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 April 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 April 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. At the forefront of 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 and its theoretical rationales and advantages for usage in empirical research, discuss recent advances that broaden the understanding, applicability, and robustness of copula correction, and examine implementation aspects of copula correction such as construction of copula control functions and handling of 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
Sauder School of Business
University of British Columbia
2053 Main Mall
Vancouver, BC V6T 1Z2
and NBER
yi.qian@sauder.ubc.ca

Hui Xie
Department of Biostatistics
Faculty of Health Sciences
Simon Fraser University
huixie@uic.edu

Anthony Koschmann
Eastern Michigan University
121 Hill Hall
College of Business
Ypsilanti, MI 48178
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 the presence of regressor-error dependence, such as the ordinary least squares (OLS) method, can lead to severe bias in the 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 to correct for endogeneity bias (Wooldridge 2010). This method relies on the existence of valid and strong IVs to satisfy the stringent requirement of exclusion restriction (ER), which makes IVs difficult to find and justify in practice (Ebbes et al. 2005; Ebbes, Wedel, and Böckenholt 2009; Park and Gupta 2012). When there exists theory or knowledge about the underlying mechanism of endogeneity, an alternative approach is to model the exact process of yielding the observed values of the endogenous regressors, which is then estimated jointly with the structural model of primary interest. For instance, in estimating a consumer demand model, a supply-side model reflecting researchers' beliefs about the managerial decisions determining the supply-side marketing mix variables (such as price and promotions) can be specified and jointly estimated with the demand model (e.g., Sudhir 2001; Yang, Chen, and Allenby 2003; Manchanda, Rossi, and Chintagunta 2004). When the supply-side model is specified correctly, this approach can successfully correct for endogeneity bias in parameter estimates of the demand model.

Recently, there have been growing interests in endogeneity correction methods that require neither observed IVs nor knowledge to correctly specify an auxiliary supply model (Ebbes, Wedel, and Böckenholt 2009; Papies, Ebbes, and Van Heerde 2017; Rutz and Watson 2019; Papies, Ebbes, and Feit 2023; Park and Gupta 2024). These instrument-free 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 model (SORE, Qian and Xie 2024), and copulas[1] (Park and Gupta 2012; Becker, Proksch, and Ringle 2022; Christopoulos, McAdam, and Tzavalis 2021; Tran and Tsionas 2021; Eckert and Hohberger 2023; Haschka 2022; Yang, Qian, and Xie 2024a,b; Liengaard et al. 2024; Breitung, Mayer, and Wied 2024; Hu, Qian, and Xie 2025) to control for endogeneity.

Copula correction methods provide substantial advantages for addressing the prevalent and thorny issue of endogenous regressors. These methods directly address the regressor-error dependence using copulas, a widely used multivariate dependence model applicable in many practical applications (Danaher 2007; Danaher and Smith 2011). Unlike the traditional IV approach and other IV-free methods, copula correction methods do not require the endogenous regressor to contain an exogenous component (either observed or latent) satisfying the stringent exclusion restriction condition that is hard to justify in practical applications. Thus, copula correction is feasible in many situations under appropriate conditions. 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 versatile and general framework for further development (e.g., Liengaard et al. 2024; Yang, Qian, and Xie 2024b; Hu, Qian, and Xie 2025).

Furthermore, one can implement copula correction by including copula control functions

[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).

derived from existing regressors as additional regressors in the structural regression model to control for endogeneity. Thus, copula correction using the control function is straightforward to apply in a wide array of settings, including both linear and nonlinear models (e.g., discrete choice models) and the challenging slope endogeneity problem. We show that copula correction using control functions does not require normal structural error and copula regressor-error dependence structure as originally believed, thereby significantly increasing the applicability and robustness of copula correction.

Focusing on copula correction, the objectives of this article are: (a) to raise awareness of the importance to address endogenous regressors in empirical studies and improve understanding of theoretical rationales for using copula correction; (b) to provide a synthesis of recent advances that broaden the understanding, applicability, and robustness of copula corrections; (c) to provide practical guidance and delineate a process of checking data requirements and identification assumptions to aid appropriate usage of copula correction; and (d) to demonstrate use of copula endogeneity correction in practical applications.

With these objectives in mind, the rest proceeds as follows. The next section starts with an overview of why and when to use copula correction and how it addresses endogeneity. We also survey various disciplines and substantive marketing areas where copula correction has been used, as well as important variations in the use and implementation of copula correction. Next, we present relevant methodological background: how the copula handles endogeneity, how copula transformation is performed in the correct way, how to generalize copula correction for correlated exogenous regressors, close-to-normal endogenous regressors, in fixed-effects and mixed-effects panel data models (with and without slope endogeneity), how copulas should be used for moderated endogenous regressors, and pitfalls resulting from misuse of copula methods. We synthesize the literature to provide a theoretical and empirical foundation for appropriate use of copula correction methods. Given significant advances made since Park and Gupta (2012)'s study, clear guidelines for using the expanded copula correction toolbox are much needed. Thus, building upon recent advances and our eval-

uations of various copula implementations, we provide an updated guidance on when and how to use copula correction, accessible to academics and practitioners alike. We discuss boundary conditions, data requirements, and underlying identification assumptions for applying copula correction, and create a 'cookbook' for how copulas should be applied based on the latest research, in a flowchart with checkpoints of data requirements and identification assumptions that characterize the settings where copula correction methods are useful and where they may fail. We then provide two empirical examples to walk through this process of applying copula correction. Finally, we close with conclusions and implications.

THEORETICAL RATIONALE FOR ENDOGENEITY CORRECTION USING COPULAS

Why and When Use Copula Correction?

Empirical examples of endogenous regressors abound, as described in the next section. For concreteness, consider here the running example of estimating the following linear structural model using nonexperimental data:

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

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 1.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 1.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 (Villas-Boas and Winer 1999) contained in E_i, leading to the dependence between P_i and E_i (i.e., the arrow connecting E to P in Figure 1.a).

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 discrete endogenous regressors.
Copula	Control function (P&G[d], 2sCOPE[e]); Likelihood-based estimation (Haschka 2022, SORE[f]). See flowchart in Fig. 5.	Address the regressor-error dependence via copulas.	Few control variables. No requirement of ER. Either the model error or the endogenous part of error is normal. Endogenous regressors are continuous. 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. When endogenous regressors are continuous. See Table 2 for common use cases.

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 2024); f: see Qian and Xie (2024) for the SORE approach. Methods in the table can be combined as a multi-methods approach to improving the applicability, robustness and quality of causal inference.

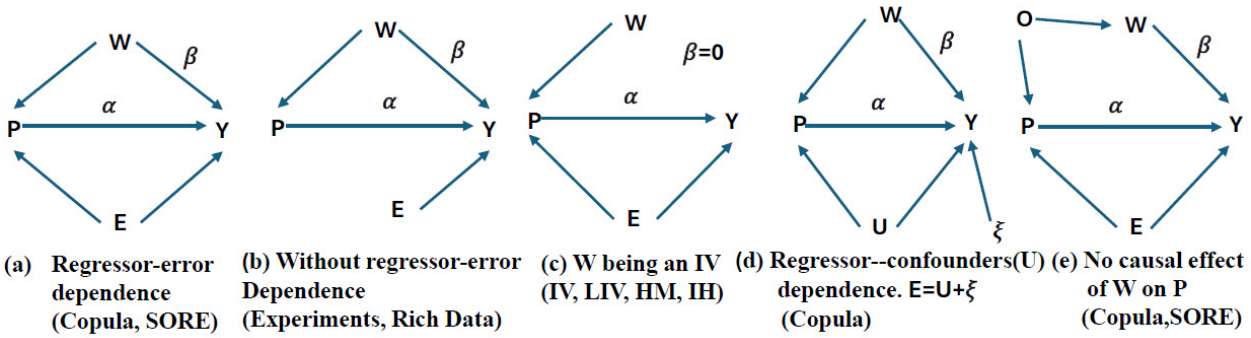

Figure 1: Directed Acyclic Graph (DAG) for Endogeneity

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 1.a explicitly includes the unobserved error term E and highlights the important role of P-E dependence. In this case, the distribution of the endogenous regressor P provides information about model parameters via its association with E. Thus, estimation methods ignoring the regressor-error dependence, such as ordinary least squares (OLS), assume the incorrect DAG in Figure 1.b and can yield severely biased model parameter estimates (i.e., endogeneity bias). Attempts to make $P - E$ independent, such as randomly assigning P via experiments or measuring and including all confounders in W, are often infeasible (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 1.a that includes the DAG in Figure 1.b as a special case and requires neither experiments nor measuring all confounders (Table 1).

As the classical approach to addressing endogeneity bias, the instrumental variable (IV) method assumes the DAG in Figure 1.c, another special case of the DAG in Figure 1.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 1.c, the untestable ER condition assuming no arrow between W and Y). The conditions of relevance and ER are typically 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 1.c) satisfying the

stringent ER condition (Park and Gupta 2012; Qian and Xie 2024). Unlike IV and these other IV-free methods, copula correction does not need to argue for any exogenous variable in W to satisfy the stringent ER condition or to causally affect P.[2] Thus, copula correction substantially increases the feasibility of endogeneity correction.

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

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 (Yang, Qian, and Xie 2024a).
- 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.

Why Does Copula Correction Work?

Copula correction first proposed by Park and Gupta (2012) is based on the idea that adequately capturing regressor-error dependence can resolve endogeneity issues and yield unbiased causal estimates. Copula correction employs Gaussian copula (GC) to link marginal distributions of regressors and the error together to obtain their joint distribution. The GC model has desirable properties, making it frequently used and widely applicable in many empirical studies to robustly capture multivariate dependence (Danaher 2007; Danaher and

[2]Empirical association between W and P (Figure 1.e) is sufficient for copula correction (Yang, Qian, and Xie 2024a).

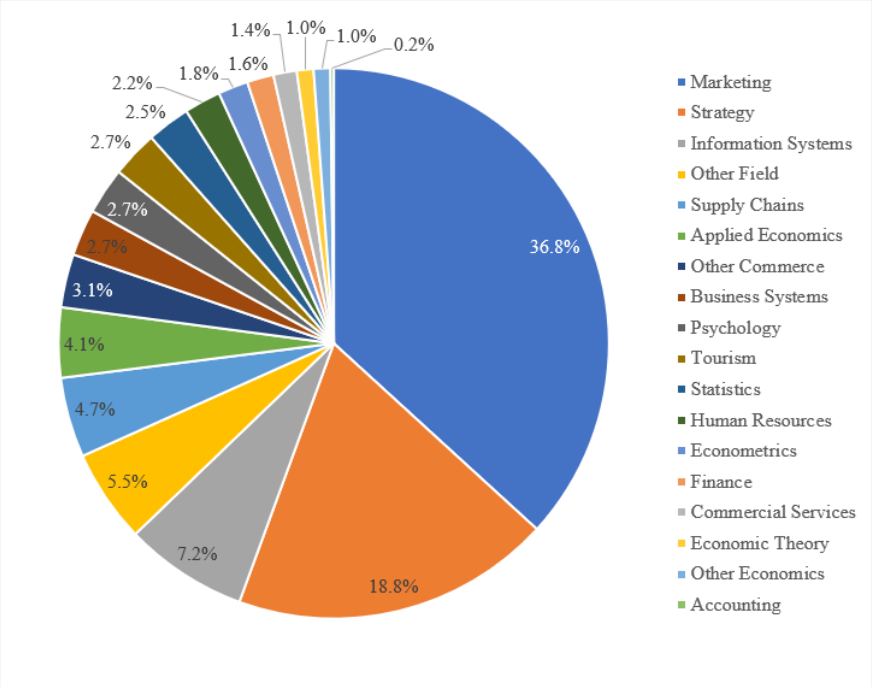

Figure 2: Disciplines using Copula Endogeneity Correction. n=511

Smith 2011). In particular, GC models with nonparametric empirical marginals capture regressor-error dependence irrespective of (potentially complex) marginal distributions while preserving the regressors' important distributional features that are critical 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 great resource in the toolkit for handling regressor endogeneity in a wide range of fields (Figure 2). In many empirical applications including those in marketing (Web Appendix Table W1), copula correction yields credible findings that are consistent with theoretical predictions, attesting to its effectiveness and applicability.

Furthermore, the later methodological background section shows that copula correction works under both data generating processes in Figure 1.a and Figure 1.d and possesses the desirable property of *double robustness*: when a GC model adequately captures either the regressor-error dependence or regressor-U dependence (where U is the endogenous part of the error term as depicted in Figure 1.d), the copula corrects endogeneity bias. Consequently, the GC regressor-error dependence assumption is only a sufficient but not necessary condition for copula correction to work. Such double robustness considerably weakens the already mild

GC regressor-error dependence assumption and increases the applicability and robustness of copula correction. Similarly, copula correction using control functions does not require the structural error to be normally distributed (Table 1), demonstrating robustness to a range of departures from the identification assumptions listed in Table 1 (Web Appendix C).

IMPACTS OF COPULA ENDOGENEITY CORRECTION

Largely due to the aforementioned advantages, copula correction has gained increasing popularity in empirical research since Park and Gupta (2012)'s study for addressing endogeneity. Although the focus here has been on marketing applications, copula correction has been extensively used in diverse fields outside marketing such as addressing potential endogeneity of a country's competitiveness measured by currency exchange rate in macroeconomic studies (e.g., Christopoulos, McAdam, and Tzavalis 2021), firms' R&D activity in finance studies (e.g., Boikos et al. 2023), and employee's feeling of control in organization management studies (e.g., Loignon et al. 2024). Figure 2 breaks down by discipline the Google Scholar citations for book chapters and journal publications (n=511) using copula endogeneity correction where each slice 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, Table 3 breaks down by various characteristics of publications that applied copula correction and appeared in leading marketing journals from 2013 to 2024 (an extensive list of papers appears in Web Appendix A).

A common use for copula correction stems from applications of the marketing mix (price, product, place, and promotion) of goods and services. A primary reason for this is such regressors are often correlated with the error term in a regression model because of uncaptured managerial knowledge in decision-making (i.e., setting prices is often related to the cost of production; advertising budgets are often set as a percentage of sales). For instance, Park and Gupta (2012) initially use copulas for price, noting "there are unmeasured product characteristics, or demand shocks, that influence not only consumer decisions but also retailer

Table 3: Publications Using Copula Correction in Leading Marketing Journals

Characteristics	Number	Characteristics	Number	Characteristics	Number
Endogenous Regressors		Outcome Type		Sample Size	
Product	19	Continuous	77	≤ 100	1
Price	31	Discrete Choice	15	101—1000	33
Place	9	Count	2	1001—5000	8
Promotion	26			5001—50000	17
Sales Force & CRM	17	Panel Data	58	≥ 50001	28
Other	33				

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 papers with their substantive areas. The total of unique journal publications is n=87.

pricing decisions" (p.582). Danaher (2023) uses copulas for price when looking at optimal advertising targeting of consumers. The concern for pricing here is that managers may set prices relative to the cost of production or perceived valuations by consumers. In their study of electronics and appliance sales, Datta et al. (2022) use copulas for line length, price, and distribution; retailers may stock more models of brands that sell better, which may get increased sales from greater distribution reach. Besides line length, product features can encompass elements like R&D spending, such as Walmart's sustainability mandate for its suppliers (Gielens et al. 2018), or in movies where the brand equity of actors may be endogenous due to the number of movie appearances, award nominations, or award wins (Mathys, Burmester, and Clement 2016). Advertising also commonly uses copulas, since managers often set advertising budgets as a percentage of sales or relative to a competitor or industry benchmark. In modeling the conversion of customers to contact an insurance agent, Guitart, Hervet, and Gelper (2020) use copulas for the focal brand's advertising, particularly in its relation to when and where the brand's primary competitor is advertising.

Another area using copula correction is salesforce and customer relationship management (CRM). Endogeneity can arise in this area because allocating particular sales personnel to particular clients or incentivizing sales personnel may be correlated with unobserved variables, like the motivation and/or ability of the sales personnel or the value of clients. Atefi et al. (2018) use copulas for salesforce training, and Burchett, Murtha, and Kohli

(2023) use copulas for salesperson's interactions with secondary items (either other people or objects like computers) when talking with customers. CRM endogeneity may occur in efforts to connect with customers, such as donation frequency and amounts (Schweidel and Knox 2013), or communications with buyers (Ludwig et al. 2022).

Copula correction can also be found in areas other than traditional marketing mix and sales efforts. A recurring explanation for the use of copula correction listed in the studies in Web Appendix Table W1 is where reverse causality or common shocks could affect the endogenous regressors. In retail research, for instance, Gijsbrechts, Campo, and Vroegrijk (2018) examine household grocery spending and use copulas for visiting hard discounters (i.e., stores with very low prices), since this becomes habit reinforcing for consumers to then spend their budget there. With social media, Fossen and Bleier (2021) use copulas to examine endogeneity when studying if online program engagement of television shows (word-of-mouth volume and deviation) affects audience size. The testing is warranted since increasing audience size may reversely cause an increase in word-of-mouth activities.

In these cases, copula correction provides a feasible approach to controlling for the thorny regressor endogeneity issue and offers opportunities for optimal managerial decision making, as further 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 diapers' category using store scanner purchase data from the IRI Academic data set for the years 2002-2006 (261 weeks) for one focal store in the Buffalo, NY market. In this instance, price was typically treated as 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 biased toward zero (i.e., less negative). As shown in a later section, the OLS price elasticity estimate in this data set is -1.367,

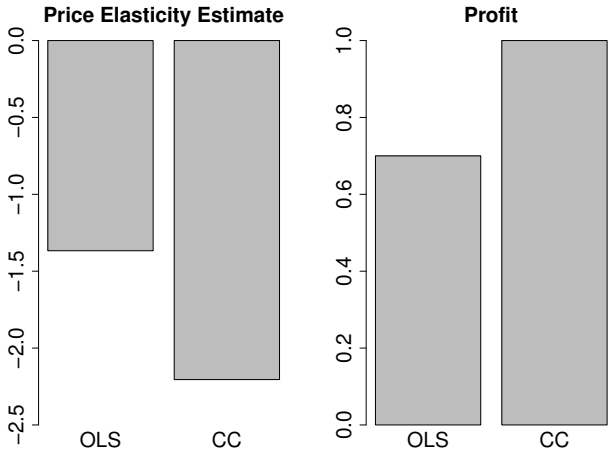

Figure 3: Example 1: Impact of copula correction on price sensitivity estimation. OLS: ordinary least squares; CC:copula correction.

which is significantly less than the price elasticity estimate of -2.205 from copula endogeneity correction (Figure 3), a 61% difference reflecting the large impact of a "wrong" estimate. Using the OLS price estimate, the manager will underestimate consumer price sensitivity and mistakenly set the price too high, resulting in lost revenue and profit. The analysis in 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).[3] 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 its correction, managers and academics likely experience under-estimated effects of pricing and advertising and over-estimated effects of salesforce.

[3]Sethuraman, Tellis, and Briesch (2011) note that the bias when not accounting for 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').

VARIATIONS IN THE USE OF COPULA CORRECTION

Appreciable variations in the use of copula endogeneity corrections exist among researchers and practitioners. These variations can substantially affect the performance of copula correction, which call for clear guidelines for optimal copula correction given the importance of endogeneity correction and the growing popularity of copula correction. Becker, Proksch, and Ringle (2022) discovered substantial bias of Park and Gupta (2012)'s copula corrected parameter estimates if the structural model contains the intercept, and cautioned the use of 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 research also shows that failure to account for exogenous regressors correlated with endogenous regressors can adversely affect copula correction's effectiveness in eliminating endogeneity bias (Haschka 2022; Qian and Xie 2024; Yang, Qian, and Xie 2024a). Originally, copula correction required sufficient nonnormality of endogenous regressors, but a recent two-stage copula correction method relaxes this requirement, and can handle endogenous regressors that are normally distributed or close-to-normal (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. Many applications in different fields are often interested in estimating structural models with higher-order terms of endogenous regressors, in order to study moderators of causal relationships or to determine optimal policy and managerial intervention (Aghion et al. 2005; Qian 2007). Considerable inconsistencies exist regarding how to handle higher-order terms of endogenous regressors in copula correction (Web Appendix Table W2). 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.

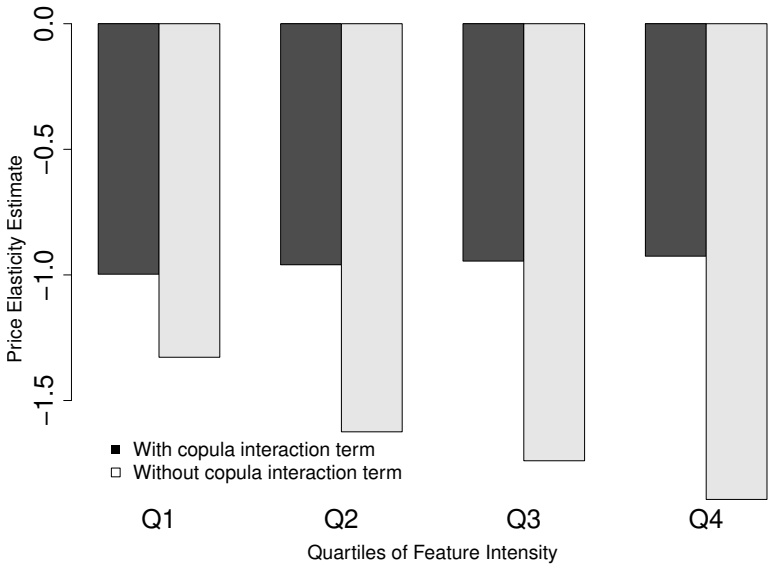

Figure 4: Mean price sensitivity estimates per quartile of feature intensity.

Example 2: Moderator of Price Sensitivity Of interest here is that price and a retail store's 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 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 presents the mean price sensitivity estimates per quartile of feature intensity for the peanut butter category, predicted from a sales demand model with an interaction term between price and feature, estimated using the IRI academic data for a store in New York City. The black (white) bars are price sensitivity estimates estimated with (without) a copula term for the interaction term. Including the copula term for the interaction term yields price sensitivity estimates that are the same across different feature intensity (meaning lack of interactive effect); excluding the copula term yields a greater magnitude of price sensitivity, and the price sensitivity estimates increase with greater feature advertisement. As shown later, adding the copula term for the interaction term can induce bias and greatly increase variability of parameter estimates.

METHODOLOGICAL BACKGROUND

In this section, we discuss the methodological aspects of the copula endogeneity correction. Our discussion aims to acquaint readers with the concepts and procedures of copula correction, to address the inconsistencies in the use of copula correction, and to inform the decision process guiding the proper use of copula correction.

Accounting for Regressor-Error Dependence Using Copula

A primer on the copula joint estimation approach

To address the endogeneity of P in Equation 1, Park and Gupta (2012) (P&G) propose two estimation methods based on a GC model for (P_i, E_i) under the assumption of a normal structural error, $E_i \sim N(0, \sigma^2)$. The first maximizes the likelihood function derived from the joint distribution of (E_i, P_i) (Park and Gupta 2012; Tran and Tsionas 2021). The second uses a generated regressor approach that is straightforward to apply and has been used in the majority of applications using copula correction. Thus, our discussion here focuses on the generated regressor approach that estimates the following augmented regression model

$$Y_i \;=\; \mu + \alpha P_i + + \beta' W_i + \gamma P_i^* + \epsilon_i, \quad \text{where} \quad P_i^* = \Phi^{-1}(F_P(P_i)); \tag{2}$$

$F_P(\cdot)$ denotes the marginal cumulative distribution function (CDF) of P, $\Phi^{-1}(\cdot)$ denotes the inverse CDF of the standard normal distribution, and γ is the coefficient parameter for P^*.

Under the GC model for (P_i, E_i), the added term P_i^* in Equation 2 captures the correlation between the endogenous regressor P and the error term E, and consequently the new error term ϵ_i in Equation 2 is independent of P_i given P_i^* in the model. Based on this result, the P&G procedure includes the copula term P_i^* as an additional control variable in the structural model to correct for the endogeneity of P. The computation of the generated regressor $P_i^* = \Phi^{-1}(F_P(P_i))$ requires an estimate of $F_P(\cdot)$, the unknown marginal CDF of the endogenous regressor P_i. The popular approach is to estimate $F_P(\cdot)$ with the empirical CDF (ECDF), $\widehat{F}_P(\cdot)$, which assigns probability mass to the uniquely observed values of P_i in the sample according to their sample frequencies.

For K continuous endogenous regressors $(P_1, \cdots, P_K)$, the P&G generated regressor ap-

proach estimates the following augmented regression model:

$$Y_i \;=\; \mu + \sum_{k=1}^{K} P_{i,k}\alpha_k + +\beta' W_i + \sum_{k=1}^{K} P^*_{i,k}\gamma_k + \epsilon_i, \quad \text{where} \quad P^*_{i,k} = \Phi^{-1}(\widehat{F}_{P_k}(P_{i,k})); \qquad (3)$$

γ_k is the coefficient parameter for P^*_k, and $\sum_{k=1}^{K} P^*_{i,k}\gamma_k$ is the linear combination of the K copula terms $\{P^*_{i,k}\}$ used to control for the endogenous regressors and thus is denoted as the copula control function (CCF).

Assumptions of the P&G procedure

For proper use of the P&G procedure, it is important to understand the assumptions behind the method. The P&G procedure makes the following assumptions.

- Assumption 1. Either the error E_i or U_i is normally distributed.
- Assumption 2. Either (P_i, E_i) or (P_i, U_i) follows a Gaussian copula.
- Assumption 3. Full rank of all regressors and $Cov(W_i, E_i) = 0$.
- Assumption 4. P_i is continuous and nonnormally distributed.
- Assumption 5: The linear combination of $P^*_{i,k}$, $\sum_{k=1}^{K} P^*_{i,k}\gamma_k$, is uncorrelated with W_i.

Contrary to the current belief, we show here that the copula control function methods (including the P&G) do not require a normal error distribution or GC regressor-error dependence and can be derived under substantially weaker conditions (Assumptions 1 and 2). In fact, the same control function procedure as above can be derived under the DAG in Figure 1.d, which decomposes the structural error as $E_i = U_i + \xi_i$, where U_i denotes the error's endogenous part representing the (mean zero) combined effects of unobserved confounders, and ξ_i is a (mean zero) exogenous disturbance term independent of U_i and all regressors. It is often plausible to assume U_i is normally distributed as a sum of many confounders' effects, satisfying Assumption 1 above. Furthermore, in many settings the GC model can adequately capture the dependence between U_i and the endogenous regressor P_i, satisfying Assumption 2 above. Then we arrive at the same augmented regression as in Equation 2, where the independent error term $\epsilon_i = \epsilon^U_i + \xi_i$, and ϵ^U_i is the remaining error term after regressing U_i on P^*_i. The message 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 in order for the augmented OLS regression to correct for bias, the identification of the P&G method does not require the structural error E_i be normally distributed or follow the GC dependence structure jointly with P_i. These weaker identification assumptions hold for other more recent copula control function methods (see Web Appendix C).

Although Assumptions 1 and 2 are used in the derivation of the generated regressors, they are not strictly required. The P&G procedure exhibits robustness to nonnormal error distributions and alternative non-Gaussian copulas (Park and Gupta 2012). Eckert and Hohberger (2023) also show that the P&G method performs on par with or better than the alternative IV estimation with a moderately skewed error distribution, but might not withstand highly skewed error distributions or arbitrary dependence structures. If a highly skewed error distribution raises questions about Assumptions 1 and 2, it is advisable to consider alternative model specifications (e.g., transforming variables).

Assumptions 3 to 5 are needed for ensuring the consistency of augmented OLS regression in Equation 3. Two important conditions are required for consistency of the augmented OLS estimates: full column rank of the regressor matrix, and zero correlation between regressors and the new error term ϵ (Wooldridge 2010). Assumption 3 is essential for common econometric methods, including OLS and IV regression. Assumption 4 is important and well established in the literature. If P approaches the normal distribution and consequently is close to a linear function of P^*, the resulting collinearity between P and P^* can lead to large standard errors and significant finite sample bias. In the extreme case when P is normally distributed, the augmented OLS regression fails by violating the full rank condition of the regressor matrix. In contrast, Assumption 5 was implicit until recently[4]. When Assumption 5 is violated, the new error term ϵ in the augmented OLS regression contains omitted variables correlated with regressors, introducing bias into estimation (Web Appendix Equation W29). Before describing recent methods that relax Assumptions 4 and 5, the next

[4]As shown in Yang, Qian, and Xie (2024a), this assumption is weaker than the assumption that exogenous and endogenous regressors are uncorrelated as suggested in Haschka (2022).

subsection discusses algorithms to produce generated regressor P^*, which can substantially affect copula correction performance.

Proper Construction of Nonparametric Rank-Based Copula Transformation

As noted above, applications of copula endogeneity correction mostly employ the nonparametric rank-based copula transformation based on the empirical marginal distributions of regressors (Equation 3). 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 construction of the empirical rank-based copula follows two steps, per Equation 3. First, the observations are ordered and mapped to a ranked percentile according to the empirical cumulative distribution, $F(\cdot)$. The second step computes the inverse normal CDF of that ranked percentile. Web Appendix Table W3 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 for the maximum value of P as follows:

$$P_i^* = \Phi^{-1}(F_P(P_i)) \quad = \quad \begin{cases} \Phi^{-1}(\texttt{Rank}(P_i)/n) & \text{if } P_i < \max(P) \\ \Phi^{-1}(n/(n+1)) & \text{if } P_i = \max(P). \end{cases} \tag{4}$$

Besides ensuring that the copula transformed values maintain the same rank order as the original regressor values for any sample size, [5], the percentile adjustment for the maximum value yields a theoretically valid maximum value of the underlying copula data. A justification of this formula 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 typical sample

[5] 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.

size (i.e., $n >> \alpha$) seen in practical studies.

To demonstrate the importance of the empirical copula transformation, consider an alternative empirical copula construction as implemented in R package `REndo` (Gui et al. 2023) and used in Becker, Proksch, and Ringle (2022) to set the percentile for the last observation to a fixed value of 0.9999999:

$$P^*_{i,Fix} = \Phi^{-1}(F_P(P_i)) \quad = \quad \begin{cases} \Phi^{-1}(\texttt{Rank}(P_i)/n) & \text{if } P_i < \max(P) \\ \Phi^{-1}(0.9999999) = 5.1999 & \text{if } P_i = \max(P), \end{cases} \tag{5}$$

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 augmented OLS regression and adversely impact 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 4 and 5 using simulation studies[6] in which the true parameter values are known. The simulation study employed the same set up as described in Becker, Proksch, and Ringle (2022) and in Web Appendix B. Data is simulated from the model $Y_i = \mu + \alpha P_i + E_i$, with a GC model between E and P and with a uniform distribution on (0,1) for P. For each simulated data, we apply both our algorithm in Equation 4 and the one in Equation 5 to obtain P^*. Park and Gupta (2012) also suggest integrating the kernel density estimate (IKDE) to obtain the marginal CDF:

$$\widehat{F}_P(p) \quad = \quad \int_{-\infty}^{p} \widehat{f}_P(u)du, \tag{6}$$

where $\widehat{f}_P(\cdot)$ is the kernel density estimate and the trapezoidal rule is used for the numerical

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

integration. We therefore also include the IKDE in the comparison.[7] For the IKDE and both ECDF algorithms, P_i^* is added as a generated regressor in the augmented OLS regression to obtain the corrected estimate of α.

The results are reported in Web Appendix Figure W2[8] and reveal that judicious handling of copula transformation is crucial for the performance of copula correction. A key finding of this study 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 4. The algorithm in Equation 4 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 Web Appendix Figure W2). Furthermore, even sample sizes as small as 100 exhibit a bias of about 0.15 for our algorithm[9], which is quite smaller than 1.0 using the algorithm in Equation 5. 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 P&G copula correction method and substantial bias in small to moderate samples. Another useful finding is that although the IKDE approach does not encounter the issue of the largest ranked value having undefined copula transformation, IDKE can experience severe estimation bias due to the boundary bias of kernel density estimation (Web Appendix B).[10]

In conclusion, our analysis provides theoretical justifications of optimal copula transformation algorithm and discovers sources of challenges in using IKDE for copula transformation. These new insights help demystify misinterpretations about copula correction and promote optimal copula transformation that greatly affects the effectiveness of copula

[7] We thank referees for the suggestion.

[8] 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.

[9] This is not surprising because the copula correction method, like instrumental variables and other IV-free methods, is a large sample procedure requiring sufficient information for satisfactory performance.

[10] Interestingly, models without intercept are robust to these methods for copula transformation. IKDE and both ECDF algorithms (Equations 4 and 5) yield unbiased estimates for models without intercept (Web Appendix Figure W3).

correction. We recommend against assigning a fixed percentile value for the largest rank, instead favoring the algorithm in Equation 4 to produce valid empirical copula construction regardless of sample size. We also caution the use of IKDE for copula transformation for the concern of potential boundary bias. 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.

Methods to Relax Assumptions of Copula Correction

The 2sCOPE control function procedure and extensions

Recent methodological developments relax key assumptions and data requirements of the P&G method, which considerably widens the applicability of copula correction (Table 4). The algorithm in Equation 4, to properly construct empirical copula transformation, should be used in these methods for optimal performance (Web Appendix B.5); however, this algorithm itself does not address the limitations of the P&G method. Yang, Qian, and Xie (2024a) propose a flexible and feasible two-stage copula endogeneity correction (2sCOPE) framework using control functions. Recent work extends 2sCOPE to enhance its capabilities and generality (Table 4). The 2sCOPE framework does not require regressor nonnormality or presume uncorrelatedness between endogenous and exogenous regressors. The methods leverage correlated exogenous regressors to sharpen structural model parameter estimates; the 2sCOPE methods include the P&G method as a special case and reduces to the P&G method when no correlated exogenous regressors exist in the model.

For the augmented OLS regression in Equation 2, the generated regressor (P^*) does not use exogenous regressors in W; this can produce biased estimates when the generated regressor P^* is correlated with the exogenous regressors in W. 2sCOPE corrects this issue by introducing a two-stage process. The idea of 2sCOPE is to remove from the P&G control function P^* the component that is correlated with the exogenous regressors, using the remaining cleaned part of P^* to control for endogeneity. Under the assumption of Gaussian

Table 4: Copula Correction Methods with Enhanced Capabilities.

Features	Methods
Handle normal endogenous regressors	2sCOPE and its extensions: Yang, Qian, and Xie (2024a,b); Liengaard et al. (2024)
Handle discrete and continuous exogenous regressors correlated with endogenous regressors and handle nonlinear terms such as interactions among regressors	Haschka (2022), SORE (Qian and Xie 2024) 2sCOPE and its extensions: Yang, Qian, and Xie (2024a,b); Liengaard et al. (2024) Breitung, Mayer, and Wied (2024)
Handle heterogeneous copula structure over levels of discrete exogenous regressors	Liengaard et al. (2024); Yang, Qian, and Xie (2024b)
Handle discrete endogenous regressors with few levels	SORE (Qian and Xie 2024)
Permit non-copula identification strategies	SORE (Qian and Xie 2024)

copula for (P, W, E) or (P, W, U) (U is the endogenous part of E in Figure 1.d), we have:

$$P_i^* = \delta' W_i^* + V_i. \tag{7}$$

where δ contains coefficient parameters, W_i^* is copula transformation of W_i, and V_i is the component of P_i^* that is unrelated to the exogenous regressors but is correlated with the structure error term E_i.[11] With a normal distribution for the error term E_i or for U_i, either (V_i, E_i) or (V_i, U_i) follows a bivariate normal distribution: the correlation coefficient captures the endogeneity of P. For instance, both E_i and V_i may contain an additive component corresponding to the combined effect of omitted variables. The above model is then obtained when the combined effect and regressors follow a GC model. One can then include the first-stage residual V_i as an additional regressor in the structural model in Equation 1 and perform the following augmented OLS regression:

$$Y_i = \mu + \alpha P_i + \beta' W_i + \gamma V_i + \omega_i. \tag{8}$$

By conditioning on the first-stage residual V_i (the component in P that causes endogeneity but uncorrelated with exogenous regressors), the new error ω_i becomes independent of all regressors (P_i, W_i, V_i), thereby ensuring the consistency of standard estimation methods.

For K continuous endogenous regressors $(P_1, \cdots, P_K)$, the 2sCOPE procedure described

[11] Although Equation 7 includes no intercept, the implementation of 2sCOPE includes the intercept, which is more general and performs well in simulation studies.

Table 5: Summary of the 2sCOPE Estimation Procedure

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 4;
- Regress $P^*_{i,k}$ on W^*_i and obtain residual $C_{i,k} = P^*_{i,k} - \delta'_k W^*_i$ (Equation 9), which removes the component related to exogenous regressors.

Stage 2:

- Add $C_{i,k}$ 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 9 (or Equation 11 when the model contains higher-order or interaction terms of regressors).

in Table 5 estimates the following augmented regression model:

$$Y_i \;=\; \mu + \sum_{k=1}^{K} P_{i,k}\alpha_k + \beta' W_i + \sum_{k=1}^{K} C_{i,k}\gamma_k + \omega_i, \quad \text{where } C_{i,k} = V_{i,k} = P^*_{i,k} - \widehat{\delta}'_k W^*_i; \quad (9)$$

$P^*_{i,k} = \Phi^{-1}(\widehat{F}_{P_k}(P_{i,k}))$, $W^*_{i,l} = \Phi^{-1}(\widehat{F}_{W_l}(W_{i,l}))$ for the lth ($l = 1, \cdots, L$) variable in W, and $\sum_{k=1}^{K} C_{i,k}\gamma_k$ is the linear combination of the K residual terms $\{V_{i,k}\}$ used to control for the endogenous regressors. The algorithm in Equation 4 is used for copula transformation of regressors including discrete exogenous regressors in W. Evaluations show 2sCOPE performs well with multiple continuous endogenous regressors and multiple exogenous regressors consisting of both continuous and discrete control covariates (Web Appendices E2 and E3 in Yang, Qian, and Xie 2024a).[12]

This two-step procedure (2sCOPE) first regresses each $P^*_{i,k}$ on W^*_i and then adds these first-stage residual terms $\{V_{i,k}\}$ to control for endogeneity. In this aspect, $\sum_{k=1}^{K} V_{i,k}\gamma_k$ serves as a control function to correct for endogeneity bias in a similar manner to the control function approach of Petrin and Train (2010). Unlike Petrin and Train (2010), 2sCOPE requires no IVs that must satisfy the stringent condition of exclusion restriction, a much stronger requirement than exogeneity. Furthermore, no arguments for the nature and direction of correlation between W and P are needed: empirical association is sufficient when using

[12] One could also eliminate discrete control covariates from the structural 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 apply Stage 1 of 2sCOPE to only group-demeaned continuous regressors and include residuals as generated regressors, while leaving the outcome unchanged. Our experience shows these approaches yield similar results.

2sCOPE. Thus, 2sCOPE greatly increases the practicality of endogeneity correction.

The 2sCOPE method 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^*. As a result, the control function in 2sCOPE accounts for correlated exogenous regressors. Second, 2sCOPE does not require endogenous regressors to have a nonnormal distribution. Even if the endogenous regressor is normally distributed, 2sCOPE can identify the model as long as one correlated W is continuous[13] and nonnormally distributed, which is feasible in many empirical applications. Third, while exogenous regressors are not used for generating the CCF 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 the CCF, 2sCOPE can eliminate P&G's finite sample bias caused by insufficient nonnormality of endogenous regressors, and increase the accuracy of the parameter estimates.

The 2sCOPE is derived based on the following assumptions:

- Assumption 1. The error E_i or its endogenous component U_i is normally distributed.
- Assumption 2. Either (P_i, W_i, E_i) or (P_i, W_i, U_i) follows a Gaussian copula.
- Assumption 3. Full rank of all regressors and $Cov(W_i, E_i) = 0$.
- Assumption 4. Either the continuous P_i or one correlated and continuous regressor in W_i is nonnormal.

Assumption 1 shows that the error term does not need to be normally distributed. Assumption 2 means that 2sCOPE continues to have the double robustness property: regressor-error dependence does not need to follow a GC relationship as long as GC adequately captures the dependence between the regressor and U_i, the endogenous part of the error. 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. Assumption 3 is not specific to 2sCOPE, but a standard

[13]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.

assumption invoked in other commonly used econometric methods, such as OLS, 2SLS using IVs, and the P&G method. When W is not exogenous (i.e., $Cov(W_i, E_i) \neq 0$), bias may arise in the coefficient estimates of endogenous regressors for all these methods (Web Appendix E.11 in Yang, Qian, and Xie 2024a). It is important to justify the exogeneity of the control variables in W for these methods (e.g., based on institutional knowledge when specifying the econometric model) or remove control variables suspected to be endogenous. Finally, Assumption 4 is less stringent than P&G's Assumption 4 (nonnormal distribution of P), while 2sCOPE eliminates Assumption 5 in P&G.

The 2sCOPE procedure assumes that the GC dependence structure is homogeneous (Assumption 2). Recent studies (Liengaard et al. 2024; Yang, Qian, and Xie 2024b) relax this assumption and provide a robustness check of 2sCOPE to the assumption. Liengaard et al. (2024) permits the GC dependence structure and the copula correction terms to vary by the levels of discrete exogenous regressors. 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 later in the Flowchart in Figure 5) within each level of combinations of discrete exogenous regressors.[14] Yang, Qian, and Xie (2024b) propose a more flexible 2sCOPE estimator based on a general-location heterogeneous GC model (see Web Appendix Table W13).

Breitung, Mayer, and Wied (2024) propose another copula correction procedure that accounts for correlated exogenous regressors. Although termed as a nonparametric control function, their approach invokes the assumptions of normality for U_i (endogenous part of the structure error E_i) and a degenerated GC dependence[15] between U_i and the error term in a

[14]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.

[15]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. Such a one-to-one deterministic relationship appears to be a strong assumption that is unlikely to hold in practice.

linear additive model, capturing the dependence of the endogenous regressor on exogenous regressors. One concern is that many endogenous regressors have complex features (e.g., bounded, truncation, discrete, or highly skewed) for which linear dependence models are known to be untrue or inadequate [16], which is why more general and plausible multivariate dependence models such as copula models are needed (Park and Fader 2004; Chen 2007; Danaher and Smith 2011; Park and Gupta 2012). We thus recommend copula correction procedures using these more flexible multivariate dependence models, such as 2sCOPE and Haschka (2022). In addition, we will describe below the SORE model (Qian and Xie 2024) and the general location copula model (Yang, Qian, and Xie 2024b), both of which nest the linear dependence models as special cases.

Likelihood-based copula correction procedures

Haschka (2022) generalizes P&G to fixed-effects (FE) linear panel data models. Because the fixed-effects transformation alters the error covariance structure, a generalized least squares (GLS) transformation is applied to address nonspherical errors and collapse panel data to pooled observations with spherical errors. Haschka (2022) then develops a copula correction method by maximizing the joint likelihood of a GC model for the error and all explanatory variables. We will detail the method in the later section "Copula Correction in Panel Data".

Qian and Xie (2024) propose an endogeneity 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 encompass a

[16] 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). Assuming linear additive models on such regressors can mismeasure dependence and produce poor model fitting and biased predictions. The linear additive equation commonly used in the first stage regression for such endogenous regressors in two-stage least squares (2SLS) using IVs is not a dependence model but simply a projection (Wooldridge 2010). 2SLS achieves identification through exclusion restriction rather than dependence modeling.

number of existing dependence models including copulas and is 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/BIC) help select proper odds ratio dependence functions, encoding regressor endogeneity and identification strategies. Distinct from other IV-free methods, SORE can handle discrete endogenous regressors with few levels (Table 4), including binary endogenous regressors or count endogenous regressors with small means, and consequently it is applicable to many applications involving these regressors. In this aspect, SORE nests as special cases the Heckman's treatment selection models (Heckman 1976) and offers alternative treatment effect identification strategies.

Optimal Copula Estimation of Endogenous Moderating and Nonlinear Effects

Many applications in different fields are interested in estimating structural models with higher-order terms of endogenous regressors to gain deeper understanding of causal mechanisms. Copula correction methods can handle these nonlinear terms (Table 4). However, considerable confusion and variation exist in how to handle these higher-order endogenous regressors. In this section we consider the best copula approach to handling these higher-order terms via both theoretical proof and empirical evaluations.

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

$$Y_i \quad = \quad \mu + \alpha_1' P_i + \alpha_2' f_1(P_i) + \alpha_3' f_2(P_i, W_i) + \beta' W_i + \eta f_3(W_i) + E_i, \tag{10}$$

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 10 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 augmented OLS regression, including only copula correction terms for the first-order endogenous terms (i.e., main effects):

$$Y_i = \mu + \alpha_1' P_i + \alpha_2' f_1(P_i) + \alpha_3' f_2(P_i, W_i) + \beta' W_i + \eta f_3(W_i) + \gamma' C_{i,main} + \epsilon_i, \quad (11)$$

where $C_{i,main} = (C_{i,1}, \cdots, C_{i,K})$ contains copula correction terms for main terms P_i only, and $C_{i,k} = V_{i,k}, k = 1, \cdots K$, are the first-stage residual terms defined in Equation 9. Because the new error term ϵ is independent of P and W under the GC model, 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 11, 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.

Although it is unnecessary to add the copula correction terms for higher-order terms,[17] 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 D.

Additionally, simulation studies in Web Appendix E demonstrate substantial harmful effects if correction terms for higher-order terms are added to control for their endogeneity. These effects include large magnitude of finite sample bias and inflated variability of

[17] 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 copula methods (see Equation 11 and Web Appendix Tables W7 and W10 for 2sCOPE) and (2) demonstrates a stronger result that adding the unnecessary high-order copula correction terms is suboptimal and has significant adverse effects using both theoretical proof and empirical evaluation.

structural model parameter estimates, as predicted by the theoretical results in the above.

Copula Correction in Panel 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

$$y_{it} = \mu_i + P_{it}'\alpha + W_{it}'\beta + e_{it}, \tag{12}$$

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 capture 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 $Cov(P_{it}, e_{it}) \neq 0$ 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_is, all regressors in (P_{it}, W_{it}) must be time-varying. Since fixed-effect parameters μ_is 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 data to the pooled case with spherical errors $\widetilde{\xi}_{it} \overset{iid}{\sim} N(0, \sigma_\xi^2)$. Haschka (2022) then 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).

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:

$$y_{it} = \mu_i + P_{it}'\alpha_i + W_{it}'\beta_i + e_{it}, \tag{13}$$

where individual-specific parameters $(\mu_i, \alpha_i, \beta_i)$ can be treated as fixed-effects, random-effects, or a mixture of fixed-effects and random-effects. The model includes the FE panel model in Equation 12 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.

To account for regressor endogeneity, Yang, Qian, and Xie (2024b) capture the regressor-error dependence using the following general location GC model that takes into account the panel data structure:

$$p_{it} = \alpha_{ip} + e_{it,p}, \quad \text{and} \quad w_{it} = \alpha_{iw} + e_{it,w}, \tag{14}$$

where the regressors p_{it} and w_{it} are allowed to depend on unit-specific mean levels α_{ip} and α_{iw}. The error terms in (13) and (14) then follow the GC model, capturing the regressor endogeneity of p_{it} and the dependence among endogenous and exogenous regressors. Assuming a homogeneous GC model, a two-stage copula control function approach estimates the following augmented panel regression model:

$$y_{it} = \mu_i + P_{it}'\alpha_i + W_{it}'\beta_i + \sum_{k=1}^{K} \gamma_k C_{it,k} + \omega_{it}, \tag{15}$$

where the copula term $C_{it,k} = (\widetilde{P}_{it,k})^* - \delta_k'(\widetilde{W}_{it})^*$; $\widetilde{P}_{it,k}$ and $\widetilde{W}_{it}$ are the time demeaned value of $P_{it,k}$ and W_{it} (i.e., subtracting each unit's averages over time of $P_{it,k}$ and W_{it} from the original values of $P_{it,k}$ and W_{it}).[18] Thus, the procedure is to apply the 2sCOPE in Table 5

[18]The time demeaning removes the effects of all time-constant confounders and is recommended 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.

to the time-demeaned regressors. The new error term ω_{it} is shown to be uncorrelated with all regressors in the augmented panel regression model in Equation 15, thereby eliminating the regressor-error dependence (Yang, Qian, and Xie 2024b). 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.[19] One can also treat slope coefficients as fixed-effects to account for slope endogeneity: dependence between regressor coefficients (α_i, β_i) and the regressors. For example, prices observed in historical data could be set by retailers with knowledge about their consumers' price sensitivity; retailers may charge lower prices in markets with greater price sensitivity. Yang, Qian, and Xie (2024b) employ the mean group (MG) estimator to estimate the augmented panel regression model in Equation 15 with slope endogeneity. Specifically, the MG estimator fits a separate augmented panel model to each panel, and then pools the estimates across all panels to obtain average estimates.

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 (Berry, Levinsohn, and Pakes 1995), for which copula correction for linear models can be directly applied to address regressor endogeneity.

Obtaining Standard Errors

For copula correction 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 procedure. For copula correction methods using two-step

[19] This general-location heterogeneous GC model (Yang, Qian, and Xie 2024b) for long panel can also be applied to grouped data formed by discrete exogenous regressors (Web Appendix F Equation W30) that generalizes Liengaard et al. (2024).

procedures, bootstrapping is applied to obtain proper standard errors in order to account for additional uncertainty associated with obtaining generated regressors in the first step (Park and Gupta 2012). 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 model estimates on the bootstrap sample. This simulation process is repeated many times to obtain a distribution for each model estimate. The standard deviation of this bootstrap distribution 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

As described in the preceding sections, considerable advances have been made since Park and Gupta (2012)'s study, with more flexible and general copula correction methods becoming available. We also show that variations in implementing copula correction have substantial impacts on its effectiveness to correct endogeneity. Informed by these findings and advances, this section describes a procedure guiding practical usage of copula correction methods.

Figure 5 presents a step-by-step flowchart[20] 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

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

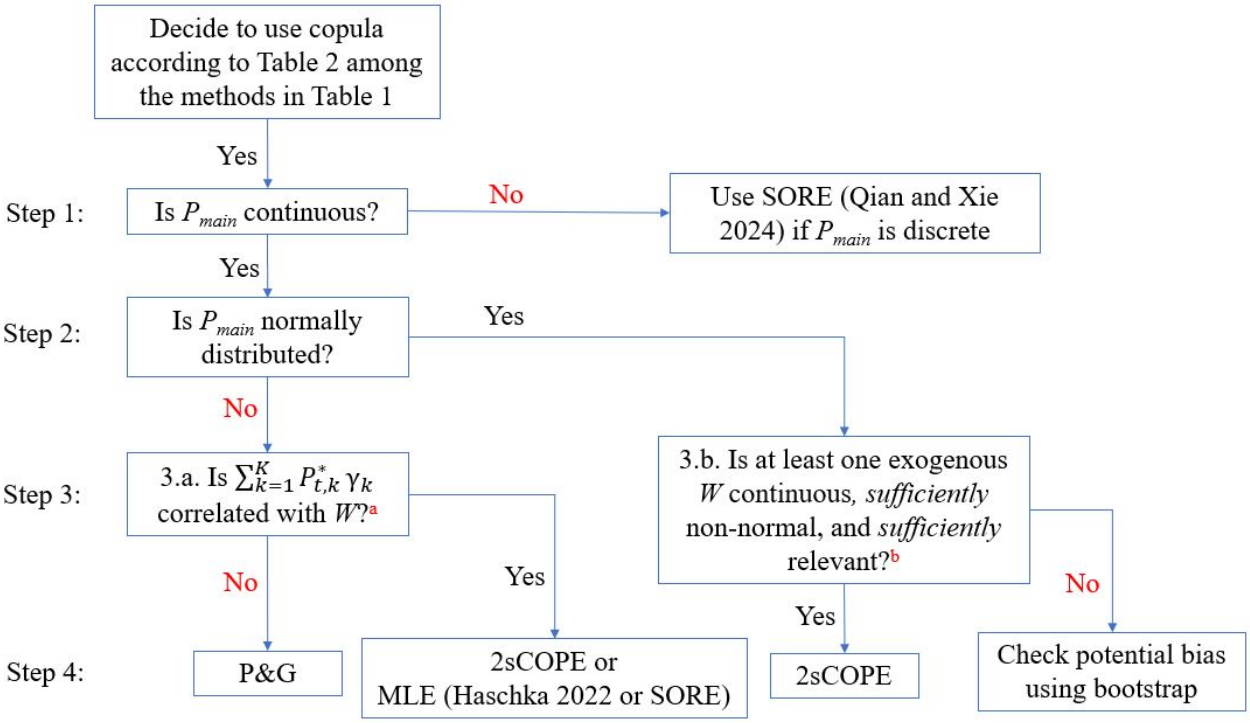

Figure 5: Flowchart for Copula Procedure.

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

[a]: For multiple endogenous regressors $(P_{main,1}, \cdots, P_{main,K})$, a less stringent condition for using P&G is no correlation between $\sum_{k=1}^{K} P_{main,k}^{*} \gamma_k$ (the linear combination of copula transformations of all the first-order endogenous regressor terms) and each W. Use the stabilized copula transformation formula in Equation 4 especially when the model includes the intercept.

[b]: W is sufficiently nonnormal if normality test $p < .001$ and sufficiently relevant to P_{main} if F statistics > 10.

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 confirmed using Table 2 (the start of the flowchart), assess the plausibility of GC dependence in the focal application. 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 the dependence between regressors and U (the combined effects of all unobserved confounders). 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 determine appropriate copula correction methods. As shown previously, copula correction only needs to include CCFs corresponding to the first-order terms P_{main} of endogenous regressors, 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 4 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. The copula procedures is designed to handle sufficiently continuous endogenous regressors. Use SORE (Qian and Xie 2024) to handle binary or discrete endogenous regressors with only a few levels, or nominal endogenous regressors whose levels have no natural ordering.

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, the 2sCOPE procedure shows even if P_{main} is normally distributed, it 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). The KS test compares the focal empirical CDF distribution - a quantity linked to copula transformation - with the reference CDF, and is an overall comprehensive measure to quantify nonnormality. Furthermore, as the performance

of copula correction improves with sample size when everything else is fixed, measures for sufficient regressor nonnormality should depend on sample size: a minor departure from normality that is insufficient for a small sample can become sufficient when sample size is large. The p-value from the KS normality test satisfies this condition. Thus, the p-value from the KS test is used to inform sufficient nonnormality of regressors.

Step 3. This step marks one of the biggest shifts in copula usage since Park and Gupta (2012), 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 to determine if the P&G method can be used. As noted previously, the P&G method requires its control function (i.e., $\sum_{k=1}^{K} P^*_{main,k}\gamma_k$ for K endogenous regressors) be uncorrelated with exogenous regressors. The correlation between P&G's control function and each exogenous regressor can be checked using Fisher's Z test for correlation. When this condition is met and sample size is small, the P&G method may be preferred because a simpler and valid model is more efficient than a more general method[21]. Otherwise, one should use 2sCOPE 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. Since P_{main} already has sufficient nonnormality, there is no need for correlated exogenous regressors to be nonnormally distributed.

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 to leverage correlated exogenous regressors to achieve model identification. In order to compensate for the lack of nonnormality of endogenous regressor P, at least one exogenous and continuous regressor W needs to satisfy the following two conditions: (1) sufficient

[21] For a large sample size, 2sCOPE has negligible effciency loss relative to P&G and is the preferred method.

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, even when P_{main} is normally distributed, 2sCOPE is expected to yield estimates with negligible bias. 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.

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 if we need to find a useful W to compensate for the lack of nonnormality of P. In 3.b, 2sCOPE 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 final step is to apply the appropriate copula procedure using either control functions or likelihood-based joint estimation. For control functions, 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).

COPULA IMPLEMENTATION EXAMPLES

In this section, we illustrate use of the flowchart to guide the implementation of copula correction 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:

$$\text{Volume}_t = \mu + \alpha P_t + \beta' W_t + E_t. \tag{16}$$

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 thus are unlikely to be correlated 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 an IV approach, 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. 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 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 the 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 the 2sCOPE method can be used to handle the price endogeneity. The

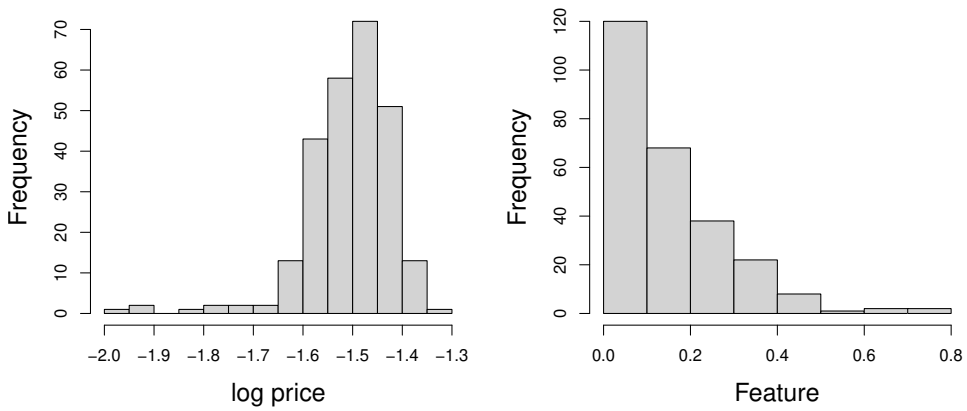

Figure 6: Distributions of Price and Feature in Example 1.

standard errors are obtained using 500 bootstrap samples.

Table 6 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 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,

Table 6: Estimation Results for Example 1

Parameters	OLS	2SLS	2sCOPE
Intercept	6.005 (0.205)***	4.371 (0.978)***	4.763 (0.668)***
Price	-1.367 (0.137)***	-2.470 (0.661)***	-2.205 (0.446)***
Feature	0.298 (0.095)***	0.059 (0.178)	0.124 (0.124)
Week	-0.002 (0.000)***	-0.002 (0.000)***	-0.002 (0.000)***
Q_2	-0.019 (0.031)	-0.014 (0.035)	-0.018 (0.036)
Q_3	-0.018 (0.032)	-0.034 (0.036)	-0.029 (0.035)
Q_4	-0.018 (0.032)	-0.061 (0.041)	-0.044 (0.035)
C_{price}			0.077 (0.037)**
ρ			0.366 (0.160)**

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

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

We now examine 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 16 to include an interaction term ($P_t * F_t$) between price and feature as follows:

$$\text{Volume}_t \quad = \quad \mu + \alpha * P_t + \beta' W_t + \phi P_t * F_t + E_t, \tag{17}$$

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

Table 7: 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 $p < 0.01$

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. Web Appendix F describes the walk-through of these steps, which concludes that 2sCOPE should be used.[22]

Table 7 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 also compare the results to OLS and 2SLS, as well as to 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

[22]Web Appendix Table W12 presents the results of the P&G method from the two examples, even though P&G did not satisfy the data requirements according to the flowchart in Figure 5. The results there show appreciable differences in some model estimates, more so for Example 2.

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 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 7), 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 W15) and lead to substantially varied and biased estimates.

Managerial and Academic Implications

The two examples highlight both 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 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 the sale 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 approximately 100% 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 has 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).

Example 2 presented the case of the interaction between an endogenous and exogenous regressor. Like Example 1, price elasticity in the absence of feature was substantially underestimated 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. 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 7). 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, as demonstrated in Example 1.

CONCLUSION

The instrument-free copula correction has been increasingly used to address endogeneity bias given its practical advantages and feasible implementation. Yet, like all other causal estimation procedures designed for use with nonexperimental data, the validity of copula correction requires correct implementation of the method and demands boundary conditions and data requirements to be met in its empirical applications.

This study contributes to the field in three areas. One, we discuss 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 (Table 1 and 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 is optimal and considerably outperforms including these copula terms. To our knowledge, these are the first theoretical results justifying the optimal implmentation 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 requirements of the prior copula correction approach are not met by the data at hand. For cases where the endogenous regressors have insufficient nonnormality, and the traditional method (Park and Gupta 2012) 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 related and nonnormally distributed 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., Yang, Qian, and Xie 2024b; Liengaard et al. 2024; 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; Hu, Qian, and Xie 2025), improving computational efficiency especially for computationally intensive procedures (e.g., the MLE procedures), to name a few. Hu, Qian, and Xie (2025)

propose a two-stage nonparametric copula control function (2sCOPE-np) that generalizes and robustifies the existing copula correction methods. Another research direction is the new empirical applications of copula correction. A great variety of quantitative models are used in empirical studies with new ones emerging constantly. Opportunities to adapt copula correction to new types of data or models abound.

REFERENCES

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

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

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.

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.

Berry, Steven, James Levinsohn, and Ariel Pakes (1995), "Automobile Prices in Market Equilibrium,," *Econometrica*, 63(4), 841–90.

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.

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

Boikos, Spyridon, Ioannis Bournakis, Dimitris Christopoulos, and Peter McAdam (2023), "Financial reforms and innovation: A micro–macro perspective," *Journal of International Money and Finance*, 132, 102820.

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.

Burchett, Molly R, Brian Murtha, and Ajay K Kohli (2023), "Secondary Selling: Beyond the Salesperson–Customer Dyad," *Journal of Marketing*, 87(4), 575–600.

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. (2007), "Modeling Page Views Across Multiple Websites with An Application to Internet Reach and Frequency Prediction," *Marketing Science*, 26(3), 422–437.

Danaher, Peter J (2023), "Optimal microtargeting of advertising," *Journal of Marketing Research*, 60 (3), 564–584.

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).

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.

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

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.

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.

Gui, Raluca, Markus Meierer, Patrik Schilter, and René Algesheimer (2023), "REndo: Internal Instrumental Variables to Address Endogeneity," *Journal of Statistical Software*, 107, 1–43.

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.

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 SSRN: https://ssrn.com/abstract=4235194 or http://dx.doi.org/10.2139/ssrn.4235194*.

Heckman, James J (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5(4), 475–492.

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.

Loignon, Andrew C, Michael A Johnson, Marlies Veestraeten, and Terrance L Boyd (2024), "A tale of two offices: The socioeconomic environment's effect on job performance while working from home," *Group & Organization Management*, 49 (1), 183–214.

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.

Manchanda, Puneet, Peter E Rossi, and Pradeep K Chintagunta (2004), "Response Modeling with Nonrandom Marketing-mix Variables," *Journal of Marketing Research*, 41 (4), 467–478.

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.

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*.

Park, Y. and P.S. Fader (2004), "Modeling Browsing Behavior at Multiple Websites," *Marketing Science*, 23(3), 280–303.

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 (2007), "Do National Patent Laws Stimulate Domestic Innovation in a Global Patenting Environment? A Cross-Country Analysis of Pharmaceutical Patent Protection, 1978-2002," *The Review of Economics and Statistics*, 89, 436–453.

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.

Schweidel, David A and George Knox (2013), "Incorporating direct marketing activity into latent attrition models," *Marketing Science*, 32 (3), 471–487.

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.

Sudhir, Karunakaran (2001), "Competitive Pricing Behavior in the Auto Market: A Structural Analysis," *Marketing Science*, 20, 42–60.

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

These materials have been supplied by the authors to aid in the understanding of their paper. The AMA is sharing these materials at the request of the authors.

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		

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

List of Publications in Table W1

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.

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.

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.

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.
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.
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.
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.

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: 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: 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 interfaced 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.

Simulation Study Setup

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

$$\begin{bmatrix} E_t^* \\ P_t^* \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.50 \\ 0.50 & 1 \end{bmatrix}\right) \tag{W1}$$

$$E_t = \Phi^{-1}(\Phi(E_t^*)) \tag{W2}$$

$$P_t = \Phi(P_t^*) \tag{W3}$$

$$Y_t = \mu + \alpha P_t + E_t = -1P_t + E_t, \tag{W4}$$

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, 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 5 to obtain generated regressor, the P&G method using the algorithm in Equation 4, 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 W2 for α (discussed further below).

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 W3. The construction of the empirical rank-based copula follows two steps, per Equation 3. 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 W3 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 4. Besides ensuring that the copula transformed values maintain the same rank order as the original regressor values for any sample size [23], 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 W3: 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.

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

[23] 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.

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

$$\widehat{f}_P(P = p) \quad = \quad \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{p - P_i}{b}\right), \tag{W5}$$

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

$$\widehat{F}_P(p) \quad = \quad \int_{-\infty}^{p} \widehat{f}_P(u)du, \tag{W6}$$

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 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

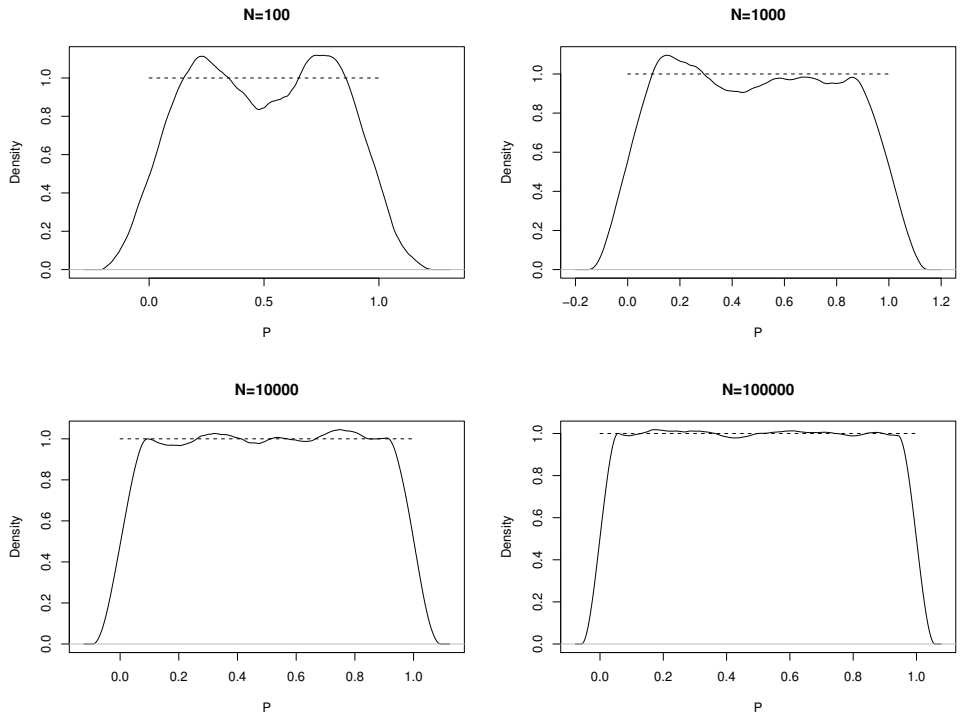

Figure W1: Boundary Bias of Nonparametric Kernel Density Estimates. Dotted line denotes the true density function of the uniform distribution on [0, 1]. Solid line denotes the KDE estimates.

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 W1 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.

Figure W2 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 W2 x-axis). OLS, as the curve with circles in Figure W2, exhibits substantial bias (> 1.5) in the coefficient estimate α for

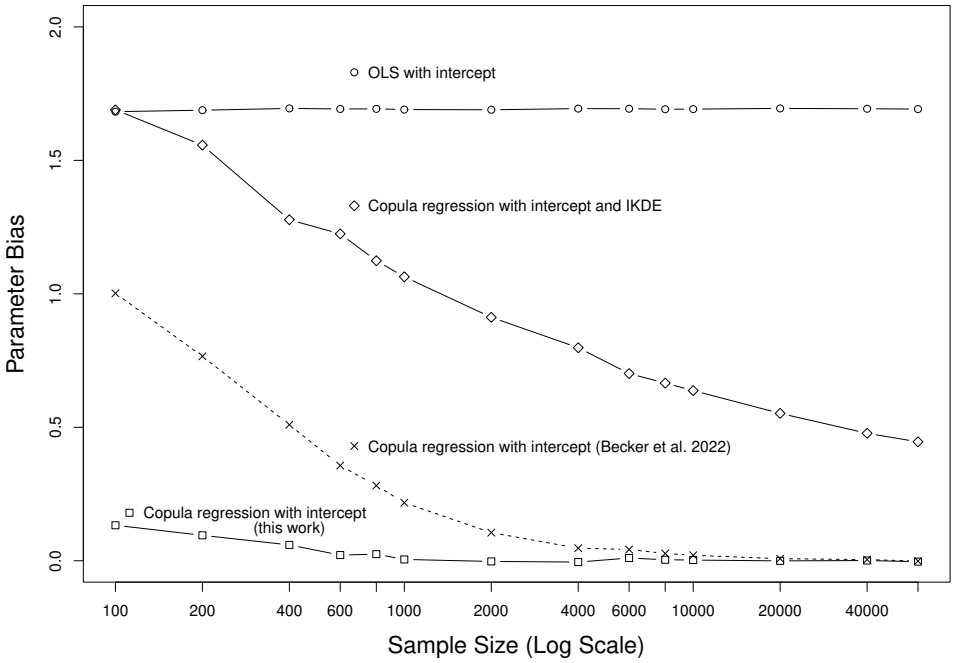

Figure W2: Bias of the endogenous regressor.

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 5 (the curve with cross marks in Figure W2) 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 W2 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 4. The algorithm in Equation 4 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 W2). Furthermore, even sample sizes as small as 100 exhibit a bias of about 0.15 for our algorithm[24], which is quite smaller than 1.0 using the algorithm in Equation 5.

Figure W2 examines the impacts of boundary bias of IKDE on copula correction using the same DGP as specified in Equations 1-4 in Becker, Proksch, and Ringle (2022) (i.e., Equations W1 to W4). 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 W2 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 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 W4 without intercept. All settings remain the same as those when estimating the models with

[24]This is not surprising because the copula correction method, like instrumental variables and other IV-free methods, is a large sample procedure requiring sufficient information for satisfactory performance.

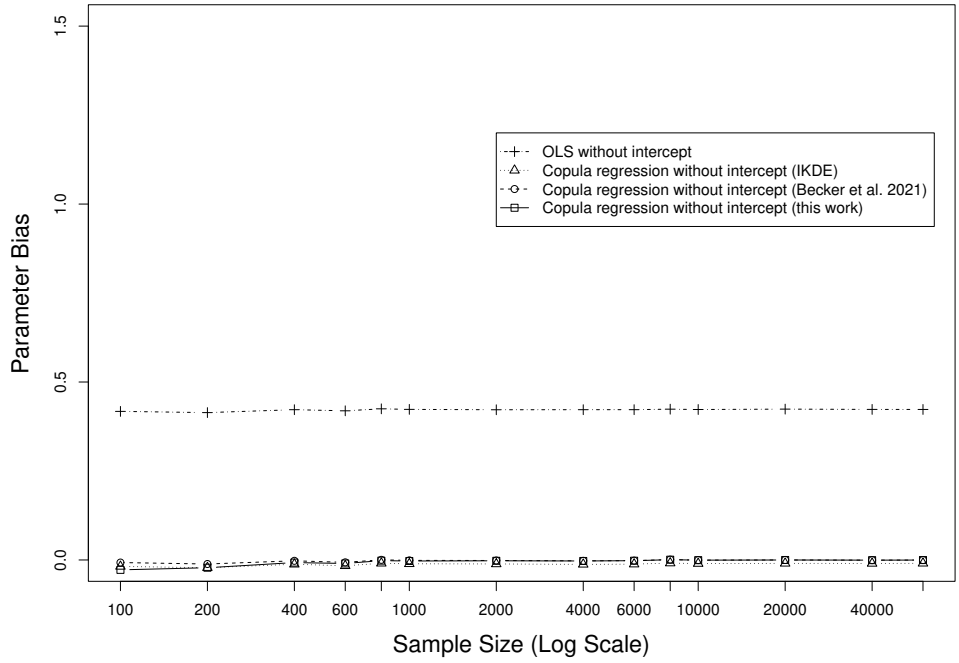

Figure W3: Bias of the endogenous regressor without intercept.

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, 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 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 data generating process (DGP) is summarized below:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ E_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix}\right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}\right), \quad \text{(W7)}$$

$$E_t = G^{-1}(U_{E,t}) = G^{-1}(\Phi(E_t^*)) = \Phi^{-1}(\Phi(E^*)) = 1 \cdot E_t^*, \quad \text{(W8)}$$

$$P_t = H^{-1}(U_{P,t}) = H^{-1}(\Phi(P_t^*)), \quad W_t = L^{-1}(U_{W,t}) = L^{-1}(\Phi(W_t^*)), \quad \text{(W9)}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + E_t = 0 + (-1) \cdot P_t + 1 \cdot W_t + E_t, \quad \text{(W10)}$$

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 (W10); (2) 2sCOPE using the fixed algorithm for copula transformation of P and W (Equation 5); (3) 2sCOPE using the adaptive algorithm for copula transformation of P and W (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.

Figure W4 shows that the fixed algorithm also negatively affects the performance of copula correction, while the adaptive algorithm avoids the bias.

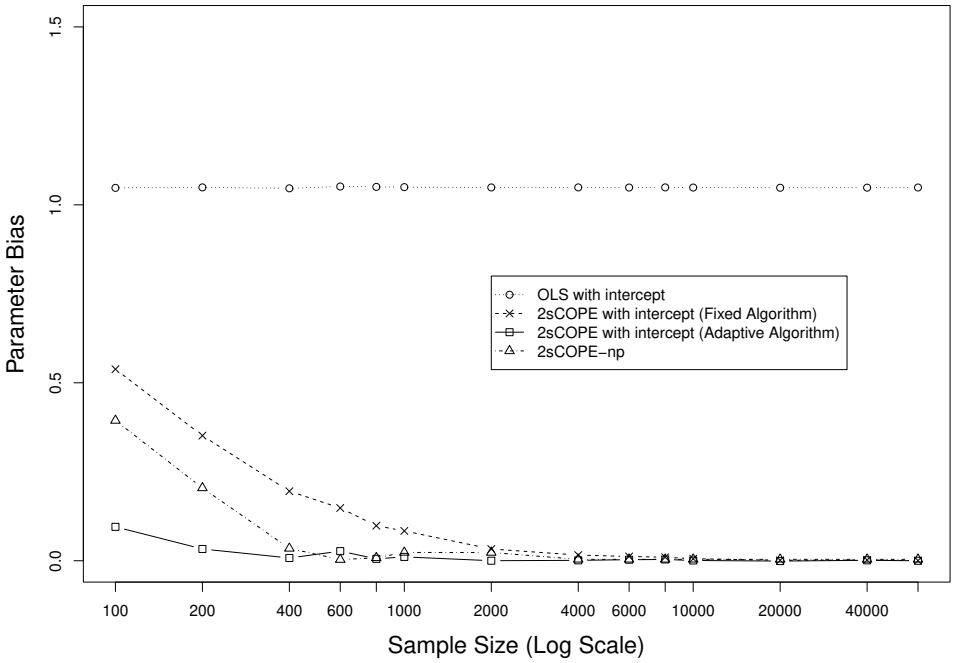

Figure W4: Method comparison with correlated endogenous and exogenous regressors

WEB APPENDIX C: DOUBLE ROBUSTNESS PROPERTY OF COPULA CORRECTION

This section demonstrates the double robustness of copula correction using control function such that 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 1.d:

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i \tag{W11}$$

$$E_i = U_i + \xi_i \tag{W12}$$

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 zero means.

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 W7 to W9 (i.e., replacing E_t with U_t in these equations). 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_i using Equation W11 with parameters values given in Table W4. Sample size is set to n=1,000 per dataset. For each dataset, we apply OLS and the 2sCOPE estimation described in Table 5. A total of 1,000 datasets were generated.

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
U[-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

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

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
U[-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

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 shows 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.

WEB APPENDIX D: 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:*

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

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:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \;\; V(\boldsymbol{\epsilon}) = \sigma_c^2 \mathbf{I}_n, \tag{W13}$$

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:

$$V(\widehat{\boldsymbol{\theta}}^{Main}) = \sigma_c^2 (\mathbf{X}'\mathbf{X})^{-1}. \tag{W14}$$

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

Equation (W13), we have:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{C}\boldsymbol{\phi} + \boldsymbol{\epsilon_1}, \;\; V(\boldsymbol{\epsilon_1}) = \sigma_c^{\prime 2}\mathbf{I}_n, \tag{W15}$$

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

$$\widehat{\boldsymbol{\theta}}^{All} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{C}\widehat{\phi}), \qquad \widehat{\phi} = (\mathbf{C}'\mathbf{R}\mathbf{C})^{-1}\mathbf{C}'\mathbf{R}\mathbf{Y}, \tag{W16}$$

$$V(\widehat{\boldsymbol{\theta}}^{All}) = \sigma_c^{\prime 2}\big[(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} + \mathbf{M}(\mathbf{C}'\mathbf{R}\mathbf{C})^{-\mathbf{1}}\mathbf{M}'\big], \tag{W17}$$

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:

$$V(\widehat{\boldsymbol{\theta}})^{All} - V(\widehat{\boldsymbol{\theta}})^{Main} \approx \sigma_c^2\big[(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} + \mathbf{M}(\mathbf{C}'\mathbf{R}\mathbf{C})^{-\mathbf{1}}\mathbf{M}' - (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\big] \tag{W18}$$

$$= \sigma_c^2\big[\mathbf{M}(\mathbf{C}'\mathbf{R}\mathbf{C})^{-\mathbf{1}}\mathbf{M}'\big]. \tag{W19}$$

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:

$$\mathtt{Var}(\widehat{\theta}_k^{All}) \geq \mathtt{Var}(\widehat{\theta}_k^{Main}) \quad \text{for } k = 1, \cdots, K. \tag{W20}$$

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 E: 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:

$$Y = \alpha_0 + \alpha_1 P_1 + \alpha_2 P_2 + \alpha_3 P_1 * P2 + E \tag{W21}$$

$$\begin{pmatrix} E^* \\ P_1^* \\ P_2^* \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho_{E1} & \rho_{E2} \\ \rho_{E1} & 1 & \rho_{12} \\ \rho_{E2} & \rho_{12} & 1 \end{bmatrix}\right)$$

$$E = H_E^{-1}(\Phi(E^*)) = \Phi^{-1}(\Phi(E^*)), \qquad P_1 = H_{P_1}^{-1}(\Phi(P_1^*)), \quad P_2 = H_{P_2}^{-1}(\Phi(P_2^*)). \tag{W22}$$

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:

$$C_{P_1} = \Phi^{-1}(\hat{H}_{P_1}(P_1)), \qquad C_{P_2} = \Phi^{-1}(\hat{H}_{P_2}(P_2)). \tag{W23}$$

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

$$C_{P_1*P_2} \quad = \quad \Phi^{-1}(\hat{H}_{P_1*P_2}(P_1 * P_2)) \tag{W24}$$

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 W6. 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 W6 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 W6, 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 W6: 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
5000	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
50000	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.

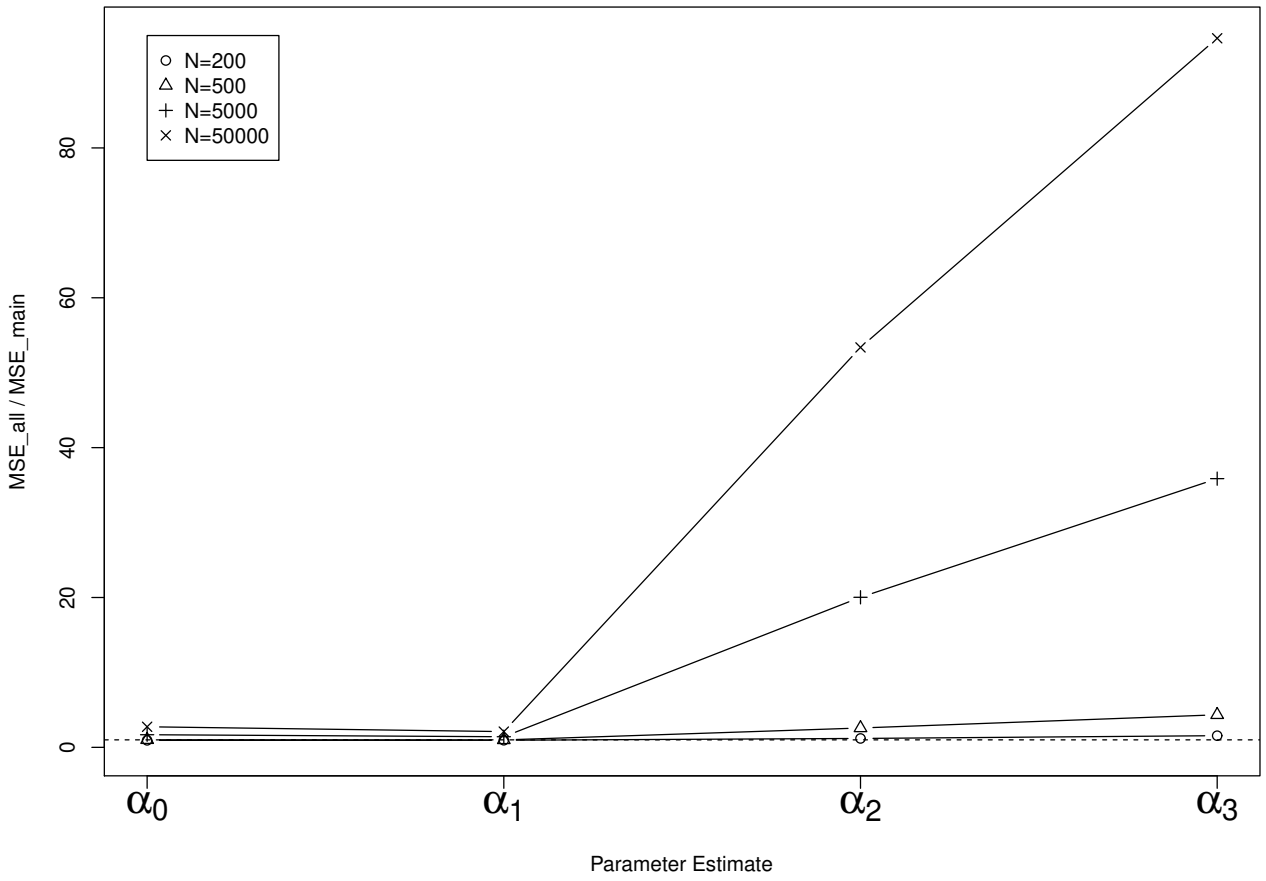

Figure W5: Ratio of mean squared errors of structural model estimates, with using the copula interaction term (Copula-All) to those without using the copula interaction term (Copula-Main).

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:

$$\begin{aligned} Y &= \alpha_0 + \beta_1 W + \alpha_1 P + \alpha_2 W * P + E \\ \begin{pmatrix} P^* \\ W^* \\ E^* \end{pmatrix} &= N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix}\right) \\ E = H_E^{-1}(\Phi(E^*)) &= \Phi^{-1}(\Phi(E^*)), \qquad P = H_P^{-1}(\Phi(P^*)), W = L_W^{-1}(\Phi(W^*)) \end{aligned} \quad \text{(W25)}$$

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 11) 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 W7.

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 W7) 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 W7). The results demonstrate the substantial adverse effects of adding unnecessary copula terms for interactions: significant finite sample estimation bias and inflated standard errors.

Table W7: 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
5000	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
50000	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 (subscript t omitted for simplicity):

$$\begin{aligned} Y &= \alpha_0 + \alpha_1 P + \alpha_2 P^2 + E, \\ \begin{pmatrix} E^* \\ P^* \end{pmatrix} &= N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \\ E = H_E^{-1}(\Phi(E^*)) &= \Phi^{-1}(\Phi(E^*)), \qquad P = H_P^{-1}(\Phi(P^*)), \end{aligned} \tag{W26}$$

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:

OLS:	P, P^2
Copula-Main:	P, P^2, C_P
Copula-All:	P, P^2, C_P, C_{P^2}

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 W8 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 W8: 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
5000	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
50000	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).

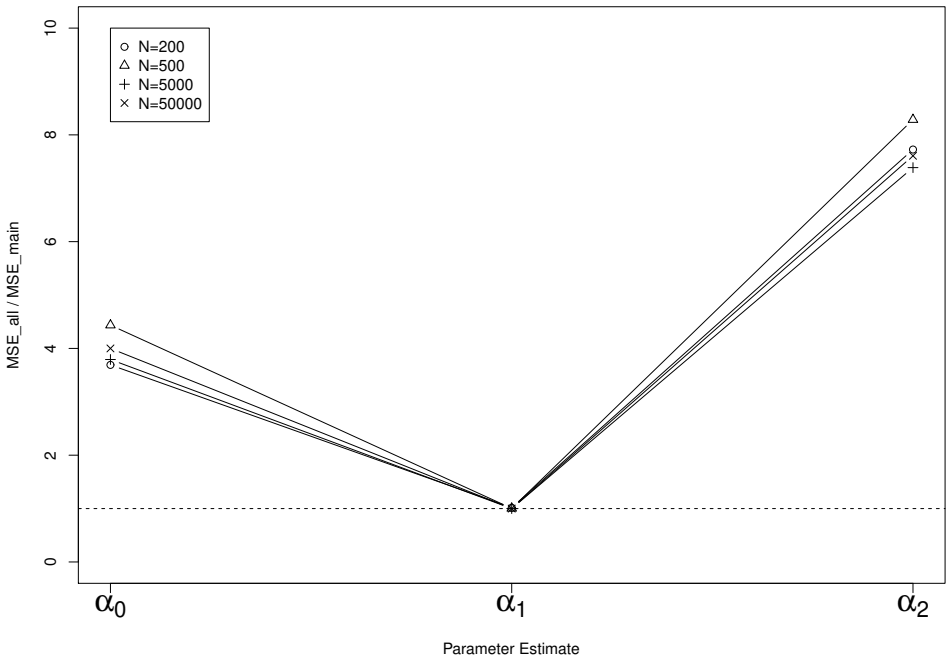

Figure W6: Ratio of mean squared errors of structural model estimates, with using the copula square term (Copula-All) to those without using the copula square term (Copula-Main).

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).

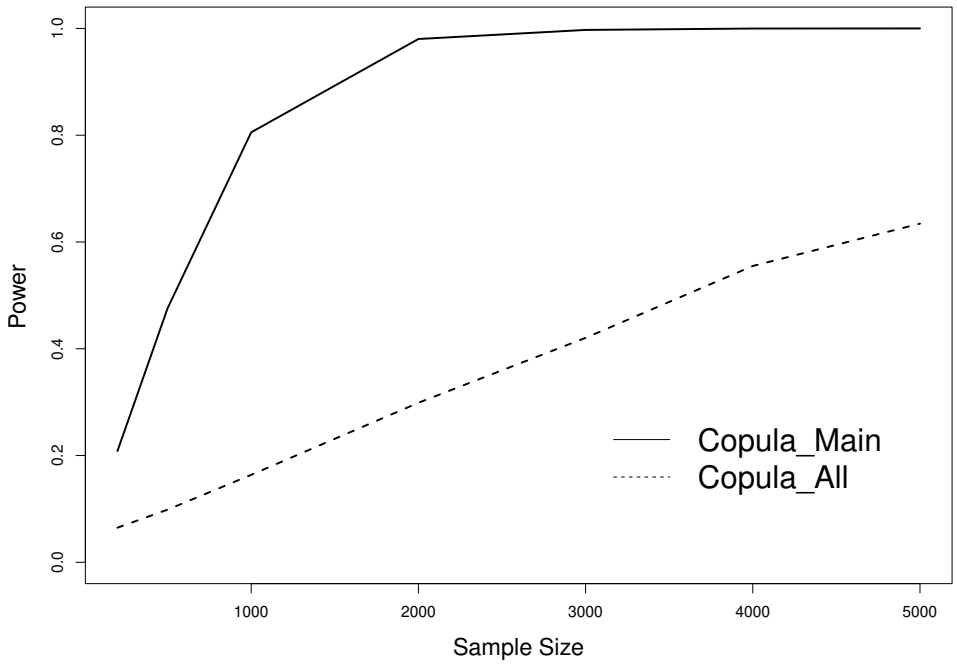

Figure W7: Statistical Power to detect the squared term P^2 with the copula squared term (Copula-All) and without the copula squared term (Copula-Main).

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:

$$Y = \alpha_0^c + \alpha_1^c(P_1 - \bar{P}_1) + \alpha_2^c(P_2 - \bar{P}_2) + \alpha_3^c(P_1 - \bar{P}_1) * (P_2 - \bar{P}_2) + E, \quad \text{(W27)}$$

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

$$\begin{aligned} \alpha_0^c &= \alpha_0 + \alpha_1\bar{P}_1 + \alpha_2\bar{P}_2 + \alpha_3\bar{P}_1\bar{P}_2 \\ \alpha_1^c &= \alpha_1 + \alpha_3\bar{P}_2 \\ \alpha_2^c &= \alpha_2 + \alpha_3\bar{P}_1 \\ \alpha_3^c &= \alpha_3. \end{aligned} \quad \text{(W28)}$$

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 W28 also holds 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 W28 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 W6, W7, and W8, 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 W9, W10, and W11, respectively. The true values for the parameters in the models after mean-centering are also listed in Tables W9 to W11. 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 W6, W7, and W8, 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 W9: 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)	(0.281)	(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)	(0.173)	(0.078)	0.0293
5000	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)	(0.052)	(0.024)	0.0028
50000	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)	(0.017)	(0.008)	0.0003

See the same note under Table W8.

Table W10: 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
5000	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
50000	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 W7.

Table W11: 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
5000	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
50000	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 W8.

WEB APPENDIX F: ADDITIONAL MATERIALS FOR THE IMPLEMENTATION EXAMPLES

In Example 2, 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.

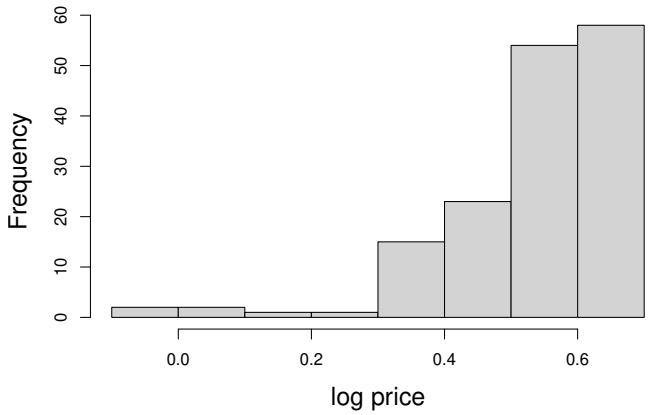

Figure W8: Price Distribution in Example 2.

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.

Although the P&G method was not selected in both examples according to the flowchart in Figure 5, Table W12 presents the results of applying P&G methods to the two implementation examples. In Example I, 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 II. 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 W12).

Theoretically, the bias of 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

Table W12: Estimation Results

	Example I		Example II	
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 $p < 0.01$

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:

$$\sigma \frac{-q\rho}{1-q^2} \left[\texttt{Cov}(P, W^*) / \texttt{Var}(P) \right], \tag{W29}$$

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 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. 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 W13) with the following augmented regression model

$$Y_i = \mu + \sum_{k=1}^{K} P_{i,k}\alpha_k + \beta' W_i + \sum_{k=1}^{K} \left\{ C_{i,k}\gamma_{k0} + \sum_{j=1}^{G-1} C_{i,k} I(g_i(w_d) = j) * \gamma_{kj} \right\} + \omega_t, \quad \text{(W30)}$$

$$\text{where } C_{i,k} = (\widetilde{P}_{i,k})^{*|g_i(W_d)} - \delta'_{g_i(W_d),k}(\widetilde{W}_{c,i})^{*|g_i(W_d)}. \quad \text{(W31)}$$

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 W31 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 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}.

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 I, 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 $F > 10$. So this means data in Example I do not satisfy the data requirement for 2sCOPE-HGC while the 2sCOPE meets data requirements. In Example II, 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 W14. We observe that 2sCOPE-HGC yielded results that largely agree with 2sCOPE 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 W13: 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 W31).

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 W30).

Table W14: 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 W15: 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.

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 Share-

holder 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.

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.

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.