for choosing between 2sCOPE and 2sCOPE-np).

As seen above, only one of 3.a or 3.b is taken in Step 3. Importantly, if P already has sufficient nonnormality that leads to 3.a, there is no need to do 3.b to check if any continuous W has sufficient nonnormality and is associated with P. These conditions are only checked to find a useful W to compensate for the lack of nonnormality of P. In 3.b, 2sCOPE/2sCOPE-np uses W to tease out an exogenous part of the endogenous regressor for model identification. A good starting place to find such W is in the exogenous control variables pre-existing in the OLS or IV regressions. Unlike IVs, these control variables (e.g., exogenous demand shocks) do not need to satisfy the stringent exclusion restriction condition. That is, these Ws do not have to be excluded from the structural model (e.g., Equation 1), and can affect the outcome directly and not through the endogenous regressors. Such Ws are more readily available than IVs, and because empirical association between the candidate W and P is sufficient, researchers using copula correction do not need to argue for the causal pathways between W and P like in the case of IVs.

Step 4. The step applies the appropriate copula procedures using either control functions or likelihood-based joint estimation. To choose between likelihood-based methods and moment-based control function methods when both can be used, see the previous section "Control Function vs Likelihood-based Correction Methods". Among copula control function methods, the 2sCOPE is relatively easy to apply and reasonably robust (e.g., assumes no particular error distribution and no specific copula structure for regressor-error dependence), which makes it well suited for a first-line method. The more general 2sCOPE-np is first-stage model-free and allows non-GC regressor-confounder dependence unconditionally on W, making it preferred when more robustness is desired or departures from GC regressor-confounder dependence are suspected. The nonparametric kernel control function estimation employed in 2sCOPE-np does require larger sample sizes (Boundary Condition c in Figure 5) and greater computation power, compared with the computationally simpler 2sCOPE method. For control function methods, if the generated regressor is not statistically significant, this suggests the endogenous regressor P_{main} is not sufficiently correlated with the error term, and endogeneity is unlikely. Thus, non-significant generated regressors should be dropped and the model re-estimated. Marketing studies have dropped copula correction terms at the $p > 0.10$ level (e.g., Datta et al. 2022), suggesting even marginally significant copula correction terms are still worth retaining. If no generated regressor is significant, the model can be estimated in a more traditional manner (i.e., OLS).

Step 5. The final step is to check the inflation of standard errors of copula corrected estimates relative to those of uncorrected estimates using the ICON statistics. An inflation of > 6 times flags potential model misspecification issues or lack of model identification.

COPULA IMPLEMENTATION EXAMPLES

This section illustrates use of the flowchart to guide copula implementation via two examples using weekly store sales data from the IRI Academic data set (Bronnenberg, Kruger, and Mela 2008). To correct for price endogeneity, the first example examines the main effect of price, while the second example examines higher-order moderating effects captured by the interaction between price and store feature (i.e., weekly store flyer promoting products).

Example 1: Main Effects Application of Copula Correction

Returning to our running Example 1, the outcome of interest is the weekly sale volume in the diaper category for one focal store in the Buffalo, NY market in the years 2002-2006, where volume is measured in diaper counts. Price is defined on an equitable volume across UPCs, since pack sizes vary in diapers per pack. IRI additionally collected information on whether UPCs were featured in the store's weekly flyer that week. Category price and feature are evaluated as market-share weighted averages of UPC-level price and feature, respectively.

Knowledge of category price elasticity is critical for retailers or category managers to set optimal pricing and increase category demand that is the first source of profitable growth, and for policymakers to design interventions (e.g., gasoline tax). Price is commonly considered endogenous in category demand models (Nijs et al. 2001; Park and Gupta 2012; Li, Linn, and Muehlegger 2014). In this example, price was treated as endogenous because of unobserved variables (e.g., retailer pricing decisions, number of shelf facings) that, when omitted from a model, become part of the structural error. For brevity, we use "Price" and "Volume" hereafter to refer to the log-transformed category price and sales volume, respectively. The impacts of price and feature advertising appear in the following model:

$$\text{Volume}_t = \mu + \alpha P_t + \beta' W_t + E_t. \tag{15}$$

In the model, P_t is the endogenous regressor as log-transformed price. W_t is a vector of control variables including feature, week, and binary variables for quarters 2, 3, and 4. We treat feature as exogenous because decisions to promote items in the store flyer are made well in advance of implementation, and are likely uncorrelated with weekly unobservables (Chintagunta 2002; Sriram, Balachander, and Kalwani 2007). The week variable is included as a control variable to account for a small but significant trend in price increases over time.

One solution to price endogeneity is to use IVs, where the diaper price of another store in the same market was used as an IV. Prices are correlated for both stores, with the belief that wholesale prices are similar for products sold by the two stores (relevance), but uncaptured product characteristics (including retailer decisions like shelf facings and shelf location) are unlikely related to wholesale prices (ER). However, the ER assumption is untestable and the IV may be not strong enough. This is one of the use cases for copula correction as listed in Table 2: use multiple methods (both IV estimation and copula correction here) to cross-validate results and increase robustness of causal inference. The omitted variables consist of multiple unmeasured product attributes (like shelf facings and locations) whose joint effect (U) can be expected to follow a normal distribution.[27] As the model includes feature and quarters to control for pre-planned promotion activities affecting sales and pricing, it appears plausible to use GC to capture the remaining dependence between price and the omitted effect U (temporaneous retailer decisions) unexplained by these control variables. We will inspect standard errors from copula correction to confirm empirical identification and check for signs of model nonidentification or regressor-error dependence misspecifications. Before we present the results, below we walk through the steps of the Figure 5 flowchart.

Step 1. Is P_{main} continuous? The endogenous regressor, Price, is a continuous measure, ranging from \$0.140 to \$0.262 per diaper, with a mean of \$0.221, median of \$0.224, and standard deviation of \$0.018.

Step 2. Is P_{main} normally distributed? Figure 6 shows somewhat skewness to the left for the price variable. However, the skewness is not strong enough to reject the KS test for normality ($D = 0.08$, $p > 0.05$) at the 0.05 level of significance. This means that the

[27] Because copula control function (CCF) is robust to symmetric nonnormal distributions of U, we checked residuals for signs of asymmetrically distributed U. The residuals from using OLS and CCF have skewness of -0.088 and -0.083, respectively, indicating no residual skewness. This provides greater assurance of copula validity, although we should note that skewed residuals do not necessarily contradict the use of CCF since CCF permits nonnormal error term.

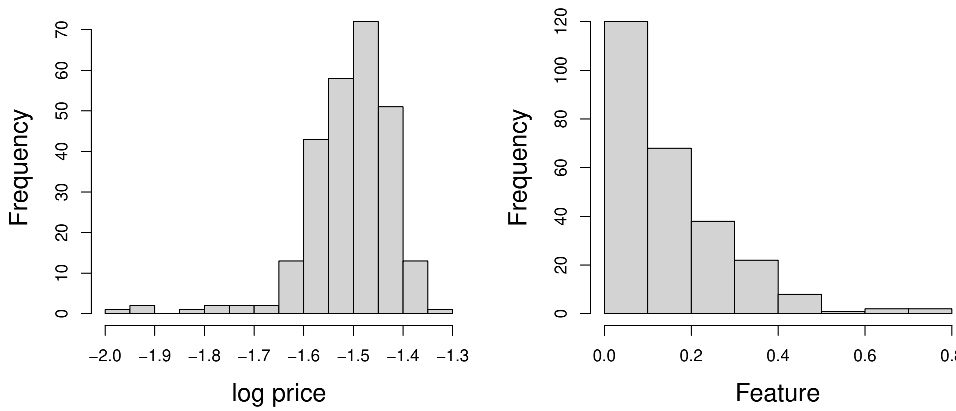

Figure 6: Distributions of Price and Feature in Example 1.

endogenous regressor may not have sufficient nonnormality. One solution is to leverage related exogenous regressors with sufficient nonnormality via 2sCOPE as described next.

Step 3.b. Is at least one W sufficiently nonnormal and correlated with P_{main}? The first-stage regression shows only one exogenous regressor is sufficiently correlated with price (F-stat > 10): feature ($F = 16.8$). The regressor, feature, is highly skewed (Figure 6) and nonnormally distributed based on the KS test ($D = 0.14$, $p < 0.0001$).

Step 4. Perform 2sCOPE estimation. The above steps show that conditions have been verified such that 2sCOPE or 2sCOPE-np can be used to handle the price endogeneity. However, the sample size (n=261) does not reach 300 to use 2sCOPE-np, so we instead apply the 2sCOPE method.[28] The standard errors are obtained using 500 bootstrap samples.

Step 5. Check the inflation of standard errors using the ICON statistics. All ICON statistics are far less than 6 (Table 8), showing no signs of weak identification.

Table 8 compares 2sCOPE to OLS and 2SLS using the IV. The 2sCOPE estimation results show that the copula correction term C_{price} (i.e., the first-stage residual) is significant (Est. $= 0.077$, SD $= 0.037$, $p < 0.05$), indicating the presence of price endogeneity, so we retain the CCF in the model to control for price endogeneity.

The results show that while price has the smallest absolute effect in the OLS model (Est. = -1.367, SE = 0.137, $p < 0.01$), the effect is greatest in the 2SLS model (Est. = -2.470, SE = 0.661, $p < 0.01$); the 2sCOPE price estimate falls in between and is much closer to the

[28] 2sCOPE-np yields similar estimation results, which are reported in Web Appendix Table W16.

Table 8: Estimation Results for Example 1

Parameters	OLS		2SLS		2sCOPE		
	Est (SE)	P	Est (SE)	P	Est (SE)	P	ICON
Intercept	6.005 (0.205)	0.000	4.371 (0.978)	0.000	4.763 (0.668)	0.000	2.9
Price	-1.367 (0.137)	0.000	-2.470 (0.661)	0.000	-2.205 (0.446)	0.000	3.3
Feature	0.298 (0.095)	0.002	0.059 (0.178)	0.738	0.124 (0.124)	0.317	1.3
Week	-0.002 (0.000)	0.000	-0.002 (0.000)	0.000	-0.002 (0.000)	0.000	1.1
Q_2	-0.019 (0.031)	0.550	-0.014 (0.035)	0.693	-0.018 (0.036)	0.617	1.2
Q_3	-0.018 (0.032)	0.567	-0.034 (0.036)	0.349	-0.029 (0.035)	0.407	1.1
Q_4	-0.018 (0.032)	0.576	-0.061 (0.041)	0.140	-0.044 (0.035)	0.209	1.1
C_{price}					0.077 (0.037)	0.037	—
ρ					0.366 (0.160)	0.022	—

Note: Table presents estimates, bootstrapped standard errors in the parentheses, and the p-values. ICON is the ratio of standard errors of 2sCOPE estimates to those of the OLS estimates.

2SLS price estimate (Est. = -2.205, SE = 0.446, $p < 0.01$). Compared to 2SLS using IV, the 2sCOPE results are not unlike that of 2SLS, within one SD of the 2SLS price estimates. The 2SLS price estimate differs somewhat from the 2sCOPE price estimate by 12.0%. Although the correlation in prices between the two stores is significant and passes the weak instruments test (F = 13.89, $p < 0.01$), the correlation is not especially strong (r = 0.218). Thus, the difference between 2sCOPE and 2SLS seen here could be because the other store's price as an IV is not particularly strong, and a strong IV is not always readily available. In such cases, cross-validating results from different methods (IV correction and IV-free copula correction) can increase the robustness of causal estimation. The 2sCOPE shows that price is positively correlated with the error term (Est. = 0.366, SE = 0.160, $p < 0.05$), indicating the presence of price endogeneity. This finding is consistent with the result of the Wu-Hausman test (H = 3.56, $p < 0.07$) from 2SLS, which also suggests endogeneity was likely present. Overall, the comparison with 2sCOPE shows that without endogeneity correction, managers would severely under-estimate price elasticity based on the OLS findings for this store, by 38.0%.

Example 2: Copula Estimation of Endogenous Interactions

Example 2 illustrates how copula correction is applied with endogenous interaction terms and examines the adverse effects (estimation bias and inflated estimation variability) of including higher-order copula terms. This empirical application extends the sales response model in Equation 15 to include an interaction term between price and feature. See Web Appendix G.2 for detailed analysis and results of Example 2.

Managerial and Academic Implications

The two examples highlight how copulas can correct for endogeneity to remove bias in estimation, as well as how copulas should be correctly specified in models with interactions. Example 1 showed that without the copula, the OLS estimate for price elasticity was severely under-estimated (Est. = -1.367) compared to both 2SLS (Est. = -2.470) and 2sCOPE (Est. = -2.205). The result showed price elasticity in OLS was 38% lower in size than 2sCOPE. We also noted that the instrument was significant but not particularly strong, attributing to the difference between 2SLS and 2sCOPE estimates.

Controlling for endogeneity in price elasticity estimates can have important managerial implications. Price elasticity estimates are often a crucial piece of information for managers to set the optimal pricing that maximizes profit. Let the profit function $p(Price) = V * (Price - Cost)$, where V is sales volume and $cost$ is the marginal cost. The maximum profit is then the value of $Price$ that satisfies the condition $\frac{\partial \ln p(Price)}{\partial Price} = 0$. Following the Amoroso-Robinson relation, the profit-maximizing price is $Price_{optim} = \frac{\alpha}{1+\alpha} Cost$, where α is the price elasticity. In Example 1, we find the optimal pricing is $Price_{ols} = \frac{-1.367}{-1.367+1} Cost = 3.72 * Cost$ if the OLS price elasticity estimate is used, and $Price_{cop} = \frac{-2.205}{-2.205+1} Cost = 1.83 * Cost$ if the 2sCOPE estimate is used. Because of the price endogeneity associated with the scanner panel data, the biased OLS estimate underestimates the size of price elasticity, meaning that OLS considers consumers less price sensitive than they actually are. Thus, the manager will set the price more aggressively; in Example 1, using the OLS price elasticity estimate means the manager will set price at 103% higher than the actual optimal price.

This considerable difference in optimal pricing based on the OLS and 2sCOPE price elasticity estimates results in a substantial profit difference as well. It can be shown that the profits achieved at the different prices have the following relationship: $\ln \frac{p_{cop}}{p_{ols}} = \alpha \ln[Price_{cop}/Price_{ols}] +$

$\ln[(Price_{cop} - Cost)/(Price_{ols} - Cost)]$, where p_{cop} and p_{ols} refer to the profit achieved when using the 2sCOPE and OLS price elasticity estimates, respectively. For Example 1, $\frac{p_{cop}}{p_{ols}} = 1.46$, which corresponds to a loss of 31% in profit when using the incorrect OLS price elasticity estimate, compared to using the correct 2sCOPE estimate (Figure 3).

CONCLUSION

Endogeneity correction is a key concern for academics and practitioners, and the instrument-free copula correction has been increasingly used to address endogeneity bias. Copula correction has practical advantages and feasible implementation. Yet, like other causal estimation procedures designed for use with nonexperimental data, the validity of copula correction requires correct implementation of the method, needing boundary conditions and data requirements to be met in its empirical applications.

This study contributes to the field in three areas. One, we advance the discussion regarding the theoretical rationales of copula correction and provide a review for how copula correction has been used in marketing and other fields to correct for endogeneity, across substantive areas, and how it has been applied (and misapplied). Two, we elucidate the identification assumptions and data requirements of copula correction and build on recent advances to provide an updated best practices "cookbook" for both managers and academics to follow in applying and implementing the copula procedures (Tables 1 and 3; Figure 5). The cookbook also informs how to modify analysis when certain conditions are not met. Three, we evaluate implementation variations (such as optimal copula transformations and higher-order effects of moderation) and demystify misconceptions of copula correction, showing theoretically and with real-world data best practices for copula correction usage.

We demonstrate that existing variations in the implementation of copula correction have substantial impacts on its performance. Our discussions on the methodological aspects of the copula method informs optimal and theoretically sound implementation for copula correction. We present a theoretically grounded way of constructing copula transformation that avoids the potential finite sample bias problem and substantially improves the performance of copula correction. We show that excluding the copula terms for higher order endogenous regressors (i.e., interactions) is optimal and considerably outperforms when these copula terms are included. To our knowledge, these are the first theoretical results justifying the optimal implementation of these aspects affecting the performance of copula correction.

We also discuss the latest extensions that expand the applicability, flexibility and robustness of copula correction, highlighting endogeneity correction when the conditions and data requirements of earlier copula correction approaches are not met (Table 3); for cases where the endogenous regressors have insufficient nonnormality and correlate with exogenous regressors (and the traditional P&G method fails to work), we describe how a two-stage copula correction (2sCOPE) and its extensions, as well as other copula correction procedures, can still work by leveraging relevant exogenous regressors.

We synthesize the above discussions into a flowchart with easy-to-follow checkpoints and data requirements. This guide is practical for researchers — in both academia and industry — to employ copula correction methods. In addition to making the copula code available, we illustrate its usage in two empirical examples for two different product categories.

Future avenues of research are teeming, such as extending the flexible 2sCOPE framework for more generality (e.g., Hu, Qian, and Xie 2025), adapting copula correction to Bayesian inference (e.g., Haschka 2024), exploring methods to further reduce the dependence on the GC assumption (e.g., Qian and Xie 2024), improving computational efficiency especially for computationally intensive procedures (e.g., the MLE procedures), to name a few. While copula correction has made advances, and a great variety of quantitative models have been using copulas, new models are regularly emerging. As such, new opportunities to adapt copula correction to new types of data or models abound.

REFERENCES

Albers, Sönke, Murali K Mantrala, and Shrihari Sridhar (2010), "Personal selling elasticities: a meta-analysis," *Journal of Marketing Research*, 47 (5), 840–853.

Ananthakrishnan, Uttara M, Cedric Xu, Maria Mitkina, and Yong Tan (2025), "Political Consumerism and the Emergence of Rare Information on User-Generated Content Platforms," *Management Science, Forthcoming, Available at SSRN: https://ssrn.com/abstract=5237716.*

Atefi, Yashar, Michael Ahearne, James G Maxham III, Todd D Donavan, and Brad D Carlson (2018), "Does Selective Sales Force Training Work?," *Journal of Marketing Research*, 55 (5), 722–737.

Becerra, Manuel and Garen Markarian (2021), "Why are firms with lower performance more volatile and unpredictable? A vulnerability explanation of the Bowman paradox," *Organization Science*, 32 (5), 1327–1345.

Becker, Jan-Michael, Dorian Proksch, and Christian M Ringle (2022), "Revisiting Gaussian Copulas to Handle Endogenous Regressors," *Journal of the Academy of Marketing Science*, 50(1), 46–66.

Bijmolt, Tammo HA, Harald J Van Heerde, and Rik GM Pieters (2005), "New empirical generalizations on the determinants of price elasticity," *Journal of marketing research*, 42 (2), 141–156.

Billingsley, Partick (1995), "Probability and Measure.," *Wiley, New York.*

Blattberg, Robert C. and Scott A. Neslin (1990), *Sales Promotion - Concepts, Methods, and Strategies* Englewood Cliffs, NJ: Prentice-Hall.

Breitung, Jörg, Alexander Mayer, and Dominik Wied (2024), "Asymptotic properties of endogeneity corrections using nonlinear transformations," *The Econometrics Journal*, 27(3), 362–383.

Bronnenberg, Bart J., Michael W. Kruger, and Carl F. Mela (2008), "Database paper - The IRI marketing data set," *Marketing Science*, 27(4), 745–748.

Chen, Hua Yun (2007), "A semiparametric odds ratio model for measuring association," *Biometrics*, 63 (2), 413–421.

Chintagunta, Pradeep K (2002), "Investigating category pricing behavior at a retail chain," *Journal of Marketing Research*, 39 (2), 141–154.

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. and Michael Smith (2011), "Modeling Multivariate Distributions Using Copulas: Applications in Marketing," *Marketing Science*, 30(1), 4–21.

Datta, Hannes, Harald J van Heerde, Marnik G Dekimpe, and Jan-Benedict EM Steenkamp (2022), "Cross-national differences in market response: line-length, price, and distribution elasticities in 14 Indo-Pacific Rim economies," *Journal of Marketing Research*, 59 (2), 251–270.

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 (2023), "Addressing Endogeneity Without Instrumental Variables: An Evaluation of the Gaussian Copula Approach for Management Research," *Journal of Management*, 49(4), 1460–1495.

Feit, Elea McDonnell and Eric T Bradlow "Fusion modeling," "Handbook of Market Research," pages 147–180, Springer (2021).

Germann, Frank, Peter Ebbes, and Rajdeep Grewal (2015), "The Chief Marketing Officer Matters!," *Journal of Marketing*, 79 (3), 1–22.

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*, 59(4), 860–881.

Haschka, Rouven E (2024), "Bayesian Inference for Joint Estimation Models Using Copulas to Handle Endogenous Regressors," *Available at http://dx.doi.org/10.2139/ssrn.4235194*.

Hu, Xixi, Yi Qian, and Hui Xie "Correcting Endogeneity via Instrument-Free Two-Stage Nonparametric Copula Control Functions," Technical report, National Bureau of Economic Research Working Paper W33607. https://www.nber.org/papers/w33607 (2025).

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*, 65, 1201–1214.

Li, Shanjun, Joshua Linn, and Erich Muehlegger (2014), "Gasoline taxes and consumer behavior," *American Economic Journal: Economic Policy*, 6 (4), 302–342.

Liengaard, Benjamin D, Jan-Michael Becker, Mikkel Bennedsen, Phillip Heiler, Luke N Taylor, and Christian M Ringle (2024), "Dealing with regression models' endogeneity by means of an adjusted estimator for the Gaussian copula approach," *Journal of the Academy of Marketing Science*, pages 1–21.

Mayer, Alexander and Dominik Wied "Endogeneity Corrections in Binary Outcome Models with Nonlinear Transformations: Identification and Inference," Technical report, arXiv:2408.06977v4. https://doi.org/10.48550/arXiv.2408.06977 (2025).

Nijs, Vincent R, Marnik G Dekimpe, Jan-Benedict EM Steenkamps, and Dominique M Hanssens (2001), "The category-demand effects of price promotions," *Marketing science*, 20 (1), 1–22.

Papies, Dominik, Peter Ebbes, and Elea McDonnell Feit *Endogeneity and Causal Inference in Marketing*, pages 253–300, World Scientific (2023).

Papies, Dominik, Peter Ebbes, and Harald J. Van Heerde *Addressing Endogeneity in Marketing Models*, pages 581–627, Cham: Springer (2017).

Park, Sungho and Sachin Gupta (2012), "Handling Endogenous Regressors by Joint Estimation Using Copulas," *Marketing Science*, 31, 567–586.

Park, Sungho and Sachin Gupta (2024), "A Review of Copula Correction Methods to Address Regressor–Error Correlation," *Impact at JMR*.

Petrin, A and K Train (2010), "A control function approach to endogeneity in consumer choice models," *Journal of Marketing Research*, 47(1), 3–13.

Qian, Yi and H Xie (2011), "No customer left behind: A distribution-free Bayesian approach to accounting for missing Xs in marketing models," *Marketing Science*, 30(4), 717–736.

Qian, Yi and H Xie (2014), "Which brand purchasers are lost to counterfeiters? An application of new data fusion approaches," *Marketing Science*, 33(3), 437–448.

Qian, Yi and H Xie (2015), "Driving more effective data-driven innovations: Enhancing the utility of secure catabases," *Management Science*, 61(3), 520–541.

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 and Hui Xie (2024), "Correcting Regressor-Endogeneity Bias via Instrument-Free Joint Estimation Using Semiparametric Odds Ratio Models," *Journal of Marketing Research*, 61(5), 914–936.

Rigobon, Roberto (2003), "Identification Through Heteroskedasticity," *Review of Economics and Statistics*, 85, 777–792.

Royston, J. Patrick (1982), "Algorithm AS 177: Expected normal order statistics (exact and approximate)," *Journal of the Royal Statistical Society. Series C (Applied statistics)*, 31 (2), 161–165.

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.

Sethuraman, Raj, Gerard J Tellis, and Richard A Briesch (2011), "How well does advertising work? Generalizations from meta-analysis of brand advertising elasticities," *Journal of Marketing Research*, 48 (3), 457–471.

Sriram, Srinivasaraghavan, Subramanian Balachander, and Manohar U Kalwani (2007), "Monitoring the dynamics of brand equity using store-level data," *Journal of Marketing*, 71 (2), 61–78.

Tran, Kien C. and Mike G. Tsionas (2021), "Efficient Semiparametric Copula Estimation of Regression Models with Endogeneity," *Econometric Reviews*, 41(5), 1–28.

Villas-Boas, J. Miguel and Russell S. Winer (1999), "Endogeneity in Brand Choice Models," *Management Science*, 45(10), 1324–1338.

Wooldridge, Jeffrey M (2010), *Econometric Analysis of Cross Section and Panel Data* Cambridge, MA: MIT Press.

Yang, Fan, Yi Qian, and Hui Xie (2024a), "EXPRESS: Addressing Endogeneity Using a Two-stage Copula Generated Regressor Approach," *Journal of Marketing Research*, 0 (ja), https://doi.org/10.1177/00222437241296453.

Yang, Liying, Yi Qian, and Hui Xie (2024b), "Handling Endogenous Marketing Mix Regressors in Correlated Heterogeneous Panels with Copula Augmented Mean Group Estimation," *NBER Working Paper w33265, https://www.nber.org/papers/w33265.*

Yang, Sha, Yuxin Chen, and Greg Allenby (2003), "Bayesian Analysis of Simultaneous Demand and Supply," *Quantitative Marketing and Economics*, 1, 251–275.

A Practical Guide to Endogeneity Correction Using Copulas

WEB APPENDIX

TABLE OF CONTENTS

WEB APPENDIX A: SUBSTANTIVE AREAS IN MARKETING WITH APPLICATIONS OF COPULA CORRECTION

See Table W1 next page.

Table W1: Examples of Substantive Areas in Marketing with Applications of Copula Endogeneity Correction

Study	Product	Price	Place	Promotion	SF[a] & CRM	Other[a]
Burmester et al 2015				x		
Datta, Foubert, and van Heerde 2015				x		
Mathys, Burmester, and Clement 2016	x			x		
Datta, Ailawadi, and van Heerde 2017		x	x	x		
Lenz, Wetzel, and Hammerschmidt 2017						x
Atefi et al 2018					x	
Gielens et al 2018	x			x		
Gijsbrechts, Campo, and Vroegrijk 2018						x
Guitart, Gonzalez, and Stremersch 2018		x		x		
Lamey et al 2018		x		x		
Lim, Tuli, and Dekimpe 2018		x				
Ter Braak and Deleersnyder 2018	x	x				x
Wetzel et al 2018					x	
Carson and Ghosh 2019					x	
Keller, Deleersnyder, and Gedenk 2019		x				
Nath et al 2019						x
Schulz, Shehu, and Clement 2019						x
Vieira et al 2019				x		x
Zhao et al 2020	x					
Bombaij and Dekimpe 2020						x
Bornemann, Hattula, and Hattula 2020	x					
Campo et al 2021	x	x				
Guitart, Hervet, and Gelper 2020				x		
Heitmann et al 2020	x	x		x		x
Homburg, Vomberg, and Muehlhaeuser 2020			x			x
Magnotta, Murtha, and Challagalla 2020					x	
Shehu, Papies, and Neslin 2020		x				
Vomberg, Homburg, and Gwinner 2020					x	
Maier and Wieringa 2021						x
Aydinli et al 2021		x				x
De Jong, Zacharias, and Nijssen 2021						x
Garrido-Morgado et al 2021	x	x				
Guitart and Stremersch 2021		x		x		x
Liu et al 2021		x				
Van Ewijk et al 2021		x		x		
Bachmann, Meierer, and Näf 2021					x	
Cron et al 2021					x	
Dhaoui and Webster 2021						x
Fossen and Bleier 2021						x

Hoskins et al 2021						x
Kidwell et al 2021						x
Lamey, Breugelmans, and ter Braak 2021						x
Sawant, Hada, and Blanchard 2021						x
Bhattacharaya, Morgan, and Rego 2022						x
Borah et al 2022	x			x		x
Cao 2022	x					x
Danaher 2022		x				
Datta et al 2022	x	x	x			
Janani et al 2022					x	
Krämer et al 2022					x	x
Ludwig et al 2022					x	
Maesen et al 2022		x	x			
Moon, Tuli, and Mukherjee 2022						
Nahm et al 2022		x				
Rajavi, Kushwaha, and Steenkamp 2022	x	x	x	x		
Scholdra et al 2022	x	x	x	x		
Van Ewijk, Gijsbrechts, and Steenkamp 2022a	x	x	x	x		
Van Ewijk, Gijsbrechts, and Steenkamp 2022b	x	x	x	x		
Widdecke et al 2022		x		x		
Zhang et al 2022		x				
Wiseman et al 2022					x	
Xu et al 2022					x	
Wiegand, Peers, and Bleier 2022				x		x
Cao et al 2023						x
Gielens et al 2023	x	x				
Umashankar, Kim, and Reutterer 2023						x
Burchett, Murtha, and Kohli 2023					x	
Dall-Olio and Vakratsas 2023	x	x		x		
Maesen and Lamey 2023	x	x				
Zhang and Liu-Thompkins 2023					x	
Kan et al 2023		x		x		
Kumar et al 2023						x
Sok, Danaher, and Sok 2023					x	
Cascio Rizzo et al 2024						x
Elhelaly and Ray 2024						x
Ma et al 2024		x	x			
Tian et al 2024						x
Geyskens et al 2024		x		x		
Wiles et al 2024				x		
Chaker et al 2024					x	
Yazdani, Gopinath, and Carson 2024						x

Kanuri, Hughes, and Hodges 2024						x
Özturan, Deleersnyder, and Özsomer 2024				x		
Sklenarz et al 2024						x
Maesen 2024	x	x		x		
Friess et al 2024					x	
Vafainia et al 2024				x		
Paschmann et al 2025						x
Yazdani, Chakravarty, and Inman 2025						x
Fang, Qian, and Xie 2025		x				
Holtrop et al 2025		x		x		
Vomberg and Gegerfelt 2025						x
Rahman et al 2025			x			
Tran et al 2025						x
Zaefarian et al 2025						x
Park and Griffith 2025						x
Van Crombrugge et al 2025	x	x		x		
Ahearne, Pourmasoudi, and Habel 2025					x	
Haschka and Herwartz 2025		x				
Weiger et al 2025						x

Note: [a] "SF" is Saleseforce; "Other" includes word-of-mouth, warranty claims, store visits

List of Publications in Table W1

Ahearne, Molly, Mohsen Pourmasoudi, and Johannes Habel (2025), "Inside sales structures and firm performance," *Journal of the Academy of Marketing Science*, forthcoming.

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.

Aydinli, Aylin, Lien Lamey, Kobe Millet, Anne ter Braak, and Maya Vuegen (2021), "How do customers alter their basket composition when they perceive the retail store to be crowded? An empirical study," *Journal of Retailing*, 97 (2), 207-216.

Bachmann, Patrick, Markus Meierer, and Jeffrey Näf (2021), "The role of time-varying contextual factors in latent attrition models for customer base analysis," *Marketing Science*, 40 (4), 783-809.

Bhattacharya, Abhi, Neil A. Morgan, and Lopo L. Rego (2022), "Examining why and when market share drives firm profit," *Journal of Marketing*, 86 (4), 73-94.

Bombaij, Nick JF, and Marnik G. Dekimpe (2020), "When do loyalty programs work? The moderating role of design, retailer-strategy, and country characteristics," *International Journal of Research in Marketing*, 37 (1), 175-195.

Borah, Abhishek, S. Cem Bahadir, Anatoli Colicev, and Gerard J. Tellis (2022), "It pays to pay attention: How firm's and competitor's marketing levers affect investor attention and firm value," *International Journal of Research in Marketing*, 39 (1), 227-246.

Bornemann, Torsten, Cornelia Hattula, and Stefan Hattula (2020), "Successive product generations: Financial implications of industry release rhythm alignment," *Journal of the Academy of Marketing Science*, 48 (6), 1174-1191

Burchett, Molly R., Brian Murtha, and Ajay K. Kohli (2023), "Secondary selling: Beyond the salesperson–customer dyad," *Journal of Marketing*, 87 (4), 575-600.

Burmester, Alexa B., Jan U. Becker, Harald J. van Heerde, and Michel Clement (2015), "The impact of pre-and post-launch publicity and advertising on new product sales," *International Journal of Research in Marketing*, 32 (4), 408-417.

Campo, Katia, Lien Lamey, Els Breugelmans, and Kristina Melis (2021), "Going online for groceries: Drivers of category-level share of wallet expansion," *Journal of Retailing*, 97 (2), 154-172.

Cao, Zixia, Reo Song, Alina Sorescu, and Ansley Chua (2023), "Innovation potential, insider sales, and IPO performance: How firms can mitigate the negative effect of insider selling," *Journal of Marketing*, 87 (4), 550-574.

Cao, Zixia (2022), "Brand equity, warranty costs, and firm value," *International Journal of Research in Marketing*, 39 (4), 1166-1185.

Carson, Stephen J., and Mrinal Ghosh (2019), "An integrated power and efficiency model of contractual channel governance: Theory and empirical evidence," *Journal of Marketing*, 83 (4), 101-120.

Cascio Rizzo, Giovanni Luca, Francisco Villarroel Ordenes, Rumen Pozharliev, Matteo De Angelis, and Michele Costabile (2024), "How high-arousal language shapes micro-versus macro-influencers' impact," *Journal of Marketing*, 88 (4), 107-128.

Chaker, Nawar N., Johannes Habel, Nathaniel N. Hartmann, Felix Johannsen, and Heinrich Rusche (2024), "Quiet sellers: When introversion drives salesperson performance," *Journal of Retailing*, 100 (3), 456-474.

Cron, William L., Sascha Alavi, Johannes Habel, Jan Wieseke, and Hanaa Ryari (2021), "No conversion, no conversation: Consequences of retail salespeople disengaging from unpromising prospects," *Journal of the Academy of Marketing Science*, 49, 502-520.

Dall'Olio, Filippo, and Demetrios Vakratsas (2023), "The impact of advertising creative strategy on advertising elasticity," *Journal of Marketing*, 87 (1), 26-44.

Danaher, Peter J. (2022), "Optimal microtargeting of advertising," *Journal of Marketing Research*, 60 (3), 564-584.

Datta, Hannes, Kusum L. Ailawadi, and Harald J. Van Heerde (2017), "How well does consumer-based brand equity align with sales-based brand equity and marketing-mix response?" *Journal of Marketing*, 81 (3), 1-20.

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.

Datta, Hannes, Harald J. van Heerde, Marnik G. Dekimpe, and Jan-Benedict EM Steenkamp (2022), "Cross-national differences in market response: Line-length, price, and distribution elasticities in 14 Indo-Pacific rim economies," *Journal of Marketing Research*, 59 (2), 251-270.

De Jong, Ad, Nicolas A. Zacharias, and Edwin J. Nijssen (2021), "How young companies can effectively manage their slack resources over time to ensure sales growth: The contingent role of value-based selling," *Journal of the Academy of Marketing Science*, 49 (2), 304-326.

Dhaoui, Chedia, and Cynthia M. Webster (2021), "Brand and consumer engagement behaviors on Facebook brand pages: Let's have a (positive) conversation," *International Journal of Research in Marketing*, 38 (1), 155-175.
Elhelaly, Nehal, and Sourav Ray (2024), "Collaborating to innovate: Balancing strategy dividend and transactional efficiencies," *Journal of Marketing*, 88 (5), 153-179.
Fossen, Beth L., and Alexander Bleier (2021), "Online program engagement and audience size during television ads," *Journal of the Academy of Marketing Science*, 49, 743-761.
Friess, Maximilian, Till Haumann, Sascha Alavi, Alexandru Ionut Oproiescu, Christian Schmitz, and Jan Wieseke (2024), "The contingent effects of innovative digital sales technologies on B2B firms' financial performance," *International Journal of Research in Marketing*, forthcoming.
Garrido-Morgado, Álvaro, Óscar González-Benito, Mercedes Martos-Partal, and Katia Campo (2021), "Which products are more responsive to in-store displays: Utilitarian or hedonic?" *Journal of Retailing*, 97 (3), 477-491.
Geyskens, Inge, Barbara Deleersnyder, Marnik G. Dekimpe, and Didi Lin (2024), "Do consumers benefit from national-brand listings by hard discounters?" *Journal of the Academy of Marketing Science*, 52 (1), 97-118.
Gielens, Katrijn, Inge Geyskens, Barbara Deleersnyder, and Max Nohe (2018), "The new regulator in town: The effect of Walmart's sustainability mandate on supplier shareholder value," *Journal of Marketing*, 82 (2), 124-141.
Gielens, Katrijn, Marnik G. Dekimpe, Anirban Mukherjee, and Kapil Tuli (2023), "The future of private-label markets: A global convergence approach," *International Journal of Research in Marketing*, 40 (1), 248-267.
Gijsbrechts, Els, Katia Campo, and Mark Vroegrijk (2018), "Save or (over-) spend? The impact of hard-discounter shopping on consumers' grocery outlay," *International Journal of Research in Marketing*, 35 (2), 270-288.
Guitart, Ivan A., Jorge Gonzalez, and Stefan Stremersch (2018), "Advertising non-premium products as if they were premium: The impact of advertising up on advertising elasticity and brand equity," *International Journal of Research in Marketing*, 35 (3), 471-489.
Guitart, Ivan A., Guillaume Hervet, and Sarah Gelper (2020), "Competitive advertising strategies for programmatic television," *Journal of the Academy of Marketing Science*, 48, 753-775.
Guitart, Ivan A., and Stefan Stremersch (2021), "The impact of informational and emotional television ad content on online search and sales," *Journal of Marketing Research*, 58 (2), 299-320.
Haschka, Rouven E., and Helmut Herwartz (2025), "Utilizing managerial beliefs for set identification of price elasticities of demand," *Journal of the Academy of Marketing Science*, forthcoming.
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.
Holtrop, Niels, Kathleen Cleeren, Kelly Geyskens, and Peter C. Verhoef (2025), "The impact of nutrition claims on purchase behavior for food products," *International Journal of Research in Marketing*, forthcoming.
Homburg, Christian, Arnd Vomberg, and Stephan Muehlhaeuser (2020), "Design and governance of multichannel sales systems: Financial performance consequences in business-to-business markets," *Journal of Marketing Research*, 57 (6), 1113-1134.

Hoskins, Jake, Shyam Gopinath, J. Cameron Verhaal, and Elham Yazdani (2021), "The influence of the online community, professional critics, and location similarity on review ratings for niche and mainstream brands," *Journal of the Academy of Marketing Science*, 49, 1065-1087

Janani, Saeed, Ranjit M. Christopher, Atanas Nik Nikolov, and Michael A. Wiles (2022), "Marketing experience of CEOs and corporate social performance," *Journal of the Academy of Marketing Science*, 50 (3), 460-481.

Kan, Christina, Yan Liu, Donald R. Lichtenstein, and Chris Janiszewski (2023), "The negative and positive consequences of placing nonpromoted products next to promoted products," *Journal of Marketing*, 87 (36), 928-948.

Kanuri, Vamsi K., Christian Hughes, and Brady T. Hodges (2024), "Standing out from the crowd: When and why color complexity in social media images increases user engagement," *International Journal of Research in Marketing*, 41 (2), 174-193.

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

Kidwell, Blair, Jonathan Hasford, Broderick Turner, David M. Hardesty, and Alex Ricardo Zablah (2021), "Emotional calibration and salesperson performance," *Journal of Marketing*, 85 (6), 141-161.

Krämer, Martin, Christina Desernot, Sascha Alavi, Christian Schmitz, Felix Brüggemann, and Jan Wieseke (2022), "The role of salespeople in industrial servitization: How to manage diminishing profit returns from salespeople's increasing industrial service shares," *International Journal of Research in Marketing*, 39 (4), 1235-1252.

Kumar, Alok, Huanhuan Shi, Jenifer Skiba, Amit Saini, and Zhi Lu (2023), "Impact of buying groups on buyer–supplier relationships: Group–dyad interactions in business-to-business markets," *Journal of Marketing Research*, 60 (6), 1197-1220.

Lamey, Lien, Barbara Deleersnyder, Jan-Benedict EM Steenkamp, and Marnik G. Dekimpe (2018), "New product success in the consumer packaged goods industry: A shopper marketing approach," *International Journal of Research in Marketing*, 35 (3), 432-452.

Lamey, Lien, Els Breugelmans, Maya Vuegen, and Anne ter Braak (2021), "Retail service innovations and their impact on retailer shareholder value: Evidence from an event study" *Journal of the Academy of Marketing Science*, 49, 811-833.

Lenz, Isabell, Hauke A. Wetzel, and Maik Hammerschmidt (2017), "Can doing good lead to doing poorly? Firm value implications of CSR in the face of CSI," *Journal of the Academy of Marketing Science*, 45, 677-697.

Lim, Leon Gim, Kapil R. Tuli, and Marnik G. Dekimpe (2018), "Investors' evaluations of price-increase preannouncements," *International Journal of Research in Marketing*, 35 (3), 359-377.

Liu, Huan, Lara Lobschat, Peter C. Verhoef, and Hong Zhao (2021), "The effect of permanent product discounts and order coupons on purchase incidence, purchase quantity, and spending," *Journal of Retailing*, 97 (3), 377-393.

Ludwig, Stephan, Dennis Herhausen, Dhruv Grewal, Liliana Bove, Sabine Benoit, Ko De Ruyter, and Peter Urwin (2022), "Communication in the gig economy: Buying and selling in online freelance marketplaces," *Journal of Marketing*, 86 (4), 141-161.

Ma, Yu, Kusum L. Ailawadi, Mercedes Martos-Partal, and Óscar González-Benito (2024), "Dual branding by national brand manufacturers: Drivers and outcomes," *Journal of Marketing*, 88 (3), 69-87.

Maesen, Stijn, and Lien Lamey (2023), “The impact of organic specialist store entry on category performance at incumbent stores,” *Journal of Marketing*, 87 (1), 97-113
Maesen, Stijn, Lien Lamey, Anne ter Braak, and Léon Jansen (2022), “Going healthy: How product characteristics influence the sales impact of front-of-pack health symbols,” *Journal of the Academy of Marketing Science*, 50 (1), 108-130.
Maesen, Stijn (2024), “Introducing specialist private labels: How reducing manufacturers’ competing assortment size affects retailer performance,” *International Journal of Research in Marketing*, forthcoming.
Magnotta, Sarah, Brian Murtha, and Goutam Challagalla (2020), “The joint and multilevel effects of training and incentives from upstream manufacturers on downstream salespeople’s efforts,” *Journal of Marketing Research*, 57 (4), 695-716.
Maier, Erik, and Jaap Wieringa (2021), “Acquiring customers through online marketplaces? The effect of marketplace sales on sales in a retailer's own channels,” *International Journal of Research in Marketing*, 38 (2), 311-328.
Mathys, Juliane, Alexa B. Burmester, and Michel Clement (2016), “What drives the market popularity of celebrities? A longitudinal analysis of consumer interest in film stars,” *International Journal of Research in Marketing*, 33 (2), 428-448.
Moon, Sungkyun, Kapil R. Tuli, and Anirban Mukherjee (2022), “Does disclosure of advertising spending help investors and analysts?” *Journal of Marketing*, 87 (3), 359-382.
Nahm, Irene Y., Michael J. Ahearne, Nick Lee, and Seshadri Tirunillai (2022), “Managing positive and negative trends in sales call outcomes: The role of momentum,” *Journal of Marketing Research*, 59 (6), 1120-1140.
Nath, Pravin, Ahmet H. Kirca, Saejoon Kim, and Trina Larsen Andras (2019), “The effects of retail banner standardization on the performance of global retailers,” *Journal of Retailing*, 95 (3), 30-46.
Özturan, Peren, Barbara Deleersnyder, and Ayşegül Özsomer (2024), “Brand advertising competition across economic cycles,” *International Journal of Research in Marketing*, 41 (2), 325-343.
Park, Soo Hyung “Ralph” and David A. Griffith (2025), “The effect of customer-centric strategy and structure alignment on new product portfolio innovativeness and firm performance,” *Journal of the Academy of Marketing Science*, forthcoming.
Paschmann, Jens W., Hernán A. Bruno, Harald J. van Heerde, Franziska Völckner, and Kristina Klein (2025), “Driving mobile app user engagement through gamification,” *Journal of Marketing Research*, 62 (2), 249-273.
Rahman, Syed Mahmudur, Jamie Carlson, Noman H. Chowdhury, Siegfried P. Gudergan, Martin Wetzels, and Dhruv Grewal (2025), “Share of time in omnichannel retailing: Definition and measurement, *Journal of Retailing*, forthcoming.
Rajavi, Koushyar, Tarun Kushwaha, and Jan-Benedict EM Steenkamp (2022), “Brand equity in good and bad times: What distinguishes winners from losers in consumer packaged goods industries?” *Journal of Marketing*, 87 (3), 472-489.
Sawant, Rajeev J., Mahima Hada, and Simon J. Blanchard (2021), “Contractual discrimination in franchise relationships,” *Journal of Retailing*, 97 (3), 405-423.
Scholdra, Thomas P., Julian RK Wichmann, Maik Eisenbeiss, and Werner J. Reinartz (2022), “Households under economic change: How micro-and macroeconomic conditions shape grocery shopping behavior,” *Journal of Marketing*, 86 (4), 95-117.

Schulz, Petra, Edlira Shehu, and Michel Clement (2019), "When consumers can return digital products: Influence of firm-and consumer-induced communication on the returns and profitability of news articles," *International Journal of Research in Marketing*, 36 (3), 454–470.
Shehu, Edlira, Dominik Papies, and Scott A. Neslin (2020), "Free shipping promotions and product returns," *Journal of Marketing Research*, 57 (4), 640-658.
Sklenarz, Felix Anton, Alexander Edeling, Alexander Himme, and Julian RK Wichmann (2024), "Does bigger still mean better? How digital transformation affects the market share–profitability relationship," *International Journal of Research in Marketing*, forthcoming.
Sok, Keo Mony, Tracey S. Danaher, and Phyra Sok (2023), "Multiple psychological climates and employee self-regulatory focus: Implications for frontline employee work behavior and service performance," *Journal of Retailing*, 99 (2), 228-246.
Ter Braak, Anne, and Barbara Deleersnyder (2018), "Innovation cloning: The introduction and performance of private label innovation copycats," *Journal of Retailing*, 94 (3), 312-327.
Tian, Min, David W. Kaufman, Saul Shiffman, and Neeraj Arora (2024), "Over-the-counter drug consumption: How consumers deviate from label instructions," *Journal of Marketing Research*, 61 (3), 430-450.
Tran, Hai-Anh, Yuliya Strizhakova, Bach Nguyen, and Samuel G. B. Johnson (2025), "Expressions of customer rumination in online posts and firm responses," *Journal of the Academy of Marketing Science*, forthcoming.
Umashankar, Nita, Kihyun Hannah Kim, and Thomas Reutterer (2023), "Understanding customer participation dynamics: The case of the subscription box," *Journal of Marketing*, 87 (5), 719-735.
Vafainia, Saeid, Robert P. Rooderkerk, Els Breugelmans, and Tammo HA Bijmolt (2024), "Decision support system development for store flyer space allocation: Leveraging own-and cross-category sales effects," *International Journal of Research in Marketing*, forthcoming.
Van Ewijk, Bernadette J., Astrid Stubbe, Els Gijsbrechts, and Marnik G. Dekimpe (2021), "Online display advertising for CPG brands: (When) does it work?" *International Journal of Research in Marketing*, 38 (2), 271-289.
Van Ewijk, Bernadette J., Els Gijsbrechts, and Jan-Benedict EM Steenkamp (2022a), "What drives brands' price response metrics? An empirical examination of the Chinese packaged goods industry," *International Journal of Research in Marketing*, 39 (1), 288-312.
Van Ewijk, Bernadette J., Els Gijsbrechts, and Jan-Benedict EM Steenkamp (2022b), "The dark side of innovation: How new SKUs affect brand choice in the presence of consumer uncertainty and learning," *International Journal of Research in Marketing*, 39 (4), 967-987.
Vieira, Valter Afonso, Marcos Inácio Severo de Almeida, Raj Agnihotri, Nôga Simões De Arruda Corrêa da Silva, and S. Arunachalam (2019), "In pursuit of an effective B2B digital marketing strategy in an emerging market," *Journal of the Academy of Marketing Science*, 47 (6), 1085-1108.
Vomberg, Arnd, Christian Homburg, and Olivia Gwinner (2020), "Tolerating and managing failure: An organizational perspective on customer reacquisition management," *Journal of Marketing*, 84 (5), 117-136.
Vomberg, Arnd and Mats von Gegerfelt (2025), "Unveiling investment vs. ownership perspectives among NFT buyers: A segmentation study exploring engagement patterns in NFT markets," *International Journal of Research in Marketing*, forthcoming.

Van Crombrugge, Michiel, Els Breugelmans, Kathleen Cleeren, and Scott A. Neslin (2025), "Retailer marketing mix response when launching a direct channel: Not all retailers are alike," *Journal of the Academy of Marketing Science*, forthcoming.

Weiger, Welf H., Wendy W. Moe, Hauke A. Wetzel, and Maik Hammerschmidt (2025), "Do social media fans walk their talk? The impact of advocacy and criticism on own purchasing," *International Journal of Research in Marketing*, forthcoming.

Wetzel, Hauke A., Stefan Hattula, Maik Hammerschmidt, and Harald J. van Heerde (2018), "Building and leveraging sports brands: Evidence from 50 years of German professional soccer," *Journal of the Academy of Marketing Science*, 46, 591-611.

Widdecke, Kai A., Wiebke IY Keller, Karen Gedenk, and Barbara Deleersnyder (2022), "Drivers of the synergy between price cuts and store flyer advertising at supermarkets and discounters," *International Journal of Research in Marketing*, 40 (2), 455-474.

Wiegand, Nico, Yuri Peers, and Alexander Bleier (2023), "Software multihoming to distal markets: Evidence of cannibalization and complementarity in the video game console industry," *Journal of the Academy of Marketing Science*, 51 (2), 393-417.

Wiles, Michael A., Saeed Janani, Darima Fotheringham, and Chadwick J. Miller (2024), "A longitudinal examination of the relationship between national-level per capita advertising expenditure and national-level life satisfaction across 76 countries," *Marketing Science*, 43 (3), 542-563.

Wiseman, Phillip, Michael Ahearne, Zachary Hall, and Seshadri Tirunillai (2022), "Onboarding salespeople: Socialization approaches," *Journal of Marketing*, 86 (6), 13-31.

Yang, Fan, Qian, Yi, and Xie, Hui (2025), "Addressing endogeneity using a two-stage copula generated regressor approach," *Journal of Marketing Research*, forthcoming.

Xu, Juan, Michel Van der Borgh, Edwin J. Nijssen, and Son K. Lam (2022), "Why salespeople avoid big-whale sales opportunities," *Journal of Marketing*, 86 (5), 95-116

Yazdani, Elham, Shyam Gopinath, and Stephen J. Carson (2024), "The role of reviewer badges in the dynamics of online reviews," *International Journal of Research in Marketing*, forthcoming.

Yazdani, Elham, Chakravarty, Anindita, and Jeffrey Inman (2025), "Racial Inequity in Donation-Based Crowdfunding Platforms: The role of facial emotional expressiveness," *Journal of Marketing*, forthcoming.

Zaefarian, Ghasem, Constantine S. Katsikeas, Zhaleh Najafi-Tavani, and Matthew J. Robson (2025), "Unpacking when and how customer involvement as co-developer affects supplier new product performance," *Journal of the Academy of Marketing Science*, forthcoming.

Zhang, Junzhou, and Yuping Liu-Thompkins (2024), "Personalized email marketing in loyalty programs: The role of multidimensional construal levels," *Journal of the Academy of Marketing Science*, 52 (1), 196-216.

Zhang, Yufei, Clay M. Voorhees, Chen Lin, Jeongwen Chiang, G. Tomas M. Hult, and Roger J. Calantone (2022), "Information search and product returns across mobile and traditional online channels," *Journal of Retailing*, 98 (2), 260-276.

Zhao, Yanhui, Yufei Zhang, Joyce Wang, Wyatt A. Schrock, and Roger J. Calantone (2020), "Brand relevance and the effects of product proliferation across product categories," *Journal of the Academy of Marketing Science*, 48, 1192-1210.

Table W2: Publications Using Copula Correction in Leading Marketing Journals

Characteristics	Number	Characteristics	Number	Characteristics	Number
Endogenous Regressors		Outcome Type		Sample Size	
Product	20	Continuous	89	≤ 100	1
Price	35	Discrete Choice	15	101—1,000	40
Place	10	Count	3	1,001—5,000	8
Promotion	28			5,001—50,000	18
Sales Force & CRM	18	Panel Data	66	$\geq 50,001$	33
Other	40				

Note: "Other" includes word-of-mouth, warranty claims, store visits, etc. The list of journals includes *Journal of Marketing, Journal of Marketing Research, Marketing Science, Journal of Consumer Research, Journal of the Academy of Marketing Science, Journal of Retailing, International Journal of Research in Marketing,* and *Journal of Consumer Psychology.* See Web Appendix Table W1 for a detailed list of these papers with their substantive areas. The total number of unique journal publications is n=100. Some columns sum to more than the number of unique papers since multiple copulas may be used or multiple models estimated in a paper using copulas.

Table W3: Examples of Applications Involving Higher-order Endogenous Terms.

Study	Higher-Order Endogenous Regressors	CHI*
Burmester eta al. (2015)	Ad Stock * Publicity Stock	Yes
Blauw and Franses (2016)	Mobile Phone Ownership^2	Yes
Lenz, Wetzel, and Hammerschmidt (2017)	Corporate Social Responsibility^2	No
Lamey et al. (2018)	Promotion Intensity * Store context	No
Gielens et al. (2018)	R& D * Retailer Power	No
Yoon et al. (2018)	Knowledge * Government Activity	Yes
Atefi et al. (2018)	Trained Percentage^2	Yes
	Trained Percentage *Performance Diversity	
Guitart, Gonzalez, and Stremersch (2018)	Advertising * Price	No
Wetzel et al. (2018)	Recruitment Spend * Brand Age	No
Keller, Deleersnyder, and Gedenk (2019)	Price Index * Price Premium	No
Heitmann et al. (2020)	Complexity *Segment Typicality	No
Vomberg, Homburg, and Gwinner (2020)	Failure Culture* Reacquisition Policies	No
Guitart and Stremersch (2021)	Ad Stock^2, Price^2, Informational^2	Yes
Magnotta, Murtha, and Challagalla (2020)	Salesperson Training*Salesperson Incentive	No
Homburg, Vomberg, and Muehlhaeuser (2020)	Direct Channel Usage*Formalization	No
Liu et al. (2021)	Price Discount^2, order Coupon^2	Yes
Kramer et al. (2022)	Industrial Service Share^2	Yes

CHI: copula correction terms for high-order terms of endogenous regressors included.

WEB APPENDIX B: DOUBLE ROBUSTNESS PROPERTY OF COPULA CORRECTION

This section demonstrates the double robustness of copula correction using control function in that the error distribution does not need to follow a normal distribution and the regressor-error dependence does not need to follow GC dependence. Consider the following structural equation model according to the data generating process from Figure 2.d:

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i \tag{W1}$$

$$E_i = U_i + \xi_i \tag{W2}$$

where U_i denotes the endogenous part of the error E_i and captures the joint effects of all unobserved confounders, and ξ_i denotes the exogenous disturbance term that is independent of P_i, W_i and U_i. With the intercept μ in the model and, without loss of generality, both U_i and ξ_i have means of zero.

As noted in the main text, the exogenous part of E_i, ξ_i, simply adds noise but does not affect endogeneity correction. Because ξ_i does not need to follow a normal distribution or any GC assumption in order for the augmented OLS regression to correct for bias, this means that the identification of the model for copula correction using control functions does not require the structural error E_i be normally distributed or follow the GC dependence structure jointly with regressors.

We illustrate this double robustness property of copula correction using a simulation study. We generate P_t, W_t, U_t using the same GC distribution as in Equations W3 to W7:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ U_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pU} \\ \rho_{pw} & 1 & 0 \\ \rho_{pU} & 0 & 1 \end{bmatrix} \right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right), \quad \text{(W3)}$$

$$U_t = \Phi^{-1}(\Phi(U^*)) = 1 \cdot U_t^*, \quad \text{(W4)}$$

$$\xi_t \sim N(0, 1), \quad \text{(W5)}$$

$$P_t = H^{-1}(U_{P,t}) = H^{-1}(\Phi(P_t^*)), \;\; W_t = L^{-1}(U_{W,t}) = L^{-1}(\Phi(W_t^*)), \quad \text{(W6)}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + E_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + U_t + \xi_t \quad \text{(W7)}$$

In the simulation, we use the Gamma (1,1) distribution for P_t and the exponential distribution Exp(1) with rate 1 for W_t. We consider two distributions for ξ_t: uniform on [-0.5,0.5] and the lognormal(0,1)-$e^{0.5}$ distribution. Thus, the error term $E_i = U_i + \xi_i$ will not follow a normal distribution because of nonnormality of ξ_i. Furthermore, E_i will not follow a GC model with regressors. However, Assumptions 1 and 2 of 2sCOPE still holds because U_i is normally distributed and follow a GC model with regressors. Thus, we expect 2sCOPE to be able to recover true parameter values. We then compute Y_t using Equation W7 with parameter values given in Table W4. Sample size is n=1,000 per dataset. For each dataset, we apply OLS and the 2sCOPE estimation described in Table 6. A total of 1,000 datasets were generated.

Table W4 reports the mean and standard deviation of the model estimates across 1,000 simulated data sets. As shown in Table W4, OLS has large bias for both distributions of ξ_t. As expected, 2sCOPE corrects for the OLS estimation bias and recovers the true parameter values despite the error term E being nonnormally distributed and does not follow a GC

Table W4: Results of the Simulation Study: Double Robustness of Copula Correction

Distribution	Skewness			OLS		2sCOPE	
of ξ_t	of E_t	Param.	True	Mean	SE	Mean	SE
Unif[-0.5,0.5]	0.00	μ	1	0.69	0.05	1.00	0.06
		α	1	1.57	0.04	1.00	0.07
		β	-1	-1.26	0.03	-1.00	0.04
		σ_E	1.04	0.91	0.02	1.04	0.04
Lnorm(0,1)- $e^{0.5}$	3.68	μ	1	0.69	0.11	1.00	0.14
		α	1	1.57	0.08	1.00	0.16
		β	-1	-1.26	0.08	-1.00	0.11
		σ_E	2.37	2.31	0.27	2.37	0.26

model with regressors, demonstrating the double robustness property of the 2sCOPE method in that a GC regressor-error dependence is not required.

Table W5 evaluates the performance of copula correction when the distribution of U_t follows a nonnormal distribution. That is, we use the same simulation set up as above except that $U_t = t_4^{-1}(\Phi(U_t^*))$ instead of $U_t = U_t^*$, where t_4 represents the CDF for the t-distribution with 4 degrees of freedom. Thus, both U_t and E_t are nonnormally distributed, violating Assumptions 1 and 2 of the 2sCOPE procedure. As shown in Table W5, 2sCOPE can still correct for the OLS estimation bias and recover the true model parameters well. The results show that although Assumptions 1 and 2 are used in the derivation of 2sCOPE, these assumptions are not strictly required; 2sCOPE demonstrates desirable robustness to the violations of Assumptions 1 and 2.

Table W5: Results of the Simulation Study: Robustness of Copula Correction with a misspecified U distribution.

Distribution	Skewness			OLS		2sCOPE	
of ξ_t	of E_t	Param.	True	Mean	SE	Mean	SE
Unif[-0.5,0.5]	0.00	μ	1	0.57	0.07	0.99	0.09
		α	1	1.78	0.06	1.01	0.13
		β	-1	-1.35	0.05	-1.00	0.07
		σ_E	1.44	1.26	0.07	1.43	0.09
Lnorm(0,1)- $e^{0.5}$	2.99	μ	1	0.57	0.12	0.99	0.16
		α	1	1.78	0.10	1.02	0.22
		β	-1	-1.35	0.09	-1.01	0.13
		σ_E	2.57	2.47	0.30	2.57	0.30

WEB APPENDIX C: ICON: AN INDEX OF COPULA-MODEL NONIDENTIFICATION

When properly applied with the underlying assumptions and data requirements being met, copula correction can be a powerful tool for addressing endogeneity bias using nonexperimental data. The main text shows that copula control function methods work under considerably weaker conditions and thus are more robust and applicable than previously believed. However, it is important to check boundary conditions to minimize the potential pitfalls of incorrect applications of copula correction.

As noted in the main text and shown in existing research, when the main identification assumptions of copula correction are violated, copula models can become weakly identified or unidentified, resulting in poor performance of copula correction (e.g., Park and Gupta 2012; Haschka 2022; Qian and Xie 2024; Yang, Qian, and Xie 2024a). Model weak identification or nonidentification can occur when regressor distributional requirements are not satisfied. When exogenous regressors lack sufficient relevance or sufficient non-normality to compensate for insufficient non-normality of endogenous regressors, copula terms become nearly collinear with existing regressors, yielding estimates with significant finite-sample bias and huge standard errors (Yang, Qian, and Xie 2024a). The evaluation using simulation studies has also shown that when the regressor-error dependence follows a linear model instead of the GC model, copula correction yields significantly biased estimates with huge standard errors (Haschka 2022; Qian and Xie 2024), which is caused by nearly singular Hessian matrices of the almost flat likelihood functions in the misspecified GC copula model. Thus, theory suggests significantly inflated standard errors of estimates as warning signs of nearly unidentified models caused by misspecified regressor-error dependence.

Because weakly identified copula models can yield significantly biased and imprecise estimates, copula correction using such models is inappropriate. In this section, we propose a simple measure, named ICON (an Index of Copula-model Nonidentification), to flag such scenarios in which the deployed copula correction model/approach is likely inappropriate or needs further refinement. The ICON measure is defined as the ratio of the standard error of a copula corrected estimate to the standard error of the corresponding uncorrected estimate. The rationale for this measure is that when the copula model becomes weakly or nonidentified, the standard errors of the copula corrected estimates will become large. Prior research has shown for such models, the standard errors of copula corrected estimates are typically more than 8-10 times of those of copula uncorrected estimates (Park and Gupta 2012; Haschka 2022; Qian and Xie 2024; Yang, Qian, and Xie 2024a). To be conservative, we suggest ICON>6 as a threshold value for weakly identified or non-identified copula models.

To illustrate the use of ICON, we first consider the case of model nonidentification caused by the lack of meeting the regressor distributional requirements. We generate P_t, W_t, U_t using the following GC model

$$\begin{pmatrix} P_t^* \\ W_t^* \\ E_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix}\right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}\right), \quad \text{(W8)}$$

$$E_t = G^{-1}(U_{E,t}) = G^{-1}(\Phi(E_t^*)) = \Phi^{-1}(\Phi(E^*)) = 1 \cdot E_t^*, \quad \text{(W9)}$$

$$P_t = H^{-1}(U_{P,t}) = H^{-1}(\Phi(P_t^*)), \quad W_t = L^{-1}(U_{W,t}) = L^{-1}(\Phi(W_t^*)), \quad \text{(W10)}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + E_t = 1 + 1 \cdot P_t + (-1) \cdot W_t + E_t, \quad \text{(W11)}$$

In the simulation, we use standard normal distributions for the marginal distributions of P_t and W_t. The GC model is unidentified because both endogenous regressor P and exogenous regressor W are normally distributed, and the population copula term is perfectly correlated with the regressors, leading to an unidentified population model. In any finite sample, however, the regressors are not exactly normally distributed and copula terms are severely (but not perfectly) collinear with the existing regressors. Thus, in finite samples, the exogenous regressor W_t lacks sufficient nonnormality to compensate for the normality of the endogenous regressor P_t. We demonstrate here that the ICON statistics can detect such nonidentified models in finite samples.

Table W6: Detect Nonidentified Copula Models with ICON: Normal Regressors

Distribution				OLS		2sCOPE		ICON
P	W	Parameters	True	Mean	SE	Mean	SE	(SE_{2scope}/SE_{ols})
Normal	Normal	μ	1	1.00	0.025	1.001	0.031 (0.033)	1.4
		α	1	1.666	0.030	1.662	0.547 (0.529)	**21.6**
		β	-1	-1.332	0.030	-1.331	0.287 (0.264)	**10.8**
		$\rho_{p\xi}$	0.5	-	-	0.001	0.363 (0.347)	**14.0**
		σ_ξ	1	0.816	0.018	0.943	0.157 (0.171)	**7.0**

Note: Table presents the means and standard deviations of parameter estimates over 1,000 simulated datasets. For 2sCOPE, the numbers within the parenthesis are the averages of bootstrapped standard error estimates over 1,000 simulated datasets. The ICON column "SE_{2scope}/SE_{ols}" presents the averages of the ratio of the bootstrapped standard error estimates for 2sCOPE to the standard error estimates for OLS over 1,000 simulated datasets.

Table W6 summarizes the results over 1,000 simulated data sets. We observe that 2sCOPE yields estimates that are, on average, very close to the OLS estimates (i.e, not correcting for endogeneity bias) and also have large estimation variability. Overall, the copula correction appears to perform worse than the OLS estimation when the copula model is nonidentified. The ICON statistic is able to detect model nonidentification because the ICON statistics are far greater than the cutoff value of 6, flagging these copula estimates as inappropriate to use. One can also detect the failure of regressor distribution and relevance requirements using the guidelines provided in Yang, Qian, and Xie (2024a) (see the boundary condition b in Flowchart 5 in the main text) that examines the normality of regressors and first-stage F statistics for the relevance of exogenous regressors). Their guideline is akin to using the first-stage F statistics to detect weak IVs in IV regressions (Staiger and Stock 1997). While their guideline is developed for use with 2sCOPE under the GC joint model for all regressors and specifically for checking regressor distribution and relevance requirements, the ICON statistics can be viewed as a more direct and general measure for detecting model nonidentification and can detect nonidentified copula models due to other causes, as shown next.

The ICON statistics can also help detect model identification due to violations of the regressor-error GC assumption. Past research has also shown that when the regressor-error dependence follows a linear model instead of the GC model, copula correction yields significantly biased estimates with huge standard errors (Haschka 2022; Qian and Xie 2024). The ICON statistics can be used to detect such misspecified copula models. To further demonstrate this point, consider the following example using simulated data. In this example [29], we consider the following data generating process (DGP):

[29]We thank one referee for suggesting the example.

$$\begin{pmatrix} P_{1t}^* \\ P_{2t}^* \\ W_t^* \\ U_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \rho_{p_1 w} & 0 \\ 0 & 1 & 0 & \rho_{p_2 U} \\ \rho_{p_1 w} & 0 & 1 & 0 \\ 0 & \rho_{p_2 U} & 0 & 1 \end{bmatrix} \right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0.5 \\ 0.5 & 0 & 1 & 0 \\ 0 & 0.5 & 0 & 1 \end{bmatrix} \right), \tag{W12}$$

$$U_t = \Phi^{-1}(\Phi(U_t^*)) = 1 \cdot U_t^*, \tag{W13}$$

$$\xi_t \sim N(0,1) \tag{W14}$$

$$P_{1t} = H_1^{-1}(\Phi(P_{1t}^*)), \quad P_{2t} = H_2^{-1}(\Phi(P_{2t}^*)), \tag{W15}$$

$$P_t = P_{1t} + P_{2t}, \quad W_t = L^{-1}(\Phi(W_t^*)), \tag{W16}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + E_t = 5 + (-1) \cdot P_t + 3 \cdot W_t + U_t + \xi_t, \tag{W17}$$

where U_t^* and P_{2t}^* are correlated ($\rho_{p_2U} = 0.5$), generating the endogeneity of P_t; W_t is exogenous and uncorrelated with U_t^*; W_t^* and P_{1t}^* are correlated ($\rho_{p_1w} = 0.5$), and therefore W_t and P_t are correlated. In this example, P_t represents the price at the market or occasion t, and the price P_t is the sum of the cost P_{1t} and the markup P_{2t}. The analyst observes the prices P_t, sales Y_t, and an exogenous variable (W_t, e.g., ad spending). There is an unobserved (omitted) variable U_t (e.g., temperature) that correlates with sales and the markup ($\rho_{p_2U} = 0.5$). Furthermore, cost P_{1t} is correlated with W_t ($\rho_{p_1w} = 0.5$). We set W to have a skewed normal distribution: $L(\cdot)$ is the CDF for skewed normal distribution with location=0, scale=0, slant parameter =10; $H_1(\cdot)$ for cost is the CDF for $Unif(-1,1)$ and $H_2(\cdot)$ for markup is the CDF for $Unif(-2,2)$, and the price is the sum of cost and markup. We also have an unobserved exogenous sales shock $\xi_t \sim N(0,1)$.

Suppose the analyst assumes (P_t, W_t, U_t) follows the GC distribution and use the following 2sCOPE procedure

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + \gamma C_{t,p|w} + \epsilon_t \tag{W18}$$

where $C_{t,p|w} = P_t^* - \hat{\delta} W_t^*$. For the above 2sCOPE to work as intended, we require the assumptions in Table 4 to be reasonably satisfied (as note above, 2sCOPE is robust to a range of departures from Assumptions 1 and 2 in Table 4). The above DGP satisfies all assumptions except Assumption 2 because (P_t, W_t, U_t) does not jointly follow the GC distribution.

To examine the performance of copula correction, we conduct a simulation study that generates 1,000 datasets from the above DGP. For each simulated data, we conduct both OLS and 2sCOPE estimation using Equation W18. As a comparison, we also generate data from a DGP that satisfies the condition that (P_t, W_t, U_t) follows the GC model below. This represents that U denotes the combined effect of many omitted variables (e.g., unobserved product attributes, retailer decisions, etc.) that affect both the cost and markup:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ U_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pU} \\ \rho_{pw} & 1 & 0 \\ \rho_{pU} & 0 & 1 \end{bmatrix} \right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right), \tag{W19}$$

$$U_t = \Phi^{-1}(\Phi(U^*)) = 1 \cdot U_t^*, \tag{W20}$$

$$\xi_t \sim N(0, 1), \tag{W21}$$

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

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + E_t = 5 + (-1) \cdot P_t + 3 \cdot W_t + U_t + \xi_t, \tag{W23}$$

where $H(\cdot)$ is the CDF for $Unif(-2, 2)$ and we set W as a skewed normal distribution: $L(\cdot)$ is the CDF for skewed normal distribution with location=0, scale=0, slant parameter =10.

Table W7 summarizes the results over 1,000 simulated data sets. The result shows that for the misspecified DGP, the 2sCOPE improves upon OLS to some extent but considerable bias remains: the α parameter has a bias of 0.25 (-0.751+1) for 2sCOPE instead of 0.36 (-0.647+1) for OLS in Table W7. Table W7 also shows the 2sCOPE α estimate for the endogenous regressor P_t has a significantly inflated standard error relative to the OLS α estimate with an average inflation of 7.8 (Table W7). This occurs because gross violations of GC dependence structure can cause loss of model identification and lead to weakly- or non-identified models, leading to significantly inflated standard errors relative to those of the uncorrected estimates (Park and Gupta 2012; Haschka 2022; Qian and Xie 2024; Yang, Qian, and Xie 2024a). Consistent with the literature, the inflated standard errors are a red flag for model nonidentification or potential gross violations of copula assumptions. The ICON statistic for the coefficient α estimate of the endogenous regressor exceeds the cutoff value of 6, indicating potential model nonidentification issues. Here, the regressor distributional and relevance requirements are satisfied, so the culprit can be attributed to the violation of GC assumption. By contrast, Table W7 shows when the underlying DGP is correctly specified, the OLS and 2sCOPE estimates have similar variability and the average ICON statistics for α is only about 3.7 (Table W7), much less than the threshold of 6.

In practice, the analyst will consider what to do if the ICON statistics indicate potential

Table W7: Detect Nonidentified Copula Model with ICON: Misspecified $P-U$ Dependence.

$P-U$			OLS		2sCOPE		ICON
Dependence	Param.	True	Mean	SE	Mean	SE	$\text{SE}_{\text{2scope}}/\text{SE}_{\text{ols}}$
Misspecified	μ	5	5.128	0.069	5.090	0.120 (0.118)	1.6
	α	-1	-0.647	0.035	-0.751	0.265 (0.262)	**7.8**
	β	3	2.840	0.070	2.887	0.139 (0.135)	1.9
	σ_E	1.44	1.342	0.030	1.386	0.068 (0.079)	2.6
Correctly Specified	μ	5	5.389	0.073	5.000	0.132 (0.132)	1.7
	α	-1	-0.457	0.039	-0.997	0.149 (0.148)	3.7
	β	3	2.509	0.078	2.998	0.156 (0.153)	2.0
	σ_E	1.44	1.298	0.030	1.416	0.067 (0.066)	2.2

Note: Table presents the means and standard deviations of parameter estimates over 1,000 simulated datasets. For 2sCOPE, the numbers within the parenthesis are the averages of bootstrapped standard error estimates over 1,000 simulated datasets. The ICON column "$\text{SE}_{\text{2scope}}/\text{SE}_{\text{ols}}$" presents averages of the ratio of the bootstrapped standard error estimates for 2sCOPE to the standard error estimates for OLS over 1,000 simulated datasets. The $P-U$ dependence is misspecified in copula correction when DGP follows Equations W12-W17 and correctly specified when DGP follows Equations W19-W23.

model nonidentification issues due to potential violations of GC dependence. The analyst can check the appropriateness of the model specifications and revise the copula correction strategies if alternative copula specifications make more sense.

We offer two potential solutions. One solution is to include relevant control variables. Note that Assumption 2 in Table 4 only requires that the unexplained dependence between P and U (or E) given exogenous regressors in W to be adequately captured by the GC model. P and U do not need to follow a GC dependence model unconditionally on exogenous regressors (Hu, Qian, and Xie 2025). In the above example, one may consider adding control variables that proxy or predict well either cost or markup so that one part of price is explained away or determined sufficiently, such that the unexplained dependence between P and U by relevant control variables can be captured by a GC relationship even if P and U do not follow GC model unconditionally on W. Thus, exogenous variables can play an important role in copula correction just like the IV approach. In many cases, IVs are plausible only after good control variables are included in the model. For example, proximity to a college or hospital is often used as an IV that is valid when the model includes important control variables to account for regional differences (e.g., Ebbes et al. 2005). Although exogenous control variables can play important roles in both copula correction and IV methods, copula correction does not require IVs.

The other solution when ICON detects potential model misspecification is that the analyst may consider collecting additional data (e.g., cost) and using alternative copula correction methods. In the above example, one may consider a revised 2sCOPE procedure with two copula correction terms for cost and markup separately.

$$Y_t = \mu + \alpha \cdot (P_{1t} + P_{2t}) + \beta \cdot W_t + \gamma_1 C_{t,p_1|w} + \gamma_2 C_{t,p_2|w} + \epsilon_t, \quad \text{(W24)}$$

where P_{1t} represent the cost part computed using the collected data, and P_{2t} represent the remaining markup part of the price (i.e., $P_t - P_{1t}$). Table W8 reports the results from 1,000 simulated data sets. The refined 2sCOPE using two copula correction terms for cost and markup separately corrects the endogeneity bias of the OLS estimates. The standard error of the price coefficient is substantially smaller now, and the ICON statistics indicate no model identification issues.

In conclusion, it is prudent to assess the plausibility of copula correction by considering the source of endogeneity to design suitable copula correction procedures. Using the ICON statistics helps identify potential model nonidentification and misspecification issues. An ICON ratio > 6 for copula corrected coefficient estimates for endogenous regressors indicates potential model identification issues and model violations. When this occurs, one

Table W8: Results of the Simulation Study: Refined Copula Correction with Two Copula Correction Terms.

		OLS		2sCOPE		ICON
Param.	True	Mean	SE	Mean	SE	SE_{2scope}/SE_{ols}
μ	5	5.128	0.069	5.001	0.092 (0.091)	1.2
α	-1	-0.647	0.035	-0.994	0.150 (0.151)	4.2
β	3	2.840	0.070	3.000	0.100 (0.100)	1.3
σ_E	1.44	1.342	0.030	1.422	0.069 (0.067)	2.0

See Note under Table W7.

can consider revising model specifications, adding relevant control variables, refining copula correction strategies, or using other endogeneity correction methods.

WEB APPENDIX D: OPTIMAL ALGORITHM FOR COPULA TRANSFORMATION

This section summarizes further results from simulation studies regarding the proper construction of copula transformation. We also provide an interactive applet supplement accessible at https://unknown8866.github.io/histogram-webpage/ for readers to visually explore the results of the simulation study with the source `R` code available at https://osf.io/by2ge/?view_only=27cc862a9c02446abbafd3a745722603.

An Example of Copula Transformation

To demonstrate how the empirical rank-based copula transformation is constructed, consider the example of the selling price of twenty goods from a small retailer, as shown in Table W9. The construction of the empirical rank-based copula follows two steps. First, the observations are ordered and mapped to a ranked percentile according to the empirical cumulative distribution, $F(\cdot)$. For example, the first observation (of twenty) is $\frac{1}{20}$, or 5% of the cumulative observations; the second observation is $\frac{2}{20}$, or 10%, and so on. The second step computes the inverse normal CDF of that ranked percentile as shown in the column "Price*": an observation in the bottom 5% (or fifth percentile) maps onto the far left end of a standard normal distribution, in this case about -1.6449 standard deviations below 0.

One item from Table W9 is of particular importance: the last observation is technically the 100th percentile, however, the inverse normal CDF of the 100th percentile is undefined. This is because the probability (reflected as F) must be between 0 and 1. The latent copula data, Price*, for the 20th observation here reflects an adjustment, where $F(\cdot)$ becomes the observation count divided by the observation count plus one (i.e., $\frac{n}{n+1} = \frac{20}{21}$). That is, we compute the copula transformation using Equation 11. Besides ensuring that the copula transformed values maintain the same rank order as the original regressor values for any sample size [30], the percentile adjustment for the maximum value yields a theoretically valid maximum value of the underlying copula data, and stabilizes the copula transformation without producing an extremely transformed value.

Table W9: Example Creation of the Rank-based Gaussian Copula

Obs	Price	F(Price)	Price*	Obs	Price	F(Price)	Price*
1	$14.00	0.05	-1.6449	11	$32.10	0.55	0.1257
2	$15.20	0.10	-1.2816	12	$33.00	0.60	0.2533
3	$16.30	0.15	-1.0364	13	$34.60	0.65	0.3853
4	$16.50	0.20	-1.0364	14	$34.90	0.70	0.3853
5	$21.00	0.25	-0.6745	15	$37.00	0.75	0.6745
6	$24.20	0.30	-0.5244	16	$42.00	0.80	0.8416
7	$27.00	0.35	-0.3853	17	$43.50	0.85	1.0364
8	$29.00	0.40	-0.2533	18	$44.10	0.90	1.2816
9	$29.50	0.45	-0.2533	19	$45.00	0.95	1.6449
10	$30.00	0.50	0.0000	20	$47.80	0.9524+	1.6684

+: To avoid generating undefined latent copula data, the rank for the maximum value of Price is changed from 1 to n/(n+1), which is 20/21=0.9524 for the sample size $n = 20$ here.

[30] By contrast, in their example of 100 observations, Papies, Ebbes, and Van Heerde (2017) set the percentile for the last observation to 0.99, which is the same as the second to last observation even though these two raw data points do not have the same rank order.

Simulation Study Setup and Findings

In this study, we use the following DGP that is the same as specified in Equations 1-4 in Becker, Proksch, and Ringle (2022):

$$\begin{bmatrix} E_t^* \\ P_t^* \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.50 \\ 0.50 & 1 \end{bmatrix}\right) \quad \text{(W25)}$$

$$E_t = \Phi^{-1}(\Phi(E_t^*)) \quad \text{(W26)}$$

$$P_t = \Phi(P_t^*) \quad \text{(W27)}$$

$$Y_t = \mu + \alpha P_t + E_t = -1P_t + E_t, \quad \text{(W28)}$$

where Y_t, P_t, and E_t represent the dependent variable, endogenous regressor, and the error term, respectively. The DGP specifies a linear model with the endogenous regressor P following a uniform distribution, and a correlation coefficient of 0.50 between P_t^* and the error term E_t. The simulation study varies in sample size N from 100 to 60,000 (100, 200, 400, 600, 800, 1,000, 2,000, 4,000, 6,000, 8,000, 10,000, 20,000, 40,000, and 60,000). For each sample size, we generate 1,000 datasets from the above DGP.

For each generated data set, we apply OLS, the Park and Gupta (P&G) method using the algorithm in Equation 12 to obtain generated regressor, the P&G method using the algorithm in Equation 11, and the integrating kernel density estimates (IKDE) to obtain the generated regressor in estimating the structural model. While the intercept term $\mu = 0$ in the DGP, the estimation does not assume this a-priori but instead estimates the intercept parameter jointly with other model parameters. The difference between the average of the estimates across 1,000 simulated datasets and its true value is the bias of an estimator, which is plotted in Figure W1 for α (discussed further below).

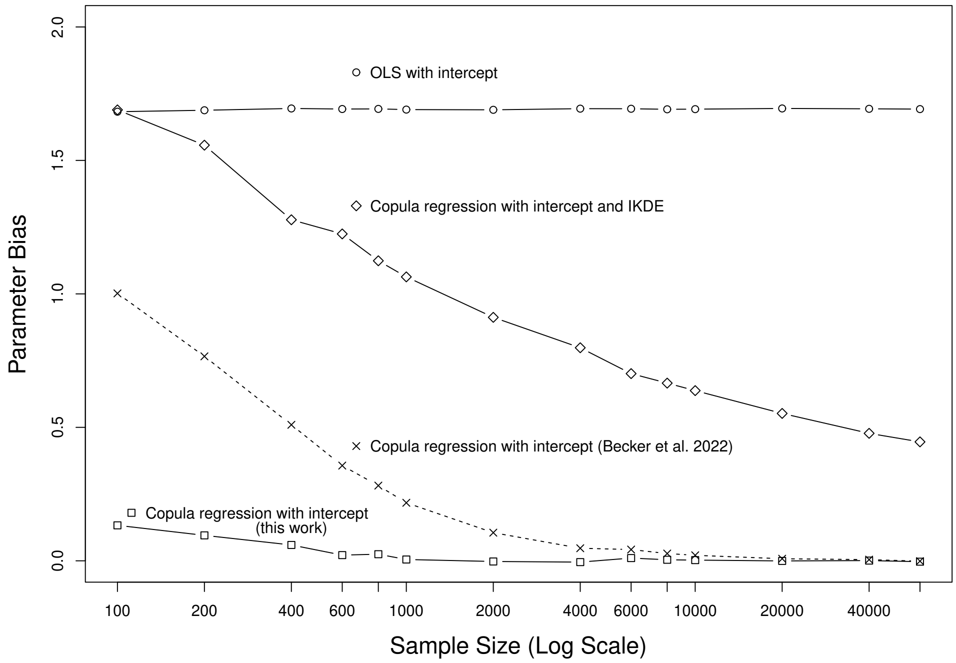

Figure W1: Bias of the endogenous regressor.

Figure W1 shows the bias of α, evaluated as the difference between the mean parameter estimate averaged over 1,000 simulated data sets and its true value, for different estimation methods at sample sizes ranging from 100 to 60,000 (Figure W1 x-axis). OLS, as the curve with circles in Figure W1, exhibits substantial bias (> 1.5) in the coefficient estimate α for endogenous regressor P. Furthermore, this bias remains the same regardless of sample size. Consistent with Becker, Proksch, and Ringle (2022), the P&G method using Equation 12 (the curve with cross marks in Figure W1) substantially reduces the bias in the OLS estimates, but does not resolve the endogeneity in many situations: substantial bias remains after copula correction in small to moderate sample sizes. The endogenous regressor's coefficient estimation bias only becomes negligible for sample sizes larger than 4,000. The finite sample bias for P&G copula regression with intercept discovered in Becker, Proksch, and Ringle (2022) is a significant problem that needs addressing, so as to ensure appropriate use of copula correction. This is relevant because prior to Becker, Proksch, and Ringle (2022), users of copula correction were unaware of such surprisingly severe bias concerns.

A key finding in Figure W1 is that the substantial bias of the P&G copula correction method for models with intercept, discovered in Becker, Proksch, and Ringle (2022), is largely solved by adjusting the largest rank using Equation 11. The algorithm in Equation 11 results in considerably improved performance of the P&G copula correction method; the endogenous regressor's coefficient estimate bias now becomes negligible when sample size reaches 400 rather than 4,000 (the curve with squares in Figure W1). Furthermore, even sample sizes as small as 100 exhibit a bias of about 0.15 for our algorithm[31], which is quite smaller than 1.0 using the algorithm in Equation 12. The theoretical reason is that constructing the empirical copula using the fixed-value percentile for the largest rank can substantially distort the distribution of generated regressor P^*, resulting in suboptimal performance of the PG copula correction method and substantial bias in small to moderate samples. In conclusion, including an intercept in the model does not cause concern as long as the last-ranked value of the empirical CDF is properly handled by using the recommended copula transformation algorithm.

[31] This is perhaps unsurprising because the copula correction method, like instrumental variables and other IV-free methods, is a large sample procedure requiring sufficient information for satisfactory performance.

Comparison with Integrating Nonparametric Kernel Density Estimation

This subsection aims to examine whether the bias problem discovered in Becker, Proksch, and Ringle (2022) can be resolved by employing the approach of integrating nonparametric kernel density estimation (IKDE) to obtain the copula correction term (Park and Gupta 2012). The IKDE method first estimates the marginal density function $f_P(p)$ of the continuous regressor P using the following Epanechnikov kernel nonparametric method

$$\widehat{f}_P(P = p) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{p - P_i}{b}\right), \quad \text{(W29)}$$

where $K(\cdot)$ is the user-supplied kernel function and b is the bandwidth parameter that exerts a strong influence on the density estimation. The optimal bandwidth value is unknown but there are some suggestions for choosing the bandwidth. When using the Epanechnikov kernel $K(x) = 0.75(1 - x^2)I(|X| \leq 1)$, the rule-of-thumb for determining the bandwidth is $b = 0.9n^{-1/5}min(s, IQR/1.34)$, where s is the sample standard deviation and IQR is the interquartile range. The IKDE approach then integrates the marginal density function estimate to obtain the marginal CDF as follows:

$$\widehat{F}_P(p) = \int_{-\infty}^{p} \widehat{f}_P(u)du, \quad \text{(W30)}$$

where the trapezoidal rule can be used for the above numerical integration (Park and Gupta 2012).

It is unclear if the IKDE approach to obtaining the copula correction terms outperforms the approach of using empirical CDF. On the one hand, the IKDE approach does not encounter the problem of the last observation having infinite value of copula latent data as

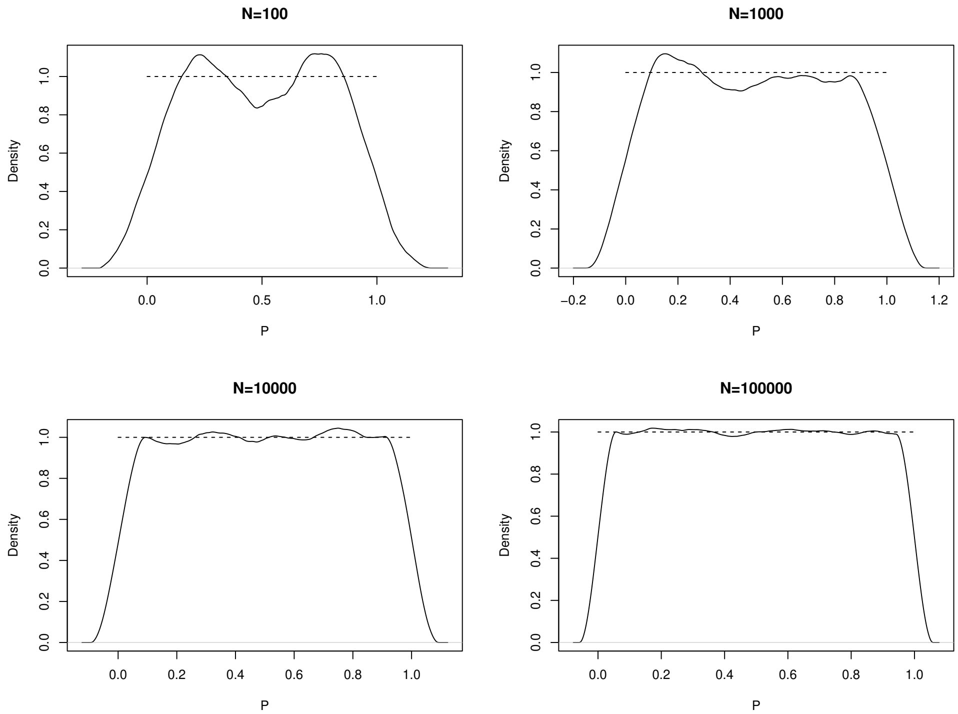

Figure W2: Boundary Bias of Nonparametric Kernel Density Estimates. Dotted line denotes the true density function of the uniform distribution on [0, 1]. Solid line denotes the KDE estimates.

empirical CDF encounters. On the other hand, the nonparametric KDE methods are subject to boundary bias (e.g., Cid and von Davier 2015, Karunamuni and Alberts 2005), which is an important drawback of KDE density estimation. The boundary bias of KDE estimation is particularly severe for variables with bounded support or for density estimation near the boundaries of the support of the density to be estimated (Karunamuni and Alberts 2005). Large sample size is required to control or mitigate the boundary bias. Figure W2 illustrates boundary bias of kernel density estimation in four simulated datasets at sample size ranging from N=100 to N=100,000 when the true density function is the uniform distribution on [0,1]. We observe density estimation bias near the two ends of the uniform distribution, although the boundary bias decreases with increasing sample size.

Returning to Figure W1, consider the estimation bias of using IKDE for copula correction with the same DGP as specified in Equations 1-4 in Becker, Proksch, and Ringle (2022) (i.e., Equations W25 to W28). We implemented the IKDE approach using the R function `density(P, kernel="epanechnikov")` for nonparametric kernel density estimation and the R function `CDF()` that integrates the KDE estimates to the cumulative distribution function using the trapezoidal rule. Figure W1 shows that copula correction using the IKDE approach has larger bias across all sample sizes than the approaches using the ECDF. This can arise from the severe boundary bias (Figure W2) of KDE for estimating the density near the boundaries of the support. By contrast, the ECDF can automatically account for the bounded support of the uniform distributions and avoid such severe boundary bias.

Models Without Intercept

Figure W3 plots the estimation results when estimating the model in Equation W28 without intercept. All settings remain the same as those when estimating the models with unknown intercept, except that the estimation now assumes the intercept parameter μ is known a-priori and consequently we estimate all the other model parameters given the a-priori known intercept value. The difference between the average of the estimates across 1,000 simulated datasets and its true value is the bias of an estimator, which is plotted in Figure W3 for α. Results in Figure W3 show large OLS estimation bias that remains constant across all sample sizes. Interestingly, in this case, there is no bias at any sample size for all algorithms to generate copula transformation (IKDE, fixed ECDF, or adaptive ECDF). This means that unlike the case of estimating models with intercept, choice of algorithms for handling the infinite value of copula transformation of the last-rank observation does

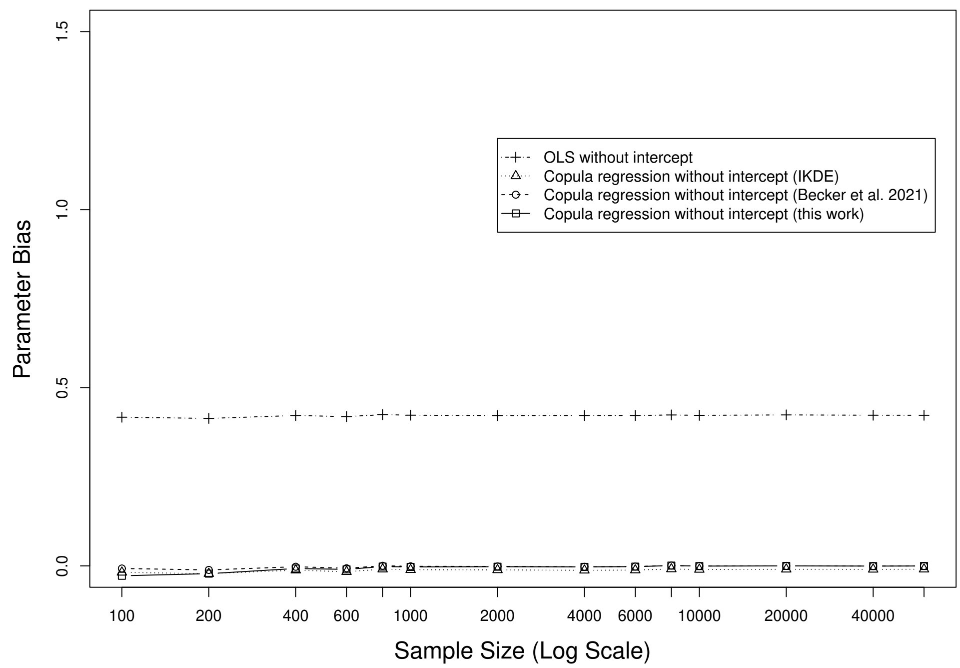

Figure W3: Bias of the endogenous regressor without intercept.

not matter, and all three algorithms work well to correct OLS estimation bias across all considered sample sizes.

Copula Transformation with Correlated Regressors

In this section, we assess the impact of copula transformation on the 2sCOPE procedure. The DGP is summarized below:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ E_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix}\right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}\right), \quad \text{(W31)}$$

$$E_t = G^{-1}(U_{E,t}) = G^{-1}(\Phi(E_t^*)) = \Phi^{-1}(\Phi(E^*)) = 1 \cdot E_t^*, \quad \text{(W32)}$$

$$P_t = H^{-1}(U_{P,t}) = H^{-1}(\Phi(P_t^*)), \quad W_t = L^{-1}(U_{W,t}) = L^{-1}(\Phi(W_t^*)), \quad \text{(W33)}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + E_t = 0 + (-1) \cdot P_t + 1 \cdot W_t + E_t, \tag{W34}$$

where E_t^* and P_t^* are correlated ($\rho_{pe} = 0.5$), generating the endogeneity problem; W_t^* is exogenous and uncorrelated with E_t^*; W_t^* and P_t^* are correlated ($\rho_{pw} = 0.5$), and therefore W_t and P_t are correlated, which calls for the use of 2sCOPE. We consider the following estimation methods: (1) OLS regression of Equation (W34); (2) 2sCOPE using the fixed algorithm for copula transformation of P and W Equation 12; (3) 2sCOPE using the adaptive algorithm for copula transformation of P and W (Equation 11), and (4) 2sCOPE-np using the nonparametric copula control function $\Phi^{-1}(F(P_i|W_i))$ in Equation 4. In the simulation, we use the uniform distribution on [0,1] for P_t and the exponential distribution $Exp(1)$ with rate 1 for W_t. Models are estimated on all generated datasets, providing the empirical distributions of parameter estimates.

We use the procedures described in Table 6 for 2sCOPE and 2sCOPE-np. For 2sCOPE-np, we use the following Nadaraya-Watson (NW) nonparametric kernel regression procedures to estimate the conditional CDF $F(P_i|W_i)$ with the following locally weighted average (Li and Racine 2008):

$$\widehat{F}^a(p|w) \quad = \quad \frac{\sum_{i=1}^n I(P_i \leq p) K_h(W_i - w)}{\sum_{i=1}^n K_h(W_i - w)}. \tag{W35}$$

where $I(P_i \leq p)$ is the indicator function for the event $P_i \leq p$; n denotes sample size; $K_h(W_i - w)$ is a weight function defined as:

$$K_h(W_i - w) \quad = \quad \frac{1}{h} k\left(\frac{W_i - w}{h}\right). \tag{W36}$$

where $k(\cdot)$ is a user-supplied smooth and symmetric kernel function and h is the bandwidth

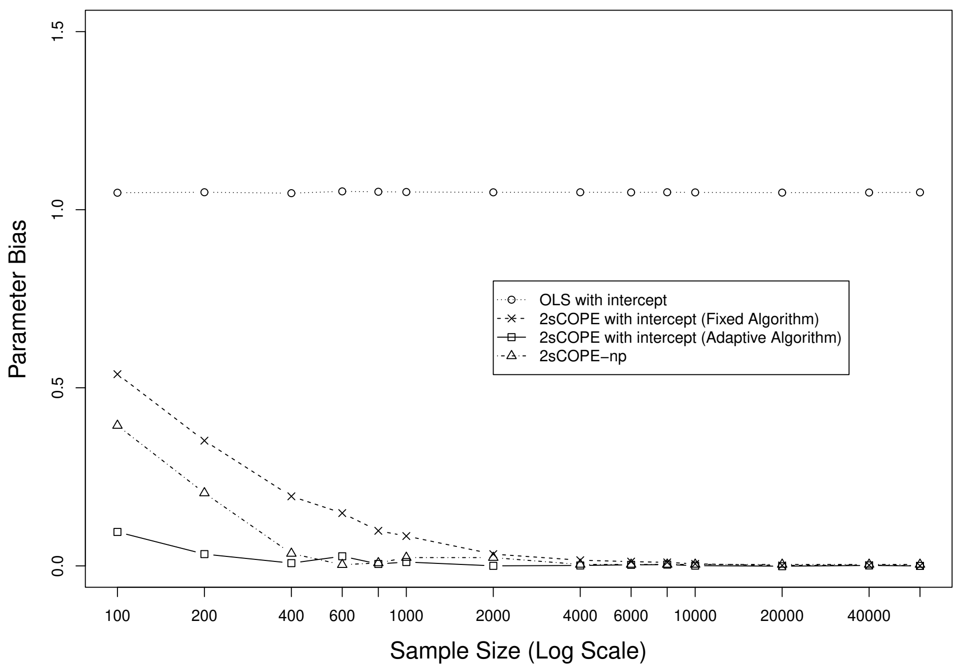

Figure W4: Method comparison with correlated endogenous and exogenous regressors

parameter. Another estimator smooths the continuous outcome Y as follows:

$$\widehat{F}^b(p|w) \quad = \quad \frac{\sum_{i=1}^n G((p - P_i)/h_0) K_h(W_i - w)}{\sum_{i=1}^n K_h(W_i - w)}, \tag{W37}$$

where $G(\cdot)$ is the CDF function defined by $G(v) = \int_{-\infty}^{v} k(u)du$ from the density function $k(u)$, and h_0 is the bandwidth for smoothing the outcome Y. See Hu, Qian, and Xie (2025) for more details about the description and implementation of these kernel conditional CDF estimators in 2sCOPE-np.

Figure W4 shows that 2sCOPE using the fixed algorithm also negatively affects the performance of copula correction, while 2sCOPE using the adaptive algorithm avoids the bias. Unlike 2sCOPE, 2sCOPE-np does not perform copula transformations directly on regressors, but instead on the smoothed conditional CDF estimate of $F(P|W)$, which takes values less than one because of smoothing. Thus, fixed algorithm or adaptive algorithm is irrelevant for 2sCOPE-np. As shown in Figure W4, the 2sCOPE using adaptive algorithm performs best with negligible bias even at the relatively small sample size $n = 100$. As expected for a nonparametric procedure, 2sCOPE-np (the curve with triangles) performs well when sample size is sufficiently large even if it does not impose any model on regressors, but it does require a larger sample size to have negligible finite sample bias than the correctly specified 2sCOPE using the adaptive algorithm (the curve with squares). Figure W4 suggests a minimum sample size of 300 for 2sCOPE-np to have negligible finite sample estimation bias (i.e., Boundary Condition 3 in Table 4). This is consistent with that the the empirical applications of the nonparametric kernel CDF estimation in Li and Racine (2008) all have a minimum sample size of 300.

WEB APPENDIX E: PROOF OF OPTIMALITY OF EXCLUDING HIGHER-ORDER COPULA TERMS.

Theorem 1. ***Optimality of excluding higher-order copula terms.*** *Let* $(\widehat{\theta}_k^{Main}), k = 1, \cdots, K$, *denote the structural model parameter estimates when only the copula terms for the main endogenous effects are included to correct for endogeneity, and* $(\widehat{\theta}_k^{All}), k = 1, \cdots, K$, *denote the corresponding estimates when copula terms for both the main effects and higher-order endogenous regressors are included. This yields:*

$$\boldsymbol{Var}(\widehat{\theta}_k^{All}) \geq \boldsymbol{Var}(\widehat{\theta}_k^{Main}) \quad \text{for } k = 1, \cdots, K.$$

Thus, $\widehat{\theta}_k^{Main}$ *yields optimal copula estimation of structural model parameters with less variance and mean squared errors than* $\widehat{\theta}_k^{All}$, *for all* k.

Proof: Consider the OLS regression of the model when only the copula main terms are included to correct for endogeneity:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad V(\boldsymbol{\epsilon}) = \sigma_c^2 \mathbf{I}_n, \tag{W38}$$

where $\mathbf{X}$ includes the intercept, the regressors in the structural model, and $\mathbf{C}_{main}$ (the copula generated regressors for the main effects); $\boldsymbol{\theta}$ collects all the coefficients of these regressors. Math symbols in bold represent matrices and vectors. The variance of the estimates using copula terms for main effects only is:

$$V(\widehat{\boldsymbol{\theta}}^{Main}) = \sigma_c^2 (\mathbf{X}'\mathbf{X})^{-1}. \tag{W39}$$

Then after introducing additional copula terms $\mathbf{C}$ for higher-order terms into the model in

Equation (W38), we have:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{C}\boldsymbol{\phi} + \boldsymbol{\epsilon_1}, \quad V(\boldsymbol{\epsilon_1}) = \sigma_c^{\prime 2}\mathbf{I}_n, \tag{W40}$$

According to linear regression theory, the new estimates after entering the copula higher-order terms $\mathbf{C}$ in the model become:

$$\widehat{\boldsymbol{\theta}}^{All} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{C}\widehat{\phi}), \qquad \widehat{\phi} = (\mathbf{C}'\mathbf{R}\mathbf{C})^{-1}\mathbf{C}'\mathbf{R}\mathbf{Y}, \tag{W41}$$

$$V(\widehat{\boldsymbol{\theta}}^{All}) = \sigma_c^{\prime 2}\big[(\mathbf{X}'\mathbf{X})^{\mathbf{-1}} + \mathbf{M}(\mathbf{C}'\mathbf{R}\mathbf{C})^{\mathbf{-1}}\mathbf{M}'\big], \tag{W42}$$

where $\mathbf{M} = (\mathbf{X}'\mathbf{X})^{\mathbf{-1}}\mathbf{X}'\mathbf{C}$, $\mathbf{R} = \mathbf{I_n} - \mathbf{P}$, and $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{\mathbf{-1}}\mathbf{X}'$. Note that $\mathbf{P}$ is the projection matrix representing the orthogonal projection that maps the responses to the fitted values, and $\mathbf{R} = \mathbf{I_n} - \mathbf{P}$ represents the orthogonal projection that maps the responses to the residuals. Given that the newly added higher-order copula terms in $\mathbf{C}$ are highly correlated with the higher-order terms in the structural model (as well as other copula terms already included in the model), the extra variability in $\mathbf{Y}$ explained by adding $\mathbf{C}$ is small. Thus, $\sigma_c^{\prime 2} \approx \sigma_c^2$ and:

$$V(\widehat{\boldsymbol{\theta}})^{All} - V(\widehat{\boldsymbol{\theta}})^{Main} \approx \sigma_c^2\big[(\mathbf{X}'\mathbf{X})^{\mathbf{-1}} + \mathbf{M}(\mathbf{C}'\mathbf{R}\mathbf{C})^{\mathbf{-1}}\mathbf{M}' - (\mathbf{X}'\mathbf{X})^{\mathbf{-1}}\big] \tag{W43}$$

$$= \sigma_c^2\big[\mathbf{M}(\mathbf{C}'\mathbf{R}\mathbf{C})^{\mathbf{-1}}\mathbf{M}'\big]. \tag{W44}$$

Since the matrix $\mathbf{M}(\mathbf{C}'\mathbf{R}\mathbf{C})^{\mathbf{-1}}\mathbf{M}'$ is positive semi-definite, all the diagonal elements are greater than or equal to zero. For each of the K structural model parameters:

$$\mathtt{Var}(\widehat{\theta}_k^{All}) \geq \mathtt{Var}(\widehat{\theta}_k^{Main}) \quad \text{for } k = 1, \cdots, K. \tag{W45}$$

The magnitude of variance inflation is inversely related to $\mathbf{C}'\mathbf{R}\mathbf{C}$, which represents the matrix of sum of squared residuals, obtained from regressing $\mathbf{C}$ on $\mathbf{X}$. Thus, the higher the correlation between the extra higher-order term $\mathbf{C}$ and existing regressors in $\mathbf{X}$, the smaller the sum of squares, which leads to greater variance inflation of $\texttt{Var}(\widehat{\theta}_k^{All})$. Q.E.D.

WEB APPENDIX F: SIMULATION STUDIES ILLUSTRATING THE HARMFUL EFFECTS OF INCLUDING HIGHER-ORDER COPULA TERMS

The theoretical proof in the preceding section shows that copula terms for higher-order effects are not only unnecessary, but also substantially inflate estimation variability: the higher the correlations between the extra higher-order copula term and other regressors, the greater the estimation variance inflation. The empirical application of peanut butter sales in the main text further demonstrates this adverse bias: omitting the higher-order copula term yields model estimates closest to that of two-stage least squares using instrumental variables; including the copula interaction term produces the opposite sign for the coefficient estimate of the endogenous interaction term, and greater estimation variability.

In addition to the above theoretical results and real data analysis, this section presents empirical evidences using simulated data to demonstrate (1) that there is no need to add correction terms for higher-order terms of endogenous regressors to control for their endogeneity, and more importantly, (2) harmful effects occur if correction terms for higher-order terms are added to control for their endogeneity. These effects include potential finite sample bias and inflated variability of structural model parameter estimates, as predicted by the theoretical results in the previous section. The simulation study below highlights the magnitude of such harmful effects: larger standard errors (by up to 5-times as shown in our simulation studies), substantial estimation bias (about 30% of parameter values), and significant loss of statistical power to detect moderating and nonlinear effects (e.g., a reduction of power from 80% to 10% in Figure W7, much further below).

Case I: Interaction Between Two Endogenous Regressors

Data were simulated from the following structural regression model with an interaction between two endogenous regressors, P_1 and P_2:

$$
\begin{aligned}
Y &= \alpha_0 + \alpha_1 P_1 + \alpha_2 P_2 + \alpha_3 P_1 * P2 + E \qquad \text{(W46)}\\
\begin{pmatrix} E^* \\ P_1^* \\ P_2^* \end{pmatrix} &= N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho_{E1} & \rho_{E2} \\ \rho_{E1} & 1 & \rho_{12} \\ \rho_{E2} & \rho_{12} & 1 \end{bmatrix}\right)\\
E = H_E^{-1}(\Phi(E^*)) &= \Phi^{-1}(\Phi(E^*)), \qquad P_1 = H_{P_1}^{-1}(\Phi(P_1^*)), \quad P_2 = H_{P_2}^{-1}(\Phi(P_2^*)). \text{(W47)}
\end{aligned}
$$

In this simulation, we set $H_{P_1}(\cdot)$ as the CDF of the uniform distribution on $[4, 6]$, $H_{P_2}(\cdot)$ as the CDF of the truncated standard normal with a lower bound of 0, and parameters $\alpha_0 = 0, \alpha_1 = 1, \alpha_2 = -1, \alpha_3 = 1, \rho_{E1} = \rho_{E2} = 0.5, \rho_{12} = -0.5$. For each simulated data set, the following three estimation procedures were applied regressing Y on the following sets of regressors:

OLS:	P_1, P_2
Copula-Main:	$P_1, P_2, C_{P_1}, C_{P_2}$
Copula-All:	$P_1, P_2, C_{P_1}, C_{P_2}, C_{P_1 * P_2}$

where $C_{P_1} = \Phi^{-1}(\widehat{F}_{P_1}(P_1))$, $C_{P_2} = \Phi^{-1}(\widehat{F}_{P_2}(P_2))$, and $C_{P_1 * P_2} = \Phi^{-1}(\widehat{F}_{P_1 * P_2}(P_1 * P_2))$ are the copula correction terms. That is, we use the P&G method for copula correction since the model contains no exogenous regressors. The OLS estimation regresses Y on P_1, P_2 and $P_1 * P_2$ without any correction for the endogeneity of these regressors. Copula-Main adds two copula correction terms, C_{P_1} and C_{P_2}, to control for the endogeneity of these three regressors, where:

$$C_{P_1} = \Phi^{-1}(\hat{H}_{P_1}(P_1)), \qquad C_{P_2} = \Phi^{-1}(\hat{H}_{P_2}(P_2)). \tag{W48}$$

In addition to C_{P_1} and C_{P_2}, Copula-All adds the copula correction term $C_{P_1*P_2}$, where:

$$C_{P_1*P_2} \quad = \quad \Phi^{-1}(\hat{H}_{P_1*P_2}(P_1 * P_2)) \tag{W49}$$

and $\hat{H}_{P_1}$, $\hat{H}_{P_2}$ and $\hat{H}_{P_1*P_2}$ denote the empirical marginal distribution functions of P_1, P_2 and $P_1 * P_2$ in the observed sample, respectively.

Bias and SEs of parameter estimates Across simulations, sample sizes (N) of 200, 500, 5,000, and 50,000 are examined. For each sample size N, we generate 5,000 data sets as replicates to systematically evaluate average performance (estimation bias and variability) for the three estimation methods. The simulation results appear in Table W10. As expected, OLS regression yields significant bias for all model parameters at all sample sizes. For example, even for a large sample size of N=5,000, the OLS regression without any correction terms yields large bias for the regression parameter estimates ($\widehat{\alpha}_1$: 2.281 [0.018]; $\widehat{\alpha}_2$: -1.549 [0.099]; $\widehat{\alpha}_3$: 1.432 [0.021]) and the error standard deviation ($\widehat{\sigma}$: 0.298 [0.006]). Copula-Main corrects for the endogenous bias ($\widehat{\alpha}_1$: 1.002 [0.058]; $\widehat{\alpha}_2$: -1.017 [0.080]; $\widehat{\alpha}_3$: 1.003 [0.015]), demonstrating that there is no need to additionally include the copula correction term, $C_{P_1*P_2}$. Furthermore, Copula-Main performs substantially better in both estimation bias and variability for all parameter estimates than Copula-All which includes $C_{P_1*P_2}$. In fact, Copula-All yields significantly biased parameter estimates, even at the large sample size of N=5,000 ($\widehat{\alpha}_0$: 0.202 [0.318]; $\widehat{\alpha}_2$: -0.713 [0.240]; $\widehat{\alpha}_3$: 0.929 [0.058]); bias decreases as sample size increases, but remains apparent even for a sample size as large as 50,000, as including the copula term for the interaction $P_1 * P_2$ causes significant estimation bias.

The same conclusion - that Copula-Main performs substantially better than Copula-All in terms of both estimation bias and variability for all parameter estimates - applies to all other sample sizes, except for the intercept parameter (α_0) at small sample size N=200. The exception likely results from both a small sample size and strong multicollinearity induced by the interaction term; however, the bias in the intercept estimate bears less practical implication, since the intercept parameter is often of less interest.

Copula-All also yields less precise estimates (larger standard errors) than Copula-Main; underlined standard errors in Table W10 highlight much larger SE for Copula-All versus Copula-Main. This imprecision includes an SE 3.00-times that for α_2 and 3.86-times that for α_3 compared to Copula-Main at a sample size of 5,000.

Overall Estimation Efficiency and Accuracy We further compare the efficiency of Copula-Main and Copula-All using the D-error measure (Arora and Huber 2001, Qian and Xie 2022). The D-error measure is defined as $|\Sigma|^{1/K}$ where Σ is the variance-covariance matrix of the regression coefficient estimates, and K is the number of explanatory variables in the structural regression model. A larger D-error value means lower efficiency, with a $\Delta\%$ increase in D-error corresponding to a $\Delta\%$ larger sample size required to achieve the same level of estimation precision. As shown in Table W10, the D-error inflation for Copula-All is about 3-times at N=5,000. In this case, Copula-All requires about 3-times the sample size in order to achieve approximately the same accuracy for estimating α_1, α_2 and α_3 jointly as Copula-Main. The variance inflation for the Copula-All estimate of α_3, the coefficient for the

Table W10: Results from Case I: Interaction of Endogenous Regressors.

N	Method	$\alpha_0(=0)$	$\alpha_1(=1)$	$\alpha_2(=-1)$	$\alpha_3(=1)$	$\sigma(=1)$	D-error
200	OLS	**-7.627**	**2.282**	**-1.546**	**1.433**	**0.294**	
		(0.464)	(0.093)	(0.501)	(0.106)	(0.031)	—
	Copula-Main	**-0.358**	1.046	**-1.187**	1.043	0.963	
		(1.363)	(0.271)	(0.417)	(0.079)	(0.121)	0.0293
	Copula-All	**-0.058**	1.012	**-0.794**	**0.930**	1.028	
		(1.364)	(0.270)	(0.468)	(0.107)	(0.134)	0.0368
500	OLS	**-7.624**	**2.281**	**-1.546**	**1.432**	**0.297**	
		(0.290)	(0.058)	(0.312)	(0.066)	(0.019)	—
	Copula-Main	**-0.119**	1.019	**-1.104**	1.024	0.99	
		(0.899)	(0.179)	(0.254)	(0.047)	(0.076)	0.0117
	Copula-All	**0.176**	0.974	**-0.702**	**0.923**	1.051	
		(0.902)	(0.178)	(0.331)	(0.077)	(0.086)	0.0165
5,000	OLS	**-7.623**	**2.281**	**-1.549**	**1.432**	**0.298**	
		(0.092)	(0.018)	(0.099)	(0.021)	(0.006)	—
	Copula-Main	-0.012	1.002	-1.017	1.003	1.000	
		(0.291)	(0.058)	(0.080)	(0.015)	(0.024)	0.0011
	Copula-All	**0.202**	0.968	**-0.713**	**0.929**	1.044	
		(0.318)	(0.061)	(0.240)	(0.058)	(0.041)	0.0031
50,000	OLS	**-7.621**	**2.281**	**-1.551**	**1.433**	**0.298**	
		(0.029)	(0.006)	(0.031)	(0.007)	(0.002)	—
	Copula-Main	0.001	1.000	-1.003	1.000	1.000	
		(0.092)	(0.018)	(0.025)	(0.005)	(0.008)	0.00011
	Copula-All	**0.064**	0.990	**-0.912**	0.978	1.013	
		(0.133)	(0.023)	(0.158)	(0.038)	(0.023)	0.00051

Table presents the averages of the estimates and standard errors in the parenthesis over the repeated samples. Bold numbers highlight the estimates with bias of at least 0.05. Underlined numbers highlight the cases where the standard errors of the estimates from Copula-All are inflated by at least 50% compared with the corresponding ones from Copula-Main. The P&G method is used for copula correction since the model contains no exogenous regressors.

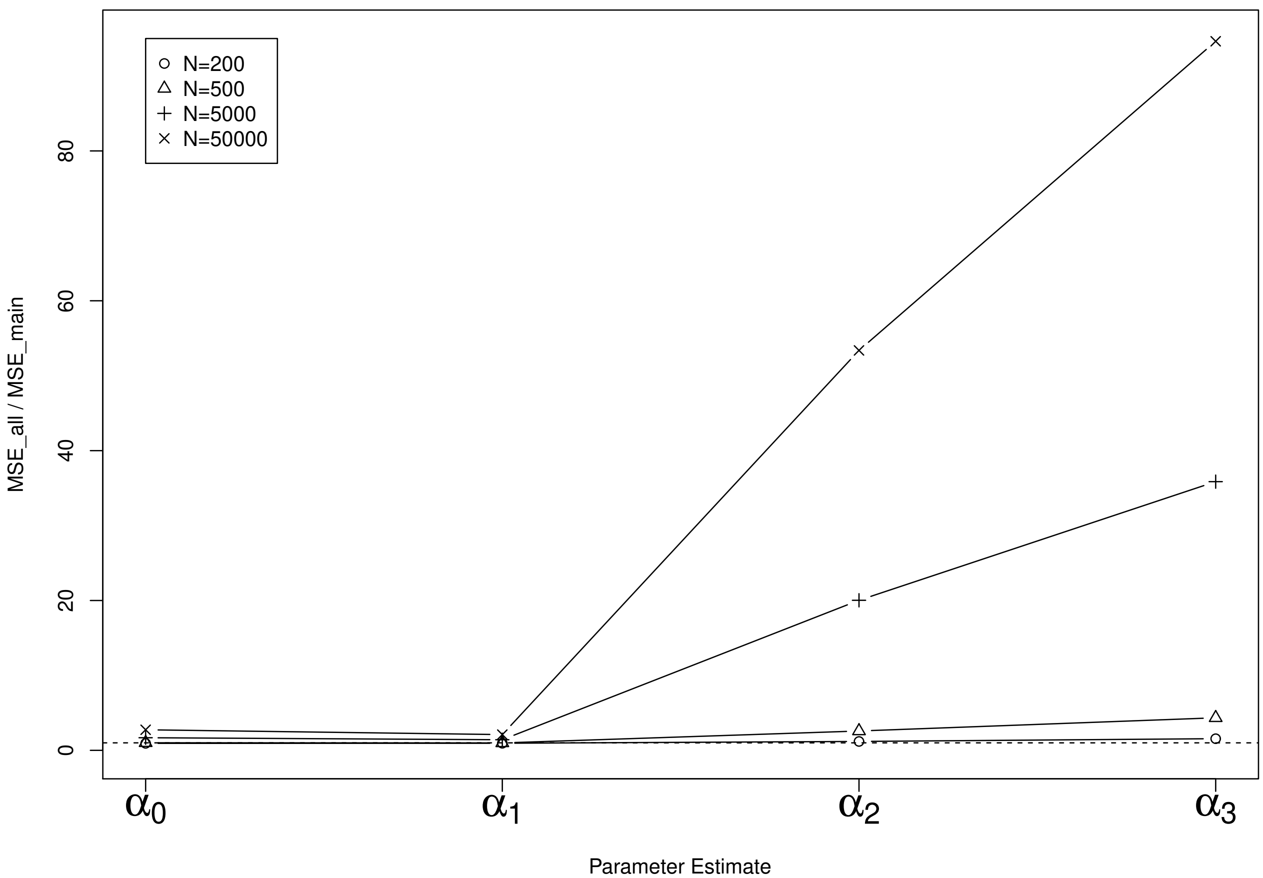

Figure W5: Ratio of mean squared errors of structural model estimates, with using the copula interaction term (Copula-All) to those without using the copula interaction term (Copula-Main).

interaction term, is much larger and equals $(\frac{0.058}{0.015})^2 \approx 15$ when N=5,000. This means 15-times the sample size is required for Copula-All to achieve the same estimation accuracy of the interaction term as Copula-Main. Regarding overall estimation efficiency, the D-error ratios for Copula-All to Copula-Main increase as sample size increases, from 1.26-times (N=200) to 1.41-times (N=500) to 2.82-times (N=5,000) to 4.64-times (N=50,000).

We also compute the ratio of mean squared error (MSE) of the structural estimate $\widehat{\alpha}_k$, comparing Copula-All to Copula-Main (where $\text{MSE}(\widehat{\alpha}_k) = \text{Bias}^2(\widehat{\alpha}_k)+\text{Var}(\widehat{\alpha}_k)$, measuring overall estimation accuracy). Notably, Copula-All increases MSEs for all model parameter estimates, with the harmful effects being largest for the interaction parameter estimate $\widehat{\alpha}_3$, whose MSE is more than 80-times that of Copula-Main when sample size N=50,000 (Figure W5).

Case II: Interaction Between an Endogenous Regressor and an Exogenous Regressor

We simulated data from the following structural regression model with an interaction term between an exogenous regressor X and an endogenous regressor P:

$$\begin{aligned} Y &= \alpha_0 + \beta_1 W + \alpha_1 P + \alpha_2 W * P + E \\ \begin{pmatrix} P^* \\ W^* \\ E^* \end{pmatrix} &= N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix} \right) \\ E = H_E^{-1}(\Phi(E^*)) &= \Phi^{-1}(\Phi(E^*)), \qquad P = H_P^{-1}(\Phi(P^*)), W = L_W^{-1}(\Phi(W^*)) \end{aligned} \tag{W50}$$

where $H_P(\cdot)$ is the CDF of the truncated standard normal on $[0, \infty]$, and $L_W(\cdot)$ is the CDF of a uniform distribution on $[4, 6]$, and we set $\alpha_0 = 0, \beta_1 = 1, \alpha_1 = -1, \alpha_2 = 1$ and $\rho_{pe} = 0.5, \rho_{pw} = -0.5$ with sample sizes of 200, 500, 5,000, and 50,000. For each sample size, we generated 5,000 repeated samples.

For each generated sample, we then apply three estimation procedures: OLS, 2sCOPE-Main and 2sCOPE-All. 2sCOPE is used to handle correlated regressors P and W. The OLS regresses Y on P, W and $W * P$ without any correction for the endogeneity of P and $W * P$. 2sCOPE-Main adds one copula correction term, $C_P = P^* - \hat{\delta}_1 W^*$ (Equation 14) to control for endogeneity of P and $W * P$, where P^* and W^* are copula transformations of P and W using the ECDFs $\hat{H}_P(\cdot)$ and $\hat{L}_W(\cdot)$ estimated from data, respectively. In addition to C_P, 2sCOPE-All adds the copula correction term $C_{W*P} = (W * P)^* - \hat{\delta}_2 W^*$, where $(W * P)^*$ is a copula transformation of the interaction term $W * P$ using its ECDF $\hat{H}_{W*P}(\cdot)$ estimated from data. $\hat{H}_P(\cdot)$, $\hat{L}_W(\cdot)$, and $\hat{H}_{W*P}(\cdot)$ denote the empirical marginal distribution functions of P, W, and $W * P$ in the observed sample, respectively. Results over 5,000 simulated samples are summarized in Table W11.

As expected, the OLS regression without any correction terms yields large bias for the regression parameter estimates and the error standard deviation σ in the structural regression model. 2sCOPE-Main corrects for the endogenous bias, demonstrating that there is no need to additionally include the correction term for the interaction term of P and W. Importantly, 2sCOPE-All, which adds the unnecessary copula correction term for the interaction term, yields less precise estimates (larger standard error of estimates as shown in Table W11) than 2sCOPE-Main, increasing the D-error by more than 100% in some cases. Furthermore, significant estimation bias in parameter estimates for α_1 exists for 2sCOPE-All which decrease as sample size increases, but still remains for a sample size as large as 50,000 (Table W11). The results demonstrate the substantial adverse effects of adding unnecessary copula terms for interactions: significant finite sample estimation bias and inflated standard errors.

Table W11: Results from Case II: Interaction between Endogenous and Exogenous Regressors

N	Method	$\alpha_0(=0)$	$\beta_1(=1)$	$\alpha_1(=-1)$	$\alpha_2(=1)$	$\sigma(=1)$	D-error
200	OLS	**-2.388**	**1.312**	**-1.281**	**1.274**	**0.829**	
		(0.902)	(0.174)	(0.876)	(0.182)	(0.041)	—
	2sCOPE-Main	**-0.126**	1.020	-1.047	1.026	0.987	
		(1.342)	(0.223)	(0.884)	(0.208)	(0.127)	0.0425
	2sCOPE-All	**-0.141**	1.028	**-0.796**	0.964	1.016	
		(1.371)	(0.229)	(1.305)	(0.315)	(0.152)	0.0651
500	OLS	**-2.351**	**1.306**	**-1.302**	**1.278**	**0.832**	
		(0.561)	(0.109)	(0.549)	(0.115)	(0.026)	—
	2sCOPE-Main	-0.013	1.000	-1.039	1.014	0.997	
		(0.842)	(0.140)	(0.543)	(0.126)	(0.083)	0.0159
	2sCOPE-All	**-0.052**	1.013	**-0.791**	0.946	1.024	
		(0.855)	(0.144)	(0.905)	(0.232)	(0.110)	0.0298
5,000	OLS	**-2.338**	**1.303**	**-1.312**	**1.280**	**0.833**	
		(0.179)	(0.034)	(0.169)	(0.035)	(0.008)	—
	2sCOPE-Main	0.018	0.997	-1.009	1.003	1.001	
		(0.242)	(0.045)	(0.165)	(0.036)	(0.025)	0.0016
	2sCOPE-All	0.025	1.002	**-0.896**	0.970	1.009	
		(0.272)	(0.057)	(0.469)	(0.112)	(0.041)	0.0039
50,000	OLS	**-2.350**	**1.305**	**-1.298**	**1.277**	**0.833**	
		(0.056)	(0.011)	(0.054)	(0.011)	(0.003)	—
	2sCOPE-Main	0.000	1.000	-1.000	1.000	1.000	
		(0.070)	(0.011)	(0.055)	(0.013)	(0.008)	0.0002
	2sCOPE-All	-0.002	1.001	**-0.948**	0.991	1.002	
		(0.083)	(0.017)	(0.166)	(0.042)	(0.014)	0.0004

Table presents the averages of the estimates and standard errors in the parenthesis over the repeated samples. Bold numbers highlight the estimates with bias of at least 0.05. Underlined numbers highlight the cases where the standard errors of the estimates from 2sCOPE-All are inflated by at least 50% compared with the corresponding ones from 2sCOPE-Main.

Case III: A Squared Term of an Endogenous Regressor

Data were simulated from the following model:

$$\begin{aligned} Y &= \alpha_0 + \alpha_1 P + \alpha_2 P^2 + E, \\ \begin{pmatrix} E^* \\ P^* \end{pmatrix} &= N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \\ E = H_E^{-1}(\Phi(E^*)) &= \Phi^{-1}(\Phi(E^*)), \qquad P = H_P^{-1}(\Phi(P^*)), \end{aligned} \tag{W51}$$

where $H_P(\cdot)$ is the CDF for the marginal distribution of P, $\alpha_0 = 0, \alpha_1 = -1, \alpha_2 = 1$ and $\rho = 0.7$. We set $H_P(\cdot)$ as the CDF of the truncated standard normal distribution on $[-0.5, 0.5]$. For each simulated data set, the following three estimation procedures were applied using OLS regression of Y on the following sets of regressors:

$$\begin{aligned} \text{OLS:} &\qquad P, P^2 \\ \text{Copula-Main:} &\qquad P, P^2, C_P \\ \text{Copula-All:} &\qquad P, P^2, C_P, C_{P^2} \end{aligned}$$

where $C_P = \Phi^{-1}(\hat{H}_P(P))$ and $C_{P^2} = \Phi^{-1}(\hat{H}_{P^2}(P^2))$ are the copula correction terms for endogenous regressors P and P^2, respectively; $\hat{H}_P$ and $\hat{H}_{P^2}$ denote the empirical marginal distribution functions of P and P^2 in the generated sample, respectively. Copula-Main indicates including copula correction terms for the main effect only, while Copula-All signifies including copula correction for all terms involving endogenous regressor P (i.e., higher-order terms). That is, we use the P&G method for copula correction since the model contains no exogenous regressors.

Across simulations, sample sizes (N) of 200, 500, 5,000, and 50,000 are examined. For each sample size N, we generate 5,000 data sets as replicates to systematically evaluate average performance (estimation bias and variability) of different estimation methods. Averages and standard deviations (SD) of parameter estimates over these 5,000 data sets are computed for each method. The difference between the average of the estimates and its true value is the bias of one estimator; the SD of the parameter estimates over these 5,000 repeated samples is the standard error (SE) of the parameter estimate, capturing estimation variability.

Table W12 presents the simulation results. For each parameter, we report the average of the estimates and SE in the parenthesis computed using 5,000 generated data sets. As expected, OLS yields significant estimation bias at all values of N. For example, when N=200, the OLS regression yields large bias in the parameter estimates ($\widehat{\alpha}_1$: 1.413 [0.188]) and the error standard deviation ($\widehat{\sigma}$: 0.726 [0.037]) in the structural regression model. Copula-Main corrects for the endogenous bias ($\widehat{\alpha}_1$: -0.964 [1.049]; $\widehat{\sigma}$: 1.013 [0.202]), demonstrating that there is no need to additionally include C_{P^2}. Meanwhile, Copula-All yields substantial bias for the coefficient parameter of P^2 ($\widehat{\alpha}_2$: 0.771 [2.214]) because adding unnecessary generated regressor C_{P^2} leads to the finite sample bias problem. In contrast, Copula-Main eliminates the majority of the bias and performs much better in this small sample size with only small bias and the SE reduced by approximately 70% ($\widehat{\alpha}_2$: 0.922 [0.797]). In a large sample size (n=5,000), the finite sample bias in Copula-All is reduced. Yet, Copula-All continues to yield less precise estimates (i.e. larger standard errors) than Copula-Main.

Table W12: Results from Case III: Endogenous Squared Terms.

N	Method	$\alpha_0(=0)$	$\alpha_1(=-1)$	$\alpha_2(=1)$	$\sigma(=1)$	D-error
200	OLS	0.000	**1.413**	0.986	**0.726**	
		(0.078)	(0.188)	(0.742)	(0.037)	—
	Copula-Main	-0.001	-0.964	**0.922**	1.013	
		(0.099)	(1.049)	(0.797)	(0.202)	0.835
	Copula-All	0.009	-0.957	**0.771**	1.020	
		(0.190)	(1.057)	(2.214)	(0.203)	2.338
500	OLS	0.001	**1.410**	0.982	**0.728**	
		(0.048)	(0.118)	(0.472)	(0.024)	—
	Copula-Main	0.001	-0.978	0.951	1.005	
		(0.057)	(0.640)	(0.483)	(0.126)	0.309
	Copula-All	0.004	-0.974	**0.889**	1.008	
		(0.120)	(0.641)	(1.393)	(0.126)	0.891
5,000	OLS	0.000	**1.413**	1.003	**0.728**	
		(0.015)	(0.036)	(0.146)	(0.007)	—
	Copula-Main	0.000	-1.000	0.994	1.001	
		(0.019)	(0.192)	(0.157)	(0.038)	0.030
	Copula-All	0.000	-1.000	0.997	1.001	
		(0.037)	(0.192)	(0.427)	(0.038)	0.082
50,000	OLS	0.000	**1.415**	1.001	**0.728**	
		(0.005)	(0.012)	(0.047)	(0.002)	—
	Copula-Main	0.000	-1.004	1.000	1.001	
		(0.006)	(0.060)	(0.050)	(0.012)	0.003
	Copula-All	0.000	-1.004	0.999	1.001	
		(0.012)	(0.060)	(0.137)	(0.012)	0.008

Table presents the averages of the estimates and standard errors in the parenthesis over the repeated samples. Bold numbers highlight the estimates with bias of at least 0.05. Underlined numbers highlight the cases where the standard errors of the estimates from Copula-All are inflated by at least 50% compared with the corresponding ones from Copula-Main. The P&G method is used for copula correction since the model contains no exogenous regressors.

We also compute the ratio of mean squared error (MSE) of the structural estimate $\widehat{\alpha}_k$, comparing Copula-All to Copula-Main (where $\text{MSE}(\widehat{\alpha}_k) = \text{Bias}^2(\widehat{\alpha}_k) + \text{Var}(\widehat{\alpha}_k)$, measuring overall estimation accuracy). Notably, Copula-All increases MSEs for all model parameter estimates, with the harmful effects being greatest for the squared term estimate $\widehat{\alpha}_2$, whose MSE is more than 6-times that of Copula-Main for all sample sizes (Figure W6).

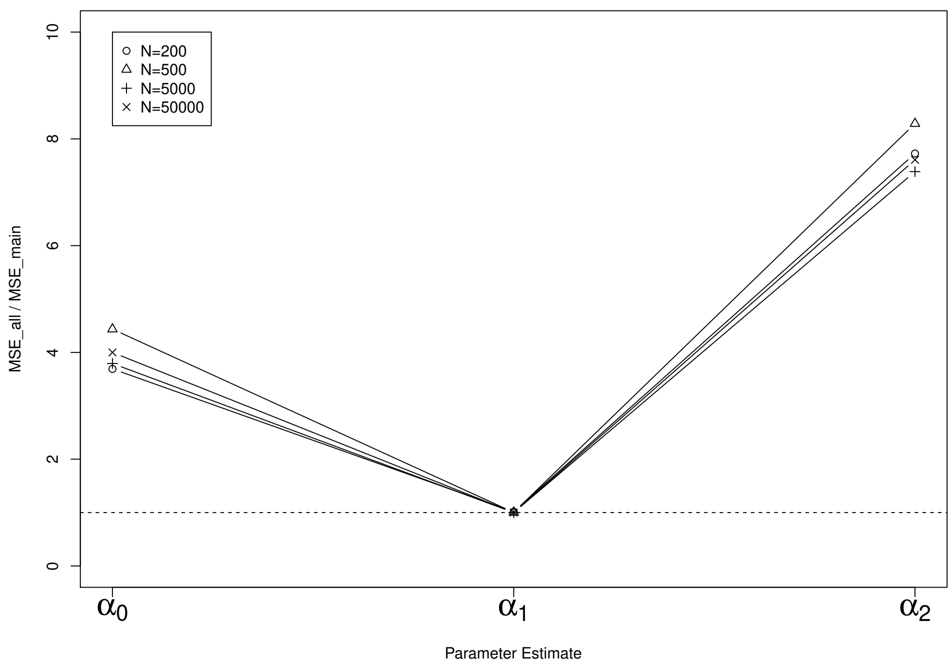

Figure W6: Ratio of mean squared errors of structural model estimates, with using the copula square term (Copula-All) to those without using the copula square term (Copula-Main).

Such a large magnitude of variance inflation has important inferential consequences and managerial implications. Figure W7 shows substantial loss of power of Copula-All to detect the presence of the squared term (P^2) for sample size up to 5,000. For example, when sample size is 1,000, the statistical power to detect the squared effect is about 8-fold for Copula-Main ($\approx 80\%$ power) of that for Copula-All ($\approx 10\%$ power).

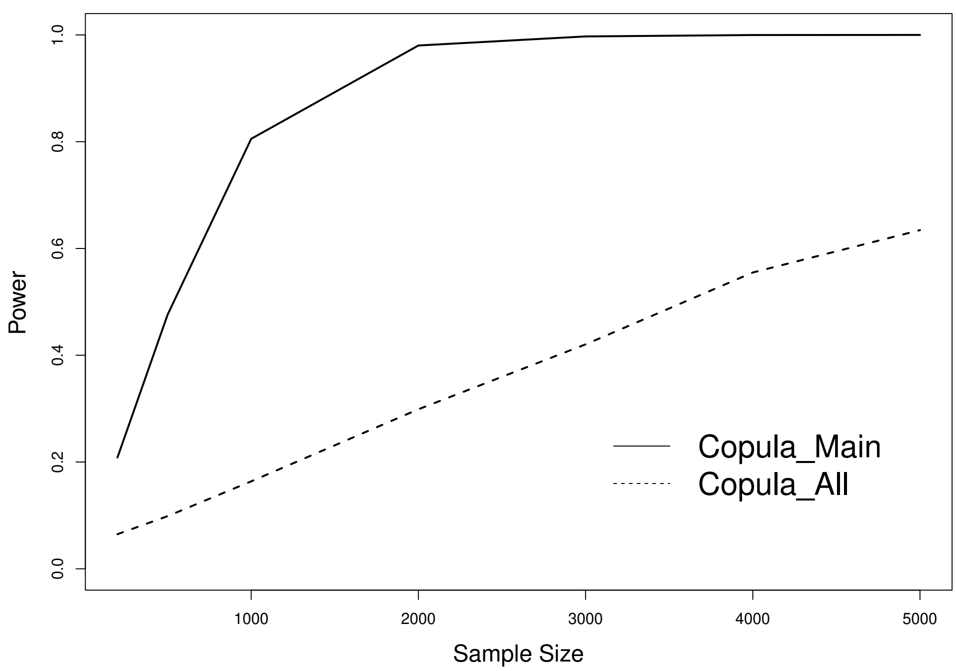

Figure W7: Statistical Power to detect the squared term P^2 with the copula squared term (Copula-All) and without the copula squared term (Copula-Main).

Mean-Centering Regressors

Lastly, we examine whether mean-centering resolves the under-performance of Copula-All. One may suspect that mean-centering might reduce the multicollinearity issue and improve the performance of Copula-All. However, as shown below, mean-centering regressors does not overturn the sub-optimal performance of adding the unnecessary copula correction for higher-order terms, demonstrating again that these unnecessary copula correction terms should be omitted from empirical models.

A common practice for researchers in economics, management, and other fields is to mean-center the regressors before estimating models with higher-order terms. One argument for this practice is that by mean-centering the regressors, the correlation - and resulting collinearity problem - between the linear and higher-order terms (e.g., quadratic terms or interaction terms) is reduced (Aiken and West 1991; Kopalle and Lehmann 2006). However, Echambadi and Hess (2007) showed that mean-centering regressors does not alleviate collinearity problems in moderated regression models. Namely, none of the parameter estimates and sampling accuracy of main effects, simple effects, interactions, or R^2 is changed by mean-centering. By main effect and simple effect, we refer to the regression coefficient for a first-order term with and without mean-centering, representing the effect of a regressor when its moderators are set at their mean values and at zero (or absence of the attribute quantified by these moderators), respectively.

To illustrate this point, consider the following structural regression model with an interaction term:

$$Y = \alpha_0 + \alpha_1 P_1 + \alpha_2 P_2 + \alpha_3 P_1 * P2 + E$$

For the purposes of ease in interpretation or reducing the correlation between the linear and interaction terms, mean-centering regressors is often employed, which leads to the following equivalent model with parameter transformation:

$$Y = \alpha_0^c + \alpha_1^c(P_1 - \bar{P}_1) + \alpha_2^c(P_2 - \bar{P}_2) + \alpha_3^c(P_1 - \bar{P}_1) * (P_2 - \bar{P}_2) + E, \qquad \text{(W52)}$$

where the parameters for the models before and after mean-centering have the following one-to-one relationship:

$$\begin{aligned}
\alpha_0^c &= \alpha_0 + \alpha_1\bar{P}_1 + \alpha_2\bar{P}_2 + \alpha_3\bar{P}_1\bar{P}_2 \\
\alpha_1^c &= \alpha_1 + \alpha_3\bar{P}_2 \\
\alpha_2^c &= \alpha_2 + \alpha_3\bar{P}_1 \\
\alpha_3^c &= \alpha_3. \qquad \text{(W53)}
\end{aligned}$$

As shown above, the regression coefficient α_1^c for the centered linear term $P_1 - \bar{P}_1$ represents the effect of P_1 when P_2 is equal to its mean value $\bar{P}_2$. Thus, α_1^c represents the main effect: the effect of P_1 when the other variables are at their mean values. In contrast, the coefficient using uncentered data, α_1, represents the simple effect: the effect of P_1 when the other variables are at zero (or absence of the attribute quantified by these other variables). The differences in estimates and standard errors between α_1 and α_1^c are due to the two coefficients having different substantive meanings, and both effects can be of substantive interest (Echambadi and Hess 2007). Quadratic terms can be considered a special case of the above model because a quadratic term can be considered as the interaction term of a regressor with itself. The relationship between parameters for models with quadratic terms before and after mean-centering can be derived similarly. Echambadi and Hess (2007) showed that the relationships in Equation W53 also hold for the OLS estimates of these model parameters.

However, our setting differs from the case of moderated regression models considered in (Echambadi and Hess 2007), since we consider the more general case of endogeneity bias correction of structural regression models with endogenous higher-order regressors. Although the relationships in Equation W53 hold exactly for OLS estimates (Echambadi and Hess 2007) for all data sets, such relationships only hold approximately for copula corrected estimates because copula generated regressors involve probability integral transformations. Specifically, we use the same data generating process for Cases I, II, and III to generate data. When estimating models, we first mean-center all the first-order terms of the regressors, and then construct the higher-order terms using these mean-centered first-order terms. Copula correction terms are then constructed using these new regressors based on centered versions of the first-order terms of regressors. Because these copula correction terms involve probability integral transformation, the estimates and sampling accuracy of main effects, simple effects, and interactions can change after mean centering, which differs from the case of Echambadi and Hess (2007) in which all regressors are exogenous.

For the models giving results in Tables W10, W11, and W12, we apply the OLS (without any correction), Copula-Main, and Copula-All to estimate the corresponding mean-centered structural regression models, with results summarized in Tables W13, W14, and W15, respectively. The true values for the parameters in the models after mean-centering are also listed in Tables W13 to W15. The mean values of the regressors $(\bar{P}_1, \bar{P}_2)$ used to compute these true parameter values are: $\frac{\phi(a)-\phi(b)}{\Phi(b)-\Phi(a)}$, where $\phi(\cdot)$ denotes the density function of the standard normal; when the marginal distribution of the regressor is the truncated standard normal on $[a, b]$, and $\frac{a+b}{2}$ when it is the uniform distribution on $[a, b]$.

Because copula correction terms for higher-order terms are not invariant to mean-centering, the ratios of the D-error for Copula-All to that of Copula-Main using mean-centered data will not be the same as those in Tables W10, W11, and W12, using uncentered data. Still, the same conclusion of inflated variability of estimates for Copula-All is apparent, and the D-error measure ratios are all above 2. This finding is consistent with that of Echambadi and Hess (2007) in that mean-centering regressors does not alleviate collinearity problems in moderated regression models. Furthermore, mean-centering seemingly shifts the variance inflation from the regression coefficient estimates of first-order terms to those of the higher-order terms, and may hurt the estimation of the higher-order terms in some cases.

It is important to note, however, that this does not imply that mean-centering affects the estimation of the *same* first-order effects. As explained above, the regression coefficients for a first-order term (with and without mean-centering) represent different effects of one regressor evaluated at different values of its moderator: these regression coefficients represent the main effects when mean-centering regressors and the simple effects when using uncentered data. As such, regression coefficients for a first-order term with and without mean-centering are not directly comparable, although both main and simple effects can be of substantive interest (Echambadi and Hess 2007). When using the parameter estimates based on the centered data to compute the simple effects, we again find finite sample bias and inflated standard errors for the estimates of simple effects (results not shown here), as occurred when using uncentered data. In sum, we conclude that mean-centering does not overturn the under-performance of Copula-All relative to Copula-Main.

Table W13: Results from Case I with Mean-Centering: Interaction of Endogenous Regressors With Mean-Centering

N	Method	$\alpha_0^c(=8.192)$	$\alpha_1^c(=1.798)$	$\alpha_2^c(=4)$	$\alpha_3^c(=1)$	$\sigma(=1)$	D-error
200	OLS	**8.259**	**3.425**	**5.619**	**1.432**	**0.294**	
		(0.208)	(0.071)	(0.084)	(0.105)	(0.031)	—
	Copula-Main	8.172	1.897	4.072	1.041	0.967	
		(0.208)	(0.279)	(0.257)	(0.080)	(0.124)	0.0316
	Copula-All	8.180	1.896	4.069	**1.101**	0.972	
		(0.215)	(0.279)	(0.266)	<u>(0.281)</u>	(0.124)	0.0734
500	OLS	**8.262**	**3.425**	**5.615**	**1.431**	**0.297**	
		(0.134)	(0.045)	(0.051)	(0.065)	(0.02)	—
	Copula- Main	8.184	1.838	4.018	1.025	0.990	
		(0.133)	(0.179)	(0.166)	(0.047)	(0.077)	0.0123
	Copula-All	8.189	1.838	4.020	**1.057**	0.992	
		(0.137)	(0.178)	(0.174)	<u>(0.173)</u>	(0.078)	0.0293
5,000	OLS	**8.263**	**3.424**	**5.612**	**1.433**	**0.298**	
		(0.042)	(0.014)	(0.017)	(0.021)	(0.006)	—
	Copula-Main	8.191	1.803	3.999	1.003	1.000	
		(0.042)	(0.057)	(0.051)	(0.015)	(0.024)	0.0011
	Copula-All	8.192	1.803	3.999	1.009	1.000	
		(0.043)	(0.057)	(0.054)	<u>(0.052)</u>	(0.024)	0.0028
50,000	OLS	**8.263**	**3.424**	**5.613**	**1.433**	**0.298**	
		(0.013)	(0.004)	(0.005)	(0.007)	(0.002)	—
	Copula-Main	8.192	1.799	3.999	1.000	1.000	
		(0.013)	(0.018)	(0.017)	(0.005)	(0.008)	0.0001
	Copula-All	8.192	1.799	3.999	1.002	1.000	
		(0.014)	(0.018)	(0.017)	<u>(0.017)</u>	(0.008)	0.0003

See the same note under Table W12.

Table W14: Results from Case II with Mean-centering: Interaction between Endogenous and Exogenous Regressors With Mean-centering.

N	Method	$\alpha_0^c(=8.192)$	$\beta_1^c(=1.798)$	$\alpha_1^c(=4)$	$\alpha_2^c(=1)$	$\sigma(=1)$	D-error
200	OLS	8.232	**2.322**	**5.088**	**1.273**	**0.831**	
		(0.195)	(0.130)	(0.129)	(0.184)	(0.041)	—
	2sCOPE-Main	8.191	1.823	4.045	1.024	0.995	
		(0.196)	(0.241)	(0.433)	(0.195)	(0.127)	0.0434
	2sCOPE-All	8.198	1.821	4.044	**1.066**	1.017	
		(0.226)	(0.250)	0.(461)	(0.704)	(0.131)	0.1459
500	OLS	8.234	**2.331**	**5.096**	**1.273**	**0.833**	
		(0.131)	(0.078)	(0.081)	(0.113)	(0.027)	—
	2sCOPE-Main	8.190	1.805	4.001	1.004	1.005	
		(0.132)	(0.159)	(0.291)	(0.127)	(0.088)	0.0169
	2sCOPE-All	8.193	1.805	4.003	1.022	1.014	
		(0.147)	(0.161)	(0.303)	(0.462)	(0.090)	0.0475
5,000	OLS	8.236	**2.325**	**5.088**	**1.276**	**0.833**	
		(0.041)	(0.024)	(0.027)	(0.036)	(0.008)	—
	2sCOPE-Main	8.191	1.798	3.999	1.000	1.001	
		(0.041)	(0.049)	(0.088)	(0.040)	(0.027)	0.0017
	2sCOPE-All	8.191	1.798	3.998	1.000	1.002	
		(0.045)	(0.050)	(0.093)	(0.148)	(0.027)	0.0044
50,000	OLS	8.237	**2.325**	**5.088**	**1.277**	**0.833**	
		(0.012)	(0.008)	(0.008)	(0.012)	(0.003)	—
	2sCOPE-Main	8.192	1.799	4.002	1.000	1.000	
		(0.012)	(0.015)	(0.027)	(0.012)	(0.008)	0.0002
	2sCOPE-All	8.191	1.799	4.002	1.002	1.000	
		(0.015)	(0.015)	(0.029)	(0.043)	(0.008)	0.0004

See the same note under Table W11.

Table W15: Results from Case III with Mean-centering: Endogenous Squared Terms With Mean-Centering

N	Method	$\alpha_0^c(=0)$	$\alpha_1^c(=-1)$	$\alpha_2^c(=1)$	$\sigma(=1)$	D-error
200	OLS	0.000	**1.414**	0.993	**0.727**	
		(0.080)	(0.188)	(0.737)	(0.037)	—
	Copula-Main	-0.001	-0.967	**0.912**	1.007	
		(0.085)	(1.008)	(0.785)	(0.193)	0.790
	Copula-All	0.000	-0.959	**0.857**	1.022	
		(0.196)	(1.019)	(2.353)	(0.194)	2.396
500	OLS	0.000	**1.414**	0.995	**0.729**	
		(0.049)	(0.117)	(0.458)	(0.024)	—
	Copula-Main	0.000	-0.993	0.949	1.005	
		(0.052)	(0.628)	(0.495)	(0.125)	0.311
	Copula-All	0.001	-0.999	**0.936**	1.011	
		(0.116)	(0.631)	(1.380)	(0.125)	0.871
5,000	OLS	-0.001	**1.413**	1.002	**0.728**	
		(0.016)	(0.038)	(0.151)	(0.007)	—
	Copula-Main	-0.001	-0.993	0.995	0.999	
		(0.017)	(0.201)	(0.159)	(0.040)	0.031
	Copula-All	-0.002	-0.993	1.008	0.999	
		(0.036)	(0.202)	(0.417)	(0.040)	0.085
50,000	OLS	-0.001	**1.415**	1.000	**0.728**	
		(0.005)	(0.013)	(0.045)	(0.002)	—
	Copula-Main	0.000	-1.003	1.000	1.001	
		(0.005)	(0.062)	(0.048)	(0.012)	0.003
	Copula-All	0.000	-1.003	0.998	1.001	
		(0.012)	(0.062)	(0.137)	(0.012)	0.009

See the same note under Table W12.

WEB APPENDIX G: ADDITIONAL MATERIALS FOR THE IMPLEMENTATION EXAMPLES

Further results of Example 1

Table W16 reports additional estimation results using 2sCOPE-np, although the sample size (n=261) is less than the recommended minimum sample size of 300. Overall, we find both 2sCOPE and 2sCOPE-np yield similar estimation results in the example.

Table W16: Additional Results for Example 1

Param.	OLS	2SLS	2sCOPE	2sCOPE-np
Intercept	6.005 (0.205)	4.371 (0.978)	4.763 (0.668)	5.102 (0.396)
	0.000	0.000	0.000	0.000
Price	-1.367 (0.137)	-2.470 (0.661)	-2.205 (0.446)	-1.977 (0.266)
	0.000	0.000	0.000	0.000
Feature	0.298 (0.095)	0.059 (0.178)	0.124 (0.124)	0.172 (0.103)
	0.002	0.738	0.317	0.095
Week	-0.002 (0.000)	-0.002 (0.000)	-0.002 (0.000)	-0.002 (0.000)
	0.000	0.000	0.000	0.000
Q_2	-0.019 (0.031)	-0.014 (0.035)	-0.018 (0.036)	-0.015 (0.034)
	0.550	0.693	0.617	0.659
Q_3	-0.018 (0.032)	-0.034 (0.036)	-0.029 (0.035)	-0.024 (0.033)
	0.567	0.349	0.407	0.467
Q_4	-0.018 (0.032)	-0.061 (0.041)	-0.044 (0.035)	-0.036 (0.036)
	0.576	0.140	0.209	0.317
C_{price}			0.077 (0.037)	0.098 (0.035)
			0.037	0.005
ρ			0.366 (0.160)	0.320 (0.105)
			0.022	0.002

Note: Table presents estimates and bootstrapped standard errors in the parentheses, followed by the p-values in the line below.

Analysis and Results of Example 2

Example 2 examines what to do when an endogenous regressor has a higher-order effect, such as a squared term or interaction (moderation) with another variable. For brevity, we speak to these higher-order effects simply as interactions. The “METHODOLOGICAL BACKGROUND” section provided studies with simulated data showing that including a copula for the interaction term may induce bias and inflated estimation variability, and that the best course is to only include copula correction terms for the main effects.

To show how copula correction is applied with interactions of endogenous regressors and examine the adverse effects of including higher-order copula correction terms in an empirical application, we extend the sales response model in Equation 15 to include an interaction term ($P_t * F_t$) between price and feature as follows:

$$\text{Volume}_t = \mu + \alpha * P_t + \beta' W_t + \phi P_t * F_t + E_t, \tag{W54}$$

where P_t and F_t are category price and feature, respectively, and W_t includes F_t, week, and binary variables for quarters 2, 3, and 4. We use the IRI academic data set for a new store and product category, a New York City store and its peanut butter sales for the years 2001-2003 (156 weeks), allowing for price and feature to work together as an interaction. Such interactions are common to both academics and managers, as marketing efforts often work together. Of interest here is that price and feature advertising likely work together to achieve interactive, synergistic effects on sales. This can be tested by estimating the interaction term between price and feature advertisement in the above sales model, with feature advertisement as a potential moderator of price. Like Example 1, we follow the same steps in Figure 5 to guide the selection of the appropriate copula method. The walk-through of these steps are as follows:

Step 1. Is P_{main} continuous? Price is a continuous measure here, ranging from \$0.957 to \$1.963 per pound, with a mean of \$1.714, median of \$1.798, and standard deviation of \$0.195.

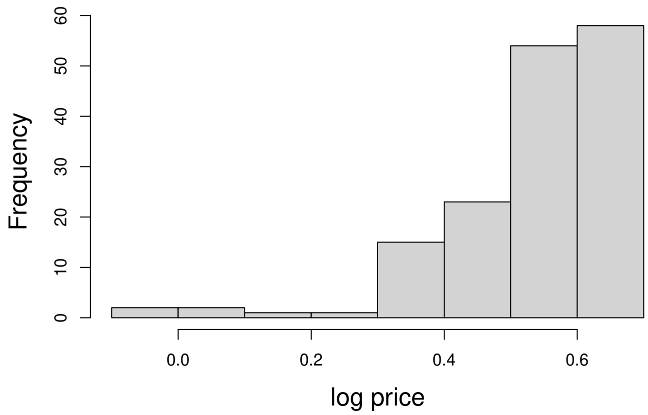

Figure W8: Price Distribution in Example 2.

Step 2. Is P_{main} normally distributed? Unlike Example 1, the price variable in Example 2 is highly skewed (Figure W8) and rejects the KS test for normality ($D = 0.23$, $p < 0.001$) at the 0.05 level of significance. The flowchart in Figure 5 show that what is needed is either P_{main} or one related W is nonnormally distributed; there is no need for both P_{main} and W to be nonnormally distributed. This means that when the endogenous regressor already has sufficient nonnormality, we do not need to check any exogenous regressor W for sufficient nonnormality and sufficient association with P, like what was needed in Figure 6 of Example 1. To determine if we should use P&G or 2sCOPE, we next check the uncorrelatedness between the linear combination of copula transformations of P_{main} with each W. When P_{main} is a scalar, this condition reduces to check the uncorrelatedness between P^*_{main} and each W.

Step 3.a. Is P^*_{main} correlated with W? The copula transformation of endogenous regressor price, P^*, is correlated with the following exogenous regressors at the 0.10 level of significance: week (r = 0.21, $p < 0.05$), feature (r = -.76, $p < 0.01$), Q3 (r = -.16, $p < 0.06$), and Q4 (r = 0.16, $p < 0.04$). This indicates we should use 2sCOPE for endogeneity correction.

Step 4. Perform 2sCOPE estimation. Until now, the steps had been met to indicate price was a candidate to use the 2sCOPE method. We will not use 2sCOPE-np in this dataset since the sample size (n=156) is well below the recommended minimum sample size of 300 (Boundary Condition 2 in Table 4).

Step 5. Check ICON statistics. The ICON statistics are the ratios of the standard errors of the 2sCOPE estimates to the standard errors of the corresponding OLS estimates. These standard errors are reported in Table W17 under columns "2sCOPE" and "OLS", and show the ICON statistics are all less than 6, so no model nonidentification issue is flagged.

Table W17 presents the 2sCOPE result with the copula correction term (i.e., the first-stage residual) for price only. The results show the price copula correction term (i.e., the first-stage residual) is significant (Est. = 0.069, SE = 0.028, $p < 0.05$), indicating the presence of endogeneity. Like Example 1, we compare the results to OLS and 2SLS, as well as when a copula correction term for the interaction term is also included (2sCOPE W/Int).

Similar to Example 1, price has the smallest absolute effect in the OLS model (Est. = -.453, SE = 0.274, $p < 0.10$) and the greatest absolute effect in the 2SLS model (Est. = -1.554, SE = 0.606, $p < 0.05$). The 2sCOPE estimate falls in between, closer to 2SLS in both effect and SE (Est. = -1.314, SE = 0.430, $p < 0.05$). The closeness to 2SLS is more expected here since the usage of another store's price is a strong instrument (r = 0.90, $p < 0.01$), as 2SLS rejects the test for weak instrument (F = 21.567, $p < 0.01$); the Wu-Hausman

Table W17: Estimation Results for Example 2

Parameters	OLS	2SLS	2sCOPE	2sCOPE W/Int
Intercept	6.038 (0.165)***	6.688 (0.359)***	6.544 (0.256)***	6.344 (0.307)***
Price	-0.453 (0.274)*	-1.554 (0.606)**	-1.314 (0.430)**	-0.999 (0.518)*
Feature	1.513 (0.234)***	0.646 (0.487)	0.837 (0.388)**	0.619 (0.420)
Price*Feature	-2.125 (0.379)***	-0.950 (0.694)	-1.167 (0.661)*	0.148 (0.825)
Week	0.001 (0.000)***	0.001 (0.000)***	0.001 (0.000)***	0.001 (0.000)***
Q_2	-0.028 (0.034)	-0.020 (0.036)	-0.022 (0.033)	-0.038 (0.041)
Q_3	-0.083 (0.035)**	-0.099 (0.038)***	-0.096 (0.034)***	-0.089 (0.045)**
Q_4	-0.090 (0.036)**	-0.081 (0.038)**	-0.080 (0.035)**	-0.066 (0.039)*
C_{price}			0.069 (0.028)**	0.058 (0.030)*
$C_{Price*Feature}$				-0.168 (0.098)*
ρ_1			0.185 (0.082)**	0.128 (0.086)
ρ_2				-0.456 (0.229)**

Note: Table presents estimates and bootstrapped standard errors in the parentheses. * is $p < 0.10$, ** is $p < 0.05$, *** is $p < 0.01$

test also suggests endogeneity ($W = 4.863$, $p < 0.03$). Without correcting for endogeneity in this example, managers would under-estimate the price elasticity by 65.5% in OLS.

Importantly, the 2sCOPE results point to a contrast with 2sCOPE when a copula correction term $C_{Price*Feature}$ is included for the interaction between price and feature. Here, the price estimate is substantially smaller and becomes insignificant (Est. = -.999, SE = 0.518, $p > 0.05$ under column "2sCOPE W/Int" in Table W17), which can lead to the incorrect conclusion that price had no significant effect on sales. A more striking difference regards the estimate of the interaction term Price*Feature. The Price*Feature estimates from 2SLS and 2sCOPE (excluding the copula interaction term) are both negative and close: the 2SLS Est. = -0.950 (SE = 0.694, $p > 0.10$) and 2sCOPE Est. = -1.167 (SE = 0.661, $p < 0.10$). By contrast, 2sCOPE including the copula term for Price*Feature yields an interaction estimate with the opposite sign and larger SE (Est. = 0.148, SE = 0.825, $p > 0.10$). These results mark an important point: when adding copula correction terms, only copula terms for the main effects should be included, and no copula terms for higher-order terms should be included. Adding the unnecessary higher-order copula terms can exacerbate the multicollinearity issue (Web Appendix Table W21) and lead to substantially varied and biased estimates.

Although the P&G method was not selected in both examples according to the flowchart in Figure 5, Table W18 presents the results of applying P&G methods to the two implementation examples. In Example 1, the parameter estimates of 2sCOPE and P&G are similar except the coefficient estimate for Feature (0.124 for 2sCOPE vs 0.276 for P&G vs 0.059 for 2SLS). The differences between P&G and 2sCOPE estimates are more pronounced in Example 2. Besides the Feature coefficient estimate, we observed differences for Price (-1.314 for 2sCOPE vs -0.999 for P&G) and Price*Feature (-1.167 for 2sCOPE vs -1.621 for P&G). Furthermore, in agreement with the 2SLS result, 2sCOPE identifies the presence of price endogeneity (0.069 for the coefficient of copula term C_{price}, p-value < 0.05) while P&G does not (0.046 for the coefficient of copula term C_{price}, p-value > 0.10) (Table W18).

Table W18: Estimation Results Using P&G

	Example 1		Example 2	
Parameters	2sCOPE	P&G	2sCOPE	P&G
Intercept	4.763 (0.668)***	4.748 (0.683)***	6.544 (0.256)***	6.344 (0.346)***
Price	-2.205 (0.446)***	-2.204 (0.468)***	-1.314 (0.430)**	-0.999 (0.592)*
Feature	0.124 (0.124)	0.276 (0.092)***	0.837 (0.388)**	1.255 (0.434)***
Price*Feature			-1.167 (0.661)*	-1.621 (0.779)**
Week	-0.002 (0.000)***	-0.002 (0.001)***	0.001 (0.000)***	0.001 (0.000)***
Q_2	-0.018 (0.036)	-0.023 (0.031)	-0.022 (0.033)	-0.029 (0.033)
Q_3	-0.029 (0.035)	-0.022 (0.028)	-0.096 (0.034)***	-0.088 (0.032)***
Q_4	-0.044 (0.035)	-0.014 (0.032)	-0.080 (0.035)**	-0.086 (0.035)**
C_{price}	0.077 (0.037)**	0.078 (0.039)**	0.069 (0.028)**	0.046 (0.037)
ρ_1	0.366 (0.160)**	0.412 (0.181)**	0.185 (0.082)*	0.203 (0.226)

Note: Table presents estimates and bootstrapped standard errors in the parentheses. * is $p < 0.10$, ** is $p < 0.05$, *** is $p < 0.01$

Theoretically, the bias of the P&G method can be viewed as an omitted variable bias. With one endogenous regressor P and one exogenous regressor W in the model, the bias of the P&G method that ignores the correlation between the endogenous regressor (P) and the exogenous regressors (W) comes from the omitted variable $\sigma\frac{-q\rho}{1-q^2}W_i^*$, absorbed into the error term in the augmented regression model (Appendix of Haschka 2022). Consequently, the bias of the P&G method for α due to ignoring the correlations between P and W is:

$$\sigma\frac{-q\rho}{1-q^2}\left[\texttt{Cov}(P,W^*)/\texttt{Var}(P)\right], \tag{W55}$$

where σ is the variance of the structural error, ρ is the correlation between P and the structural error, q is the correlation between P and W, $\texttt{Cov}(P,W^*)$ is the partial association between P and the omitted variable W^* given P^* and W, and the variance of P is $\texttt{Var}$(P). The formula sheds light on the sources affecting the sign and magnitude of the bias of the P&G method. For example, if the explained part of the variation in the dependent variable is large (i.e., small σ), we can expect the bias of P&G due to ignoring the correlation between P and W to be minimal. The stronger the correlation between P and W (i.e., larger q), the larger the bias of P&G. Also, if P has a wide variation relative to the partial covariance between P and W^* given P^* and W, the bias of P&G would be small. Given a value of $\texttt{Var}(P)$, the smaller the partial covariance between P and W^* given P^* and W, the smaller the omitted variable bias of the P&G method. However, a 'too small' value of the partial covariance between P and W^* given P^* and W may mean high collinearity between P and P^* (or between W and W^*) such that the remaining partial covariance $\texttt{Cov}(P,W^*)$ given P^* and W can only take small values. This can cause P&G estimates to suffer from finite sample bias due to insufficient regressor nonnormality. Thus, the overall bias due to both ignoring the regressor dependence and insufficient regressor nonnormality can be complicated. Furthermore, in practice, the true values of ρ (the magnitude of endogeneity) is unknown, preventing an accurate assessment of the sign or magnitude of the bias for P&G.

Fortunately, the alternative 2sCOPE method is easy to apply and account for the dependence between regressors. Because 2sCOPE employs the GC models, the computational complexity increases at a much slower rate than other multivariate models as the number of dimensions increases (Danaher and Smith 2011). Thus, it is computationally feasible to run these more general copula correction methods to account for the dependence between regressors. As shown in Yang, Qian, and Xie (2024a), the estimation efficiency loss (i.e., the increase in standard errors) of 2sCOPE relative to P&G is negligible when the endogenous and exogenous regressors have no or weak correlations and 2sCOPE is the preferred method unless sample size is very small. When exogenous and endogenous regressors are correlated, 2sCOPE not only can remove the bias of P&G, but also can possibly increase estimation efficiency and reduce standard errors by leveraging correlated exogenous regressors.

Next we consider appropriateness of using 2sCOPE-HGC in the examples. The general-location heterogeneous GC (HGC) model (Yang, Qian, and Xie 2024b) for panel data can also be applied to grouped data formed by discrete exogenous regressors that generalizes Liengaard et al. (2024). Let $W = (W_c, W_d)$ where W_c and W_d denote the continuous and discrete exogenous regressors, respectively. Liengaard et al. (2024) permits the GC dependence structure and the copula correction terms to vary by the levels of discrete exogenous regressors in W_d. When the levels of combinations of all discrete regressors are not small, this approach may lead to sparse data insufficient for ECDF estimation and a larger number of copula parameters and copula correction terms than necessary, resulting in inflated estimation variance and estimation bias. Thus, it is important to have sufficient sample size and meet data requirements (shown in the Flowchart in Figure 5) within each level of combinations of discrete exogenous regressors.[32] Yang, Qian, and Xie (2024b) propose a more flexible 2sCOPE estimator based on a general-location heterogeneous GC model (see Web Appendix Table W19).

The general-location HGC model permits the location and the GC dependence of the error term and continuous regressors to vary by W_d in different ways. The 2sCOPE-HGC procedure follows a modified two-stage estimation process (Web Appendix Table W19) with the following augmented regression model

$$\begin{aligned} Y_i &= \mu + \sum_{k=1}^{K} P_{i,k}\alpha_k + \beta' W_i + \sum_{k=1}^{K} \left\{ C_{i,k}\gamma_{k0} + \sum_{j=1}^{G-1} C_{i,k} I(g_i(w_d) = j) * \gamma_{kj} \right\} + \omega_t, \quad \text{(W56)} \\ \text{where } C_{i,k} &= (\widetilde{P}_{i,k})^{*|g_i(W_d)} - \delta'_{g_i(W_d),k}(\widetilde{W}_{c,i})^{*|g_i(W_d)}. \quad \text{(W57)} \end{aligned}$$

Inside the copula term $C_{i,k}$, $\widetilde{P}_{i,k} = P_{i,k} - \bar{P}_k^{m_i}, \widetilde{W}_{c,i} = W_{c,i} - \bar{W}_c^{m_i}$, where $\bar{P}_k^{m_i}$ and $\bar{W}_c^{m_i}$ are the group mean of P_k and W_c for observations in the same group m_i as the observation i and the groups $\{m_i\}$ are formed by the observed levels of combinations of the discrete regressors. Thus, $\widetilde{P}_k^{m_i}$ and $\widetilde{W}_k^{m_i}$ are simply within-group demeaned P_k and W_c to account for potential effects of discrete regressors on the location of continuous regressors. The model further permits the GC dependence structure of the demeaned continuous regressors and the error term to vary by the group variable $g_i(W_d)$ defined on W_d. The notation $*|g_i(W_d)$ in Equation W57 denotes empirical copula transformation using only observations within the group $g_i(w_d)$, across which the GC dependence may vary. The 2sCOPE-HGC is

[32] Simulation results (Figure 2 in Liengaard et al. 2024) show the finite sample estimation bias remains before sample size reaches between 1600 and 3200 observations for an exogenous regressor with two levels. The finite sample bias depends on the normality of regressors and correlations between endogenous and exogenous regressors.

more general than that of Liengaard et al. (2024) in that 2sCOPE-HGC allows for different sets of discrete exogenous regressors to separately affect the location and GC dependence structure. For example, two discrete exogenous regressors W_{d1} and W_{d2} may both affect the location but the dependence structure only vary by W_{d1}.

Table W19: Estimation Procedure for 2sCOPE-HGC

Stage 1:

- Do group demeaning of $P_{i,k}$ and $W_{c,i}$ and obtain the demeaned regressors $(\widetilde{P}_{i,k}, \widetilde{W}_{c,i})$.
- Within each of the subgroups $\{g_i(W_d)\}$ across which GC dependence may vary, apply Stage 1 of the 2sCOPE to the demeaned continuous regressors $(\widetilde{P}_{i,k}, \widetilde{W}_{c,i})$ and obtain residual $C_{i,k} = (\widetilde{P}_{i,k})^{*|g_i(W_d)} - \delta'_{g_i(W_d),k}(\widetilde{W}_{c,i})^{*|g_i(W_d)}$ (Equation W57).

Stage 2:

- Add $C_{i,k}$ and the interaction terms between $C_{i,k}$ and the indicator variables for the (non-reference) levels of the group variable (Equation W56).

It is important to have sufficient sample size and meet data requirements (shown in the Flowchart in Figure 5) within each level of combinations of discrete exogenous regressors in order to apply 2sCOPE-HGC. Both examples contain quarters as the discrete exogenous regressors. In Example 1, within each group of observations formed by the quarters, no data satisfy the requirement in Figure 5. The test for normality of price fails to reject normality in all groups formed by quarters, and within no group the F-stat for any W have $F > 10$. This means data in Example 1 do not satisfy the data requirement for 2sCOPE-HGC while the 2sCOPE meets data requirements. In Example 2, the price variable in Quarter 3 rejects normality ($p < 0.02$). For other quarters, the price variable fails to reject the normality assumption and no W variable is found to have sufficient relevance (F >10) with the price variable in groups formed in these quarters. Thus, strictly speaking, 2sCOPE-HGC does not satisfy all data requirements and one should be cautious about applying 2sCOPE-HGC to this example as well, although to a lesser extent. However, for illustration purposes, the result of 2sCOPE-HGC for this example is presented in Table W20. We observe that 2sCOPE-HGC yielded results that largely agree with 2sCOPE rather than with OLS. Furthermore, none of the interactions between the C_{price} and quarters (i.e., $C_{price} * Q2$, $C_{price} * Q3$, $C_{price} * Q4$) is statistically significant. Thus we conclude that no evidence supports the HGC model. Overall, the more parsimonious 2sCOPE is preferred.

Table W20: Further Estimation Results for Example 2

Parameters	OLS	2SLS	2sCOPE	2sCOPE-HGC
Intercept	6.038 (0.165)***	6.688 (0.359)***	6.544 (0.256)***	6.378 (0.353)***
Price	-0.453 (0.274)*	-1.554 (0.606)**	-1.314 (0.430)**	-1.037 (0.591)*
Feature	1.513 (0.234)***	0.646 (0.487)	0.837 (0.388)**	1.072 (0.487)**
Price*Feature	-2.125 (0.379)***	-0.950 (0.694)	-1.167 (0.661)*	-1.513 (0.740)**
Week	0.001 (0.000)***	0.001 (0.000)***	0.001 (0.000)***	0.001 (0.000)***
Q_2	-0.028 (0.034)	-0.020 (0.036)	-0.022 (0.033)	-0.024 (0.033)
Q_3	-0.083 (0.035)**	-0.099 (0.038)***	-0.096 (0.034)***	-0.093 (0.036)***
Q_4	-0.090 (0.036)**	-0.081 (0.038)**	-0.080 (0.035)**	-0.085 (0.036)***
C_{price}			0.069 (0.028)**	0.049 (0.045)
C_{price}∗Q2				0.033 (0.056)
C_{price}∗Q3				0.016 (0.069)
C_{price}∗Q4				-0.051 (0.050)

Note: Table presents estimates and bootstrapped standard errors in the parentheses. * is $p < 0.10$, ** is $p < 0.05$, *** is $p < 0.01$.

Table W21: VIF Results in Example 2

Parameters	2sCOPE		2sCOPE W/Int	
	Est. (SE)	VIF	Est. (SE)	VIF
Intercept	6.544 (0.256)***	—	6.344 (0.307)***	—
Price	-1.314 (0.430)**	27.9	-0.999 (0.518)*	29.1
Feature	0.837 (0.388)**	59.3	0.619 (0.420)	61.5
Price*Feature	-1.167 (0.661)*	18.8	0.148 (0.825)	29.1
Week	0.001 (0.000)***	1.2	0.001 (0.000)***	1.2
Q_2	-0.022 (0.033)	1.5	-0.038 (0.041)	1.6
Q_3	-0.096 (0.034)***	1.7	-0.089 (0.045)**	1.7
Q_4	-0.080 (0.035)**	1.7	-0.066 (0.039)*	1.7
C_{price}	0.069 (0.028)**	3.2	0.058 (0.030)*	3.2
$C_{Price*Feature}$			-0.168 (0.098)*	6.2

Note: Table presents estimates and bootstrapped standard errors in the parentheses. * is $p < 0.10$, ** is $p < 0.05$, *** is $p < 0.01$. Regression models with interaction terms will often yield high VIF values because of high correlations between variables and their interactions. Such high VIF values do not imply problems in terms of estimation and inference for models with interaction terms (Kalnins and Hill 2023, p.72, and Echambadi and Hess 2007). However, in the case of copula correction, adding the unnecessary copula term $C_{Price*Feature}$ for interaction term exacerbates the multicollinearity issue that substantially increases the VIF for the interaction term estimate from 18.8 to 29.1, cause inflated standard errors, and introduce potential finite sample bias as shown in our simulation studies.

Implications in Example 2

Example 2 presented the case of the interaction between an endogenous and exogenous regressor (Web Appendix Table W17). Like Example 1, price elasticity in the absence of feature was substantially under-estimated in OLS (Est. = -0.453) than 2SLS (Est. = -1.554) or 2sCOPE (-1.314). The OLS price elasticity estimate was nearly a third that of 2sCOPE.

Furthermore, 2sCOPE including a copula term for the interaction term biased the price elasticity estimate downwards (Est. = -0.999), about 30% lower as compared with the estimate of -1.314 from 2sCOPE excluding this copula term (Web Appendix Table W17). This bias in the price elasticity estimate becomes even larger as feature intensity increases. Including the copula term for the endogenous interaction term of Price*Feature yields a severely biased interaction effect estimate; while 2sCOPE without this unnecessary copula term had a negative estimate of -1.176, 2sCOPE including this term (2sCOPE W/Int) produced a positive estimate of 0.148 (Table W17). As shown in Figure 4, including the unnecessary copula term for Price*Feature yields price sensitivity estimates that are the same across different feature intensity (meaning lack of interactive effect); excluding this copula term yields much greater magnitude of price sensitivity that increases with greater feature advertisement. Such drastic differences in price elasticity estimates can have substantive managerial implications, including the optimal price setting and profit maximization, like in Example 1.

REFERENCES FOR WEB APPENDIX

Arora, Neeraj and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments.," *Journal of Consumer Research*, **28**, 273–83.

Atefi, Yashar, Michael Ahearne, James G Maxham III, Todd D Donavan, and Brad D Carlson (2018), "Does Selective Sales Force Training Work?," *Journal of Marketing Research*, **55 (5)**, 722–737.

Aiken, Leona S and Stephen G West (1991), *Multiple Regression: Testing and Interpreting Interactions*, Newbury Park: Sage Publications.

Blauw, Sanne and Philip Hans Franses (2016), "Off the Hook: Measuring the Impact of Mobile Telephone Use on Economic Development of Households in Uganda using Copulas," *Journal of Development Studies*, **52(3)**, 315–330.

Burmester, Alexa B, Jan U Becker, Harald J van Heerde, and Michel Clement (2015), "The Impact of Pre-and Post-launch Publicity and Advertising on New Product Sales," *International Journal of Research in Marketing*, **32 (4)**, 408–417.

Cid, Jaime A and von Davier, Alina A (2015), "Examining potential boundary bias effects in kernel smoothing on equating: An introduction for the adaptive and Epanechnikov kernels," *Applied Psychological Measurement,* **39 (3)**, 208–222.

Echambadi, Raj and James D Hess (2007), "Mean-Centering Does Not Alleviate Collinearity Problems in Moderated Multiple Regression Models," *Marketing Science*, **26(3)**, 438–445.

Gielens, Katrijn, Inge Geyskens, Barbara Deleersnyder, and Max Nohe (2018), "The New Regulator in Town: The Effect of Walmart's Sustainability Mandate on Supplier Shareholder Value," *Journal of Marketing*, **82(2)**, 124–141.

Guitart, Ivan A, Jorge Gonzalez, and Stefan Stremersch (2018), "Advertising Non-premium Products as if They Were Premium: The Impact of Advertising Up on Advertising Elasticity and Brand Equity," *International Journal of Research in Marketing*, **35 (3)**, 471–489.

Guitart, Ivan A and Stefan Stremersch (2021), "The Impact of Informational and Emotional Television Ad Content on Online Search and Sales," *Journal of Marketing Research*, **58 (2)**, 299–320

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**, 257–277.

Homburg, Christian, Arnd Vomberg, and Stephan Muehlhaeuser (2020), "Design and Governance of Multichannel Sales Systems: Financial Performance Consequences in Business-to-business Markets," *Journal of Marketing Research*, **57 (6)**, 1113–1134.

Keller, Wiebke IY, Barbara Deleersnyder, and Karen Gedenk (2019), "Price Promotions and Popular Events," *Journal of Marketing*, **83 (1)**, 73–88

Kalnins, Arturs, and Kendall Praitis Hill (2023), "The VIF Score. What is it Good For? Absolutely Nothing," *Organizational Research Methods,* **28 (1)**, 58–75.

Karunamuni, Rhoana J., and Tom Alberts (2005), "On boundary correction in kernel density estimation," *Statistical Methodology,* **2 (3)**, 191–212.

Kopalle, Praveen K. and Donald R. Lehmann (2006), "Setting Quality Expectations When Entering a Market: What Should the Promise Be?," *Marketing Science*, **25(1)**, 8–24.

Kramer, Martin, Christina Desernot, Sascha Alavi, Christian Schmitz, Felix Bruggemann, and Jan Wieseke (2022), "The Role of Salespeople in Industrial Servitization: How to

Manage Diminishing Profit Returns From Salespeople's Increasing Industrial Service Shares," *International Journal of Research in Marketing*, **39(4)**, 1235–1252.

Lamey, Lien, Barbara Deleersnyder, Jan-Benedict EM Steenkamp, and Marnik G Dekimpe (2018), "New Product Success in the Consumer Packaged Goods Industry: A Shopper Marketing Approach," *International Journal of Research in Marketing*, **35 (3)**, 432–452.

Lenz, Isabell, Hauke A Wetzel, and Maik Hammerschmidt (2017), "Can Doing Good Lead to Doing Poorly? Firm Value Implications of CSR in the Face of CSI," *Journal of the Academy of Marketing Science*, **45 (5)**, 677–697.

Li, Q. and Racine, J. S. (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, **26(4)**:423–434.

Liu, Huan, Lara Lobschat, Peter C Verhoef, and Hong Zhao (2021), "The Effect of Permanent Product Discounts and Order Coupons on Purchase Incidence, Purchase Quantity, and Spending," *Journal of Retailing*, **97 (3)**, 377–393.

Magnotta, Sarah, Brian Murtha, and Goutam Challagalla (2020), "The Joint and Multilevel Effects of Training and Incentives From Upstream Manufacturers on Downstream Salespeople's Efforts," *Journal of Marketing Research*, **57 (4)**, 695–716.

Staiger, Douglas O and Stock, James H (1997), "Instrumental variables regression with weak instruments," *Econometrica*, **65**, 557-586.

Vomberg, Arnd, Christian Homburg, and Olivia Gwinner (2020), "Tolerating and Managing Failure: An Organizational Perspective on Customer Reacquisition Management," *Journal of Marketing*, **84 (5)**, 117–136.

Wetzel, Hauke A, Stefan Hattula, Maik Hammerschmidt, and Harald J van Heerde (2018), "Building and Leveraging Sports Brands: Evidence From 50 Years of German Professional Soccer," *Journal of the Academy of Marketing Science*, **46 (4)**, 591–611.

Yoon, Hyungseok David, Namil Kim, Bernard Buisson, and Fred Phillips (2018), "A Cross-national Study of Knowledge, Government Intervention, and Innovative Nascent Entrepreneurship," *Journal of Business Research*, **84**, 243–252.