Journal of the Academy of Marketing Science (2025) 53:279–299
https://doi.org/10.1007/s11747-024-01055-4

METHODOLOGICAL PAPER

Dealing with regression models' endogeneity by means of an adjusted estimator for the Gaussian copula approach

Benjamin D. Liengaard[1] · Jan-Michael Becker[2] · Mikkel Bennedsen[1] · Phillip Heiler[1] · Luke N. Taylor[1] · Christian M. Ringle[3,4]

Received: 12 June 2023 / Accepted: 22 September 2024 / Published online: 30 October 2024

Abstract

Endogeneity in regression models is a key marketing research concern. The Gaussian copula approach offers an instrumental variable (IV)-free technique to mitigate endogeneity bias in regression models. Previous research revealed substantial finite sample bias when applying this method to regression models with an intercept. This is particularly problematic as models in marketing studies almost always require an intercept. To resolve this limitation, our research determines the bias's sources, making several methodological advances in the process. *First*, we show that the cumulative distribution function estimation's quality strongly affects the Gaussian copula approach's performance. *Second*, we use this insight to develop an adjusted estimator that improves the Gaussian copula approach's finite sample performance in regression models with (and without) an intercept. *Third*, as a broader contribution, we extend the framework for copula estimation to models with multiple endogenous variables on continuous scales and exogenous variables on discrete and continuous scales, and non-linearities such as interaction terms. *Fourth*, simulation studies confirm that the new adjusted estimator outperforms the established ones. Further simulations also underscore that our extended framework allows researchers to validly deal with multiple endogenous and exogenous regressors, and the interactions between them. *Fifth*, we demonstrate the adjusted estimator and the general framework's systematic application, using an empirical marketing example with real-world data. These contributions enable researchers in marketing and other disciplines to effectively address endogeneity problems in their models by using the improved Gaussian copula approach.

Keywords Endogeneity · Bias · Estimator · Intercept · Gaussian copula · Regression model · Guidelines

Introduction

Endogeneity in regression models is a fundamental marketing research concern (e.g., Jean et al., 2016; Sande & Ghosh, 2018; Shugan, 2004; Zaefarian et al., 2017). Park and Gupta (2012) make a key contribution to addressing endogeneity problems by introducing the Gaussian copula approach as an instrumental variable (IV)-free method to reveal and reduce endogeneity bias in regression models. This is done by parameterizing the error term and the endogenous

Bulent Menguc served as Area Editor for this article.

✉ Christian M. Ringle
c.ringle@tuhh.de

Benjamin D. Liengaard
benlien@econ.au.dk

Jan-Michael Becker
jan-michael.becker@bi.no

Mikkel Bennedsen
mbennedsen@econ.au.dk

Phillip Heiler
pheiler@econ.au.dk

Luke N. Taylor
lntaylor@econ.au.dk

[1] Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, Aarhus V 8210, Denmark

[2] Department of Marketing, BI Norwegian Business School, Nydalsveien 37, Oslo 0484, Norway

[3] Department of Management Sciences and Technology, Hamburg University of Technology, Am Schwarzenberg-Campus 4, 21073 Hamburg, Germany

[4] College of Business, Law and Governance, James Cook University, Townsville, Australia

regressor's joint distribution by means of a copula. More specifically, the Gaussian copula approach decomposes the endogenous independent variable and the error term's joint distribution into their marginal distributions, which are thereafter used as the basis to calculate a control or likelihood function. Due to its straightforward way of identifying and correcting endogeneity bias in regression models (Rutz & Watson, 2019), the method is now frequently applied in empirical applications in premier marketing journal publications but also in other disciplines (Becker et al., 2022; Park & Gupta, 2024).[1]

For the Gaussian copula approach to be valid, it must meet specific identification requirements (Park & Gupta, 2012): (1) the error term distribution must be normal, (2) the endogenous regressor distribution must be non-normal, and (3) the copula dependency structure must be Gaussian (see also Park & Gupta, 2024). While the first and third requirements (i.e., the error term distribution and the Gaussian copula correlation structure) are inherently unobservable and cannot therefore be directly evaluated, researchers can evaluate the second requirement (the regressor's non-normality) to ensure that the Gaussian copula approach is applied appropriately (Becker et al., 2022; Haschka, 2022).

The Gaussian copula approach is deemed advantageous because researchers often have difficulty finding suitable IVs, which are also available in empirical studies (Falkenström et al., 2023). Despite the potential benefits associated with the Gaussian copula approach, a substantial finite sample bias and low statistical power are critical problems if the model includes an intercept (Becker et al., 2022). Simply estimating a model without an intercept is not a solution because the results can also be heavily biased if the data-generating process (DGP) requires an intercept but it is omitted from the model estimation (Becker et al., 2022). Consequently, current recommendations regarding applying the Gaussian copula approach to regression models with an intercept require a large sample size and the endogenous regressor's strong non-normality, especially if the skewness is high (Becker et al., 2022; Eckert & Hohberger, 2023). These demanding requirements pose a threat to the approach's valid application in practice. This is particularly problematic, since almost all regression models include an intercept in empirical studies (e.g., see the review presented by Becker et al., 2022). Also, variables' mere standardization does not solve this problem due to the copula approach's intrinsic non-linearities.

[1] The literature review by Becker et al. (2022) reveals 69 publications that use the Gaussian copula approach by the end of 2020, whereby 40 of the these applications (58%) appeared in premier marketing journals such as International Journal of Research in Marketing (11), Journal of the Academy of Marketing Science (7), Journal of Marketing Research (6), Journal of Retailing (6), and Journal of Marketing (5).

Prior research (e.g., Becker et al., 2022; Eckert & Hohberger, 2023; Falkenström et al., 2023; Haschka, 2022) cannot answer the question of why regression models with an intercept perform poorly, while those without an intercept perform extremely well when the Gaussian copula approach is applied. Our research explains the difference between models with and without an intercept, and provides analytical derivations regarding the bias's sources. Further, these insights allow us to make an important research contribution by developing an adjusted empirical cumulative distribution function (ECDF) that reduces the finite sample bias when the Gaussian copula approach is applied to regression models with an intercept.

Another open methodological research question pivotal for empirical studies pertains to the Gaussian copula approach's ability to deal with multiple endogenous and exogenous regressors with different types of scales as well as non-linearities (e.g., interaction effects). Haschka (2022) reveals that (in addition to the three initial assumptions by Park & Gupta, 2012) the Gaussian copula approach has a *fourth requirement*. The original approach requires uncorrelated endogenous and exogenous regressors to ensure unbiased estimation of the endogenous regressor's coefficient. Based on this finding, Haschka (2022) develops a modified maximum likelihood-based estimator for panel data that takes the correlations between endogenous and exogenous variables into account. Yang et al. (2022) build on this finding by employing a control function technique and proposing a two-stage estimator that also takes the correlation between continuous endogenous and exogenous variables into account. Qian and Xie (2024) propose a method that allows discrete (e.g., binary) and continuous regressors (endogenous and exogenous), transformations of them, and interactions between them. However, unlike our method, these regressors cannot handle normally distributed endogenous regressors, and their estimator is computationally expensive and complex to implement.

Our research extends these findings by establishing a general and flexible framework permitting several correlated endogenous and exogenous variables – with the former being continuous and the latter being discrete or continuous – and allowing non-linearities, such as interactions between any variables. The proposed framework and methodological advances focus on the control function technique, which only requires adding additional copula terms to the model of interest. It is therefore easier for researchers to adapt the control function technique to a large variety of models, because this does not require determining and implementing the model's specific likelihood function. Consequently, in practice, most researchers use the control function technique when applying the Gaussian copula approach in empirical studies (Becker et al., 2022). Overall, our research includes the following main contributions:

First, we provide analytical derivations that show the finite sample bias's sources. This bias comprises two main components. One component is the endogenous regressor's estimated cumulative distribution function's (CDF's) quality. In finite samples, the CDF's estimation leads to errors in the copula's estimation, which we show is proportionally related to bias in the parameters of interest. Previous research on the method (Becker et al., 2022; Eckert & Hohberger, 2023) and its application in empirical (marketing) studies neglects the CDF estimators' crucial role. In their original article, Park and Gupta (2012) propose a non-parametric, kernel-based density estimation with numerical integration, while most research refers to using an ECDF to estimate the endogenous regressor's distribution, but does not explain how to do so exactly. We, on the other hand, show that using different methods to estimate the CDF could, in finite samples, produce very different results. The bias's second component arises from collinearity in the predictor matrix, which inflates the bias from the error in the CDF estimation. Our analytical derivations show that the bias is inversely proportional to a determinant based on the predictor matrix. In the simple case of only one endogenous variable, this determinant reduces to one minus the squared empirical correlation between the endogenous regressor and the copula term. Since the determinant (or the squared correlation) can be calculated from the observed data, it could provide researchers with a general measure for assessing the potential bias in the Gaussian copula approach.[2]

Second, we build on the analytical findings regarding the role of the CDF estimator in the Gaussian copula approach's performance (bias) to develop an adjusted ECDF estimator. This estimator substantially reduces the finite sample bias that can arise when employing the Gaussian copula approach in regression models with an intercept.[3]

Third, while Haschka's (2022) and Yang et al.'s (2022) Gaussian copula approaches can handle correlations between continous endogenous regressors and exogenous covariates, we extend these methods to a more flexible framework. It accommodates multiple endogenous regressors on continuous scales and exogenous regressors on both discrete (e.g., binary) and continuous scales. Our framework also accounts for non-linearities (e.g., interaction effects) in the model and allows the copula structure to vary across categories of discrete exogenous covariates.

Fourth, based on our methodological contributions, extensive simulations that mimic realistic application situations in marketing studies show (i) that the adjusted ECDF estimator that we propose outperforms related techniques previously used in research (i.e., Becker et al., 2022; Eckert & Hohberger, 2023; Park & Gupta, 2012) in order to mitigate the finite sample bias. Additional simulation results substantiate the assumption that (ii) our more general framework enables researchers to validly deal with multiple endogenous and exogenous regressors, along with the interactions between them.

Fifth, we provide guidelines for the application of the Gaussian copula approach in the proposed framework using the adapted ECDF estimator and apply them to an empirical example with real data that Park and Gupta (2012) used in their original publication. Furthermore, we extend Park and Gupta's (2012) empirical example by considering multiple endogenous and exogenous regressors as well as a varying copula structure across time/quarters.

This research's results support the Gaussian copula approach's application to endogeneity in regression models under a more relaxed non-normality requirement than those that Becker et al. (2022) and Eckert and Hohberger (2023) present. Our proposed CDF estimator has lower finite sample bias even for mildly non-normal distributions. In addition, by building on Yang et al. (2022), our more general Gaussian copula model also remains valid with normally distributed endogenous regressors if one or more exogenous regressors are sufficiently non-normal. Our proposed framework and estimator therefore allow researchers to use the Gaussian copula approach for regression models with and without an intercept in a much wider range of data constellations. We show that the use of our adjusted estimator could lead to unbiased results even when only relatively small sample sizes are available. Consequently, this research contributs to a valid Gaussian copula approach's application to correct endogeneity bias in regression models.

The remainder of this article is structured as follows: To begin with, we use a simple regression model with one endogenous variable to demonstrate the importance of the CDF estimator for the Gaussian copula approach's performance. Building on these foundations, we extend the Gaussian copula approach to accommodate scenarios with multiple endogenous variables, discrete, and continuous exogenous variables, as well as non-linearities (e.g., interaction terms) and varying copula structures. Next, we validate this more general framework by means of an extensive simulation study. To highlight the applicability of our research, we present a flowchart outlining the application of the Gaussian copula approach with our proposed enhancements. We conclude this article with a discussion of the results and suggestions for future research.

[2] While this is an interesting and important contribution of our research, the comprehensive development of generally applicable guidelines is beyond the scope of this article. Future research can therefore utilize these findings to develop guidelines based on the predictor matrix's determinant in order to help researchers assess a system of variables' required non-normality when applying Park and Gupta's (2012) Gaussian copula approach.

[3] In independent work, a recent working paper by Qian et al. (2024) suggest a rank-based modification of the copula to reduce the bias problem in finite samples. However, they only provide finite sample results for a single restrictive DGP and no analytic results or discussion regarding the various components of bias inflation.

CDF estimators' relevance for the Gaussian copula approach's performance and an adjusted ECDF estimator

In the Gaussian copula approach, an important step is estimating the endogenous variable's CDF. This is required regardless of whether the control function technique or the maximum likelihood-based copula approach are used. There are different ways of estimating a random variable's CDF (e.g., using a kernel approach or an ECDF). To date, researchers often neglect the role of the CDF estimator in the Gaussian copula approach's performance. Consequently, to begin with, we will highlight the importance of selecting an appropriate CDF estimator. We first analyze a simple case with only a single endogenous regressor as in Park and Gupta (2012) and Becker et al. (2022). In Online Appendix OA.7, we show that these findings are also transferable to the more general, multivariate framework that we develop later in this article.

Characterizing the bias in the simple bivariate model

Assume the data generation of an observation i in line with the following model:

$$y_i = \alpha + \beta P_i + \epsilon_i, \tag{1}$$

where y_i is the dependent variable, $\alpha \in \mathbb{R}$ is the intercept, $\beta \in \mathbb{R}$ is the slope coefficient of interest, and P_i is a one-dimensional regressor with a strictly monotonic CDF and is potentially correlated with the error term ϵ_i. Owing to the endogeneity of P_i, ordinary least squares (OLS) estimation produces a biased estimate of β. The Gaussian copula approach to address endogeneity uses the following three assumptions: (1) non-normal endogenous regressor P_i, and (2) the Gaussian copula structure

$$\begin{pmatrix} \epsilon_i \\ P_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{bmatrix} \right), \tag{2}$$

where $P_i^* = \Phi^{-1}\left(F_P\left(P_i\right)\right)$, $F_P(\cdot)$ is the CDF of P_i, $\Phi^{-1}(\cdot)$ is the inverse of the standard normal CDF, and ρ is the correlation between ϵ_i and P_i^*. This implies (3) that $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma^2 > 0$ (i.e., a normal error distribution). Under these assumptions, Park and Gupta (2012) suggest that the Gaussian copula approach provides an unbiased, IV-free estimator of β, using either a maximum likelihood or a control function technique. While our research primarily centers on the control function technique due to its prominence in empirical studies (Becker et al., 2022), our contributions and proposed adjustments are also applicable to maximum likelihood estimation.

The Gaussian copula's control function technique adds a copula term to Eq. 1. In this single endogenous regressor case, the copula term is P_i^*, which we symbolize as $C\left(P_i\right)$ to denote its dependence on P_i. Online Appendix OA.1 describes how including $C\left(P_i\right)$ acts as a control function that removes the endogeneity problem. In practice, $C\left(P_i\right)$ is unknown, so it is estimated by $\widehat{C}\left(P_i\right) = \Phi^{-1}\left(\widehat{F}_P\left(P_i\right)\right)$, which requires estimating the CDF of P_i, $\widehat{F}_P(\cdot)$. The Gaussian copula approach endeavors to control the endogeneity in P_i by adding $\widehat{C}\left(P_i\right)$ as a control function to the model in Eq. 1. Consequently, we consider the OLS estimation of the following model:

$$y_i = \alpha + \beta P_i + \gamma \widehat{C}\left(P_i\right) + u_i,\ i = 1, 2, \ldots, n, \tag{3}$$

where, according to Eq. 2, $\gamma = \sigma\rho$. In finite samples, $\widehat{C}\left(P_i\right) \neq C\left(P_i\right)$, and the estimation error in $\widehat{C}\left(P_i\right)$ results in a generated regressor bias in $\widehat{\beta}$ (Pagan, 1984). As Becker et al. (2022) show, the bias in $\widehat{\beta}$ can be large in regression models with an intercept. Equation 4 depicts the analytical bias (see Online Appendix OA.2 for details) in regression models with an intercept

$$bias\left(\widehat{\beta}|P\right) \approx \frac{\gamma}{\left(1 - \widehat{Corr}\left(P_i, \widehat{C}\left(P_i\right)\right)^2\right)} \cdot \left[\text{estimation error of } \widehat{C}(P_i)\right], \tag{4}$$

where $P = \left\{\left(P_i\right)\right\}_{i=1}^n$ represents all of the endogenous regressor's sample values. In general, the bias increases as γ increases, and as the estimation error of $\widehat{C}\left(P_i\right)$ increases (e.g., when we use a less accurate estimator of $\widehat{C}\left(P_i\right)$). If the collinearity between P_i and $\widehat{C}\left(P_i\right)$ increases (i.e., an increase in $\widehat{Corr}(P_i, \widehat{C}(P_i))$), any bias stemming from γ, as well as from the estimation error from $\widehat{C}\left(P_i\right)$, is amplified. Online Appendix OA.2 shows that the construction of this amplification effect is much less pronounced when there is no intercept and P_i has a non-zero mean. This situation corresponds to the simulation results in Park and Gupta (2012). Equation 4 also reveals that the non-normality of the endogenous regressor P_i only affects bias through its correlation with $\widehat{C}\left(P_i\right)$. Reducing the bias of $\widehat{\beta}$ also requires accurate estimation of $C\left(P_i\right)$. In order to do so, we outline an adjusted ECDF estimator, which demonstrates considerably less bias in finite samples than the CDF estimator that Park and Gupta (2012) use in their original study (Table 1), as well as the ECDF estimators that Becker et al. (2022) and Eckert and Hohberger (2023) employ.

Different CDF estimators for the copula model

While the non-normality in P_i determines the correlation between $C\left(P_i\right)$ and P_i, Eq. 4 shows that researchers could reduce the bias in $\widehat{\beta}$ by estimating $\widehat{C}\left(P_i\right)$ accurately. Given that $\widehat{C}\left(P_i\right) = \Phi^{-1}\left(\widehat{F}_P\left(P_i\right)\right)$, we should focus on the estimator $\widehat{F}_P(\cdot)$

to improve $\widehat{C}(P_i)$. We therefore outline common ways of estimating $F_P(\cdot)$ and propose an adjusted ECDF estimator (Table 1).

Park and Gupta (2012) *first* use an estimator of $F_P(\cdot)$, namely F_1, that utilizes a kernel-based estimator of the probability density function h, which was subsequently used to estimate the CDF via

$$\widehat{F}_1(x) = \int_{-\infty}^{x} \widehat{h}(y)dy, \tag{5}$$

where $\widehat{h}$ is a kernel-based estimator of h. That is:

$$\widehat{h}(x) = \frac{1}{n \cdot b} \sum_{i=1}^{n} K\left(\frac{x - P_i}{b}\right), \tag{6}$$

where $b > 0$ is a bandwidth parameter, and K represents a kernel function. Park and Gupta (2012) use an Epanechnikov kernel and chose the bandwidth that Silverman (1998) proposes. However, this bandwidth choice was not the most sensible, since it is only mean-squared error (MSE) optimal if the underlying variable P_i has a Gaussian distribution, while, on the other hand, the Gaussian copula approach requires P_i to be non-Gaussian. Consequently, and for comparison purposes, we use an Epanechnikov kernel with a cross-validated bandwidth selection to estimate the CDF (Li et al., 2017). In contrast to F_1, this *second* F_2 estimator (Table 1), which uses a cross-validated bandwidth choice, does not require a Gaussian distribution for optimality.

The ECDF is another candidate for estimating $F_P(\cdot)$:

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(P_i \le x). \tag{7}$$

In the context of the Gaussian copula approach, a key concern about ECDF estimators is that they require the inverse standard normal function $\Phi^{-1}()$ of the estimated ECDF, and that $\Phi^{-1}(1) = \infty$, because the ECDF will be one for the largest value in P_i (i.e., $\widehat{F}\left(\max_i [P_i]\right) = 1$), giving us $\max_i \left[\widehat{C}(P_i)\right] = \infty$. To alleviate this problem, several researchers (e.g., Becker et al., 2022; Eckert & Hohberger, 2023) and software implementations (e.g., the R package REndo, Gui et al., 2022) use a *third* estimator, F_3, which adjusts the ECDF as

$$\widehat{F}_3(x) = \begin{cases} 10^{-7} & \text{if } \widehat{F}(x) = 0 \\ 1 - 10^{-7} & \text{if } \widehat{F}(x) = 1 \\ \widehat{F}(x) & \text{else.} \end{cases} \tag{8}$$

However, $\Phi^{-1}(10^{-7}) = 5.2$ and $\Phi^{-1}(1 - 10^{-7}) = -5.2$ are extreme values to draw from a standard Gaussian distribution. In fact, each of these draws occurs at a rate of 1 in 10 million. Since $\widehat{F}\left(\max_i [P_i]\right) = 1$, there is always one value of $\widehat{C}(P_i)$ that deviates considerably from a Gaussian distribution. In small samples, this characteristic can generate substantial bias. Although there are alternative estimators based on the ECDF (e.g., Hill & Mann, 2000), they also suffer from $\widehat{F}\left(\max_i [P_i]\right) = 1$, therefore rendering them unsuitable for the Gaussian copula approach.

Since the estimators F_1, F_2, and F_3 often introduce substantial bias to regression models with an intercept (see the results of simulation Study 1), we propose a *fourth* adjusted ECDF estimator (F_4). To facilitate its use in applied settings, the proposed estimator is simple to implement, while the desirable asymptotic properties of the ECDF are maintained. The F_4 estimator is given as

$$\widehat{F}_4(x) = \frac{1}{2n} + \frac{(n-1)}{n^2} \sum_{i=1}^{n} I(P_i \le x). \tag{9}$$

This adjusted ECDF estimator is aimed at better approximating the true CDF in small samples in which the Gaussian copula approach has the largest bias. The new correction term is obtained by minimizing the true CDF's MSE, using the standard ECDF as a predictor in a linear regression. The resulting coefficients are subsequently approximated in terms of the sample size (see Online Appendix OA.3 for details). This CDF estimator resolves the previously noted problem of $\widehat{F}\left(\max_i [P_i]\right) = 1$, because $\widehat{F}_4\left(\max_i [P_i]\right) = \frac{1}{2n} + \frac{(n-1)}{n}\widehat{F}\left(\max_i [P_i]\right) = \frac{n-0.5}{n}$. Moreover, $\widehat{F}_4(x) = \frac{1}{2n} + \frac{n-1}{n}\widehat{F}(x)$, so $\widehat{F}_4$ inherits the ECDF's many desirable large sample properties, such as its (uniform) convergence to F.

Table OA.3-1 in the Online Appendix graphically compares the estimated values of the CDF and the copula obtained by applying the four different CDF estimators to a sample of 10 observations from a uniform and Chi-squared distribution.

Simulation Study 1: Showcasing the CDF estimator's importance

Based on the analytical derivations in the previous section, the simulation results should substantiate and extend our findings and knowledge regarding (1) the CDF estimators' relevance for the Gaussian copula approach's performance (bias), and (2) the adjusted ECDF estimator's performance and usefulness compared to the alternatives outlined in Table 1.

We illustrate our analytical findings' consequences by showing that the choice of the CDF estimator leads to performance (bias) differences when the Gaussian copula approach is applied to regression models with an intercept. Specifically, we show that the proposed CDF estimator significantly reduces the approach's bias in various data constellations, regardless of whether the regression model has an intercept or not. Moreover, we show that as, for instance,

Table 1 Different estimators for the Gaussian copula approach

Estimator		Specification	Reason for including and software implementation
Based on an Epanechnikov kernel	F_1: The estimation of CDF using an Epanechnikov kernel with a Silverman bandwidth selection. *Reference*: Silverman (1998).	$\widehat{F}_1(x) = \int_{-\infty}^{x} \widehat{h}(y)dy$ with $\widehat{h}(p) = \frac{1}{n \cdot b} \sum_{i=1}^{n} K\left(\frac{p-P_i}{b}\right)$, $K(x) = 0.75 \cdot \left(1 - x^2\right) \cdot I(\lvert x \rvert \leq 1)$, and the bandwidth as $b = 0.9 \cdot n^{-\frac{1}{5}} \cdot \min(s, IQR/1.34)$, with s the sample standard deviation of $P_{1:n}$, and IQR the interquartile range of $P_{1:n}$	Park and Gupta (2012) use this CDF estimator in the original study to show the method's performance for regression models without an intercept. *Implementation*: R package np (Hayfield & Racine, 2008).
	F_2: The estimation of CDF using an Epanechnikov kernel with a cross-validated bandwidth selection. *Reference*: Li et al. (2017).	$\widehat{F}_2(x) = \int_{-\infty}^{x} \widehat{h}(y)dy$ with $\widehat{h}(p) = \frac{1}{n \cdot b} \sum_{i=1}^{n} K\left(\frac{p-P_i}{b}\right)$, $K(x) = 0.75 \cdot \left(1 - x^2\right) \cdot I(\lvert x \rvert \leq 1)$, and determining the bandwidth by minimizing the cross-validation function of Li et al. (2017)	Contrary to the bandwidth used in $\widehat{F}_1(x)$, a cross-validated bandwidth does not assume the endogenous regressor P's normality. *Implementation*: R package np (Hayfield & Racine, 2008).
Based on an empirical cumulative distribution function (ECDF)	F_3: The empirical cumulative distribution function (ECDF) with an endpoint adjustment. *Reference*: Becker et al. (2022).	$\widehat{F}(x) = \frac{1}{n} \sum_{i=}^{n} I(P_i \leq x)$ with $\widehat{F}_3(x) = \begin{cases} 10^{-7} & \text{if } \widehat{F}(x) = 0 \\ 1 - 10^{-7} & \text{if } \widehat{F}(x) = 1 \\ \widehat{F}(x) & \text{else} \end{cases}$	Becker et al. (2022) use this adjustment to show that the Gaussian copula's performance deteriorates if an intercept is included. Eckert and Hohberger (2023) also use this estimator in their simulation study on the Gaussian copula approach's performance. *Implementation*: R package REndo (Gui et al., 2022).
	F_4: Adjusted ECDF. *Reference*: This article.	$\widehat{F}_4(x) = \frac{1}{2n} + \frac{(n-1)}{n^2} \sum_{i=1}^{n} I(P_i \leq x)$	A sample size-dependent adjustment of the ECDF, which prevents $\widehat{F}(x)$ from being 0 or 1, but still retains the ECDF's desirable large sample properties. *Implementation*: Our own R code implementation.

$I(\cdot)$ is the indicator function, n the sample size, P_i the endogenous regressor for observation $i = 1, \ldots, n$, and $P_{1:n}$ is the vector of all of the endogenous regressor's values

Becker et al. (2022) and Eckert and Hohberger (2023) disclose, other established estimators exhibit the Gaussian copula approach's known problems.

Goal and design This simulation study's goal is to assess the importance of the CDF estimator choice for the Gaussian copula approach's performance. We therefore apply the four different estimators to determine the performance differences based on bias. The aim is to achieve an unbiased result. We use the basic design and model of Park and Gupta's (2012) Study 1, but include an intercept in both the DGP and in the regression specification. Specifically, we consider the DGP

$$y_i = 3 - 1P_i + \epsilon_i, \tag{10}$$

where P_i is endogenous.[4] The model is estimated with the corresponding regression specification

$$y_i = \alpha + \beta P_i + \gamma \widehat{C}\left(P_i\right) + u_i, \tag{11}$$

where we will investigate how the $\widehat{\beta}$ coefficient's bias changes when estimating the copula term $\widehat{C}\left(P_i\right)$ with each of the CDF's four estimators outlined in Table 1. We consider nine different sample sizes (i.e., 100, 200, 400, 800, 1,600, 3,200, 6,400, 12,800, and 25,600), and use 3,000 replications of the 36 different specifications (i.e., nine different sample sizes and four alternative CDF estimators). In Online Appendix OA.4, we extend these simulations by considering, for example, models and regression specifications without intercepts, and an endogenous regressor with a zero and non-zero mean.

Since different P_i distributions could change the endogeneity bias's level in a regression model without a copula term, we compare the bias in a regression model *with* a copula term to the bias in a regression *without* a copula term. The relative bias is given as

$$relative\ bias = \frac{E\left[\widehat{\beta}\right] - \beta}{E\left[\widehat{\beta}_{naive}\right] - \beta}, \tag{12}$$

where $\widehat{\beta}_{naive}$ is the estimate of the endogenous regressor's coefficient when a copula term is excluded, $\widehat{\beta}$ is the estimate of the endogenous regressor's coefficient when a copula term is included, and β is the true coefficient. Consequently, the relative bias accounts for the different levels of endogeneity under different distributions and scales of P_i, as reflected in $E\left[\widehat{\beta}_{naive}\right] - \beta$. As such, the relative bias is a measure of the bias reduction when the copula term is included in the regression model. More specifically, if $|relative\ bias| < 1$, the copula model, compared to a regression model with an untreated endogeneity problem, reduces the bias.

Results Fig. 1 shows the relative bias results of the four CDF estimators in Table 1, when considering the DGP and the estimation specification in Eqs. 10 and 11 (for the corresponding raw bias results, see Table OA.4-6 in the Online Appendix). In line with Becker et al. (2022), we find that existing methods for estimating the CDF of P_i yield a substantial bias in a regression model with an intercept comprising small to medium sample sizes. However, we find that the adjusted ECDF estimator (i.e., $\widehat{F}_4$) only has a marginal relative bias compared to other estimators with small sample sizes. Further, the newly proposed estimator converges rapidly toward unbiasedness when the sample size increases. More specifically, when using the adjusted ECDF estimator, we obtain approximately unbiased estimates (i.e., less than 5% relative bias) around a sample size of 200 observations. In contrast, the second-best estimator (i.e., $\widehat{F}_3$) needs a sample size of around 3,200 observations to be approximately unbiased.

We also compare the variance of $\widehat{\beta}$ when using the different estimators (see Table OA.4-12 in the Online Appendix), and find a general pattern that allows us to rank the CDF estimators from the lowest to the highest variance of $\widehat{\beta}$. More precisely, $\widehat{F}_3$ and $\widehat{F}_4$ produce a low $\widehat{\beta}$ variance, while $\widehat{F}_1$ and particularly $\widehat{F}_2$ have a comparatively high variance when the sample sizes are small.

When using the Gaussian copula approach, we find that two general properties affect $\widehat{\beta}$´s bias (see also our characterization of the bias in the simple bivariate model): (1) how well we estimate the copula term, $\widehat{C}\left(P_i\right)$, and (2) the correlation between the endogenous regressor and the copula term, $Corr(P, \widehat{C}(P))$. To understand how the different CDF estimators (Table 1) perform with respect to these two properties, we calculated the correlation $Corr(P, \widehat{C}(P))$ and the root mean square error of $\widehat{C}(P)$, $RMSE\left(\widehat{C}(P)\right)$, which measures the estimation error in $\widehat{C}(P)$, with larger values indicating more error (see Table OA.4-12 in the Online Appendix).

While $Corr(P, \widehat{C}(P))$ is fairly stable across the sample sizes for $\widehat{F}_1$, $\widehat{F}_2$, and $\widehat{F}_4$, the estimator $\widehat{F}_3$ produces a comparatively low $Corr(P, \widehat{C}(P))$ for small sample sizes. The latter is due to $\widehat{F}_3$ using $\widehat{F}_3\left(\max_i\left[\mathrm{P_i}\right]\right) = 1 - 10^{-7}$ and $\widehat{F}_3\left(\min_i\left[\mathrm{P_i}\right]\right) = 10^{-7}$, both of which are unlikely values to obtain when drawing a small sample from a standard normal distribution compared to a larger sample. Consequently, these extreme endpoint values reduce the correlation between the endogenous regressor and the estimated copula term. Although $\widehat{F}_3$'s low $Corr(P, \widehat{C}(P))$ values are beneficial in terms of bias, it is burdened by the relatively high $RMSE\left(\widehat{C}(P)\right)$ values for smaller samples. That is, $\widehat{F}_3$ does not estimate $C(P)$ well, which

[4] We generate $P = \Phi(\mathrm{P}^*)$, with $\begin{pmatrix} \epsilon \\ P^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$, resulting in $P \sim U(0, 1)$.

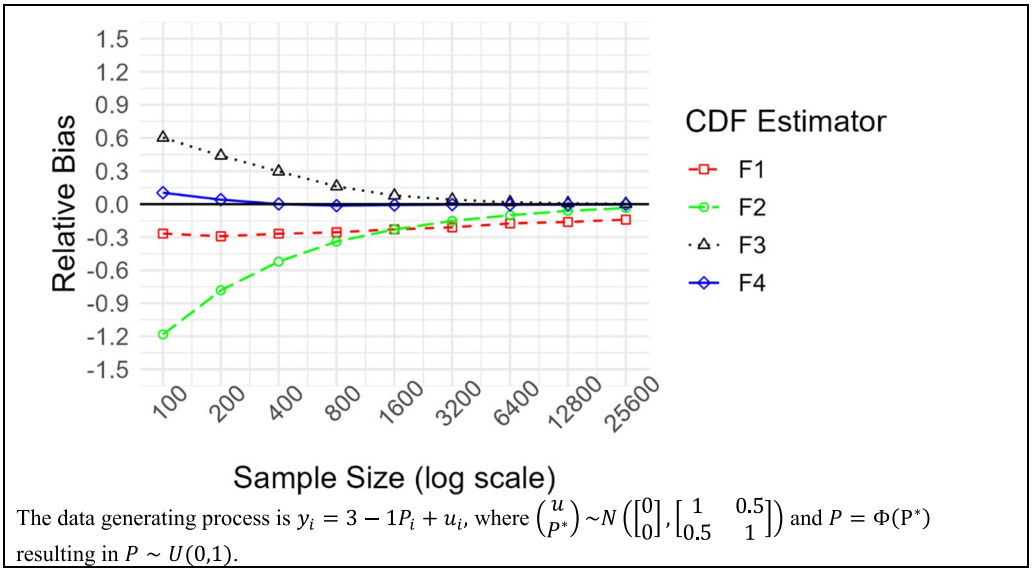

Fig. 1 Different CDF estimators' relative bias

worsens the bias of $\widehat{\beta}$. Comparing the CDF estimators, we find that the $RMSE\left(\widehat{C}(P)\right)$ is lowest (best) for $\widehat{F}_4$, but is closely followed by $\widehat{F}_1$, which has the second lowest RMSE across the simulation designs. The reason for $\widehat{F}_4$ outperforming $\widehat{F}_1$ in terms of relative bias, is due to $\widehat{F}_4$ having a lower $Corr(P, \widehat{C}(P))$, and a better $RMSE\left(\widehat{C}(P)\right)$.

Summary and discussion Study 1 demonstrates that in regression models *with* an intercept, the CDF estimators perform differently. These results support our claim that the choice of a CDF estimator is important for the Gaussian copula approach's performance. Surprisingly, the Gaussian copula literature has largely overlooked the relevance of the CDF estimator.

Study 1 also reveals that the adjusted ECDF estimator consistently outperforms other estimators. Online Appendix OA.4 extends the results in this section by considering situations where (1) data is generated either *with* or *without* an intercept, (2) the data has a mean that either *differs* from zero or is *equal* to zero (such as when using centered data), and (3) the model is estimated either *with* or *without* an intercept. The proposed CDF estimator performs better than the CDF estimators in all cases considered in Online Appendix OA.4, except in the simulation setup that corresponds to Park and Gupta's (2012) Study 1, in which neither the DGP nor the regression specification includes an intercept. In this case, all the estimators perform equally well. Furthermore, the extended simulation results in Online Appendix OA.4 indicate that researchers can include an intercept in their estimation equation if they use the adjusted ECDF – even if the data-generating process does not include an intercept.

By only covering a uniformly distributed endogenous regressor, this simulation study resembles Park and Gupta's (2012) original simulation setup. Online Appendix OA.5 extends the simulations by replicating Becker et al.'s (2022) extensive simulations in Study 4. This replication considers a range of endogenous regressor distributions, but also varies the endogeneity level and the dependent variable's explained variance. In line with the theoretical derivations (see Eq. 4 and Online Appendix OA.2), we find that neither the level of endogeneity nor the explained variance affects the Gaussian copula approach's relative bias. However, we do find that the degree of non-normality affects the bias. As argued when characterizing the bias in the simple bivariate model, the degree of non-normality is relevant for the Gaussian copula and is determined by $\widehat{Corr}(P, \widehat{C}(P))$. Online Appendix OA.6 demonstrates, by means of simulations, that in cases involving a single endogenous regressor, a researcher could gauge the relative magnitude of the bias by only considering the sample size and $\widehat{Corr}(P, \widehat{C}(P))$.

A general framework for Gaussian copula-based model estimation

Methodological extension

The previous section considered the simple case of explaining outcome y_i with a single endogenous regressor P_i (see Eq. 1). Now, we show how to extend the copula method to handle multiple continuous endogenous regressors $P_i \in \mathbb{R}^{d_P}$, additional continuous exogenous regressors $W_i \in \mathbb{R}^{d_W}$, and discrete exogenous regressors $Z_i \in \mathbb{R}^{d_Z}$, such as binary variables. We allow for multiple regressors of each type, various

transformations, and various interactions. In addition, we also allow P, W, and Z to be correlated and the copula structure to vary across different categories of the discrete exogeneous variable Z.

We leave the regression equation for researchers to specify; therefore, keeping it in its most general form

$$y_i = g\left(P_i, W_i, Z_i\right) + \epsilon_i, \tag{13}$$

where $g(P_i, W_i, Z_i)$ is the structural function of interest to the researcher. This could be a conventional linear-in-parameters function with variables (P_i, W_i, Z_i), which could include interactions between the variables and any non-linear transformations, such as quadratics or logarithms. This setting is very general and encompasses many alternative models. For example,

1. A single endogenous regressor:
$$g\left(P_i\right) = \alpha + \beta P_i, \tag{14}$$
which we introduced in the previous section.

2. The specification of the empirical example in Park and Gupta (2012):
$$\begin{aligned} g\left(P_i, W_i, Z_i\right) = \alpha &+ \beta_1 P_{1,i} + \beta_2 P_{2,i} + \beta_3 P_{3,i} \\ &+ \delta_1 Z_{1,i} + \delta_2 Z_{2,i} + \delta_3 Z_{3,i}, \end{aligned} \tag{15}$$
which we will study in-depth in the later empirical example.

3. Multiple endogenous, exogenous, and interactions:
$$\begin{aligned} g\left(P_i, W_i, Z_i\right) = \alpha &+ \beta_1 P_{1i} + \beta_2 P_{2i} + \delta_W W_i + \delta_Z Z_i \\ &+ \delta_{12} P_{1i} P_{2i} + \delta_{1W} P_{1i} W_i + \delta_{2Z} P_{2i} Z_i, \end{aligned} \tag{16}$$
which we will study in-depth with the means of simulations in the subsequent section.

One goal of this research is to develop a more general framework for the Gaussian copula approach. We therefore outline the copula-based identification assumption providing the basis of how we derive the copula terms, and how we add them to the original regression specification (see Online Appendix OA.7 for details). Like the standard Gaussian copula structure given in Eq. 2, the general copula structure also assumes a normal distribution between the transformations of observed continuous variables and the regression error term. However, it incorporates three extensions: First, it allows multiple endogenous variables to be correlated with the error term. Second, it allows multiple exogenous variables to be correlated with multiple endogenous variables. In Online Appendix OA.7, we show that this extension requires the copula terms to incorporate information from both the endogenous and exogenous continuous variables. Third, the copula structure is allowed to differ between the discrete Z variable's categories.

Since P_i is endogenous, the OLS estimator applied to Eq. 13 is biased. We next describe how to augment this model (Eq. 13) via control functions, and to mitigate bias in the parameters of $g(P_i, W_i, Z_i)'s$ estimation under a general copula-based identification assumption (see Online Appendix OA.7). To better illustrate the above, we first extend the copula-based identification assumption in Eq. 2 by allowing (potentially) multiple P and W regressors but assume that no Z regressors exist. Specifically, following our previously introduced notation, we create $P_i^* = C\left(P_i\right) = \Phi^{-1}\left(F_P\left(P_i\right)\right)$ for each endogenous regressor P; and for each exogenous regressor W, we create analogous $W^* = C\left(W_i\right) = \Phi^{-1}\left(F_W\left(W_i\right)\right)$ variables. We assume $(\epsilon_i, P_i^*, W_i^*)$ are jointly normal, P_i^* is correlated with ϵ_i and W_i^*, but, as in Yang et al. (2022), ϵ_i and W_i^* are uncorrelated.

As in the single endogenous regressor case without W, the method continues by adding a copula term to the original regression specification to mitigate the endogeneity. However, we now not only require a copula term for each endogenous regressor, but also need to account for the introduction of W in the copula terms. The copula terms take the following multivariate form:

$$C\left(P_i, W_i\right) = \left(C\left(P_i\right)' \; C\left(W_i\right)'\right) \Sigma^{-1}_{C(P_i),C(W_i)} \begin{pmatrix} I_{d_P} \\ 0_{d_W \times d_P} \end{pmatrix}, \tag{17}$$

where $\Sigma_{C(P_i),C(W_i)}$ is the variance-covariance matrix of $(C(P), C(W))$, $0_{d_W \times d_P}$ is a $\left(d_W \times d_P\right)$ matrix of 0's, with d_W and d_P being the number of W and of P regressors, respectively, and I_{d_P} is the identity matrix of dimension d_P. Note that this results in the same number of copula terms as endogenous regressors P, i.e. $C\left(P_i, W_i\right)$ is a $\left(1 \times d_P\right)$ vector of copula terms. For example, for a single endogenous P and a single exogenous W, this results in one copula term of the form $C\left(P_i, W_i\right) = \{C\left(P_i\right)\sigma^2_{C(W)} - C\left(W_i\right)\sigma_{C(P),C(W)}\}/\{\sigma^2_{C(P)}\sigma^2_{C(W)} - \sigma^2_{C(P),C(W)}\}$. From Eq. 17 we denote $C_k(P_i, W_i)$ as the k^{th} copula term (i.e., the k^{th} entry in the d_P-dimensional vector $C(P_i, W_i)$). For each of these copula terms, we either need the endogenous variable P or one of the exogenous variables W to be non-normal if, similar to Yang et al. (2022), $Corr(P^*, W^*) \neq 0$. This implies that no linear combination of P and W should be perfectly correlated with the copula term (i.e., we should have full rank of the predictor covariance matrix).

To remove the endogeneity, we must add these copula terms to the original regression model (Eq. 13), arriving at

$$y_i = g\left(P_i, W_i\right) + \sum_{k=1}^{d_P} \gamma_k C_k\left(P_i, W_i\right) + \epsilon_i. \tag{18}$$

However, as in the simple model, $C_k(P_i, W_i)$ are unknown, so they must be estimated (e.g., by using the adjusted ECDF estimator). It should be specifically noted that the copula terms depend on the covariance matrix of (P^*, W^*), which must also be estimated (e.g., by the empirical variance-covariance matrix). Using the estimated control functions $\widehat{C}_k(P_i, W_i)$, Eq. 13 then becomes:

$$y_i = g\left(P_i, W_i\right) + \sum_{k=1}^{d_P} \gamma_k \widehat{C}_k(P_i, W_i) + u_i. \tag{19}$$

Similarly to the case of a single endogenous regressor, OLS estimation of Eq. 19 will alleviate the endogeneity bias arising from the endogenous regressors P_i.

At this stage, we have not discussed how to include discrete exogenous regressors. Fortunately, this only requires a slight extension of the above methodology. Specifically, we need to allow each copula term to vary with each category/group that the discrete regressors can form. For example, let us assume we have two binary variables, $\left(V_1, V_2\right)$, indicating marital status and loyalty membership, respectively. We use Z as the categorical variable representing all possible combinations of the discrete variables.[5] In the previous example with two binary variables we could therefore have four unique combinations $\left(V_1, V_2\right) = Z \in \left\{z_1, \ldots, z_4\right\} = \{(0,0),(0,1),(1,0),(1,1)\}$ where z_1 is the label for observations that are not married and have no loyalty membership $(V_1 = 0, V_2 = 0)$, z_2 represents observations that are not married but have a loyalty membership $(V_1 = 0, V_2 = 1)$, and so on. Consequently, we must then allow the d_P-dimensional copula term to differ with respect to each of these four categories, resulting in a total of four different copula terms.

In order to allow the copula terms to differ with respect to each Z combination, we first allow P_i^* and W_i^* to depend on Z_i as follows: $P_{z,i}^* = C^z\left(P_i\right) = \Phi^{-1}\left(F_{P|Z}\left(P_i|Z_i = z\right)\right)$ and $W_{z,i}^* = C^z\left(W_i\right) = \Phi^{-1}\left(F_{W|Z}\left(W_i|Z_i = z\right)\right)$, where $F_{P|Z}(.|.)$ and $F_{W|Z}(.|.)$ are the conditional distributions of P and W given Z. Without further restrictions, this boils down to simply calculating P_i^* and W_i^* separately for each subset of data according to Z_i. For example, if there were two binary variables for Z, researchers would split their data into four sub-datasets based on the categories that Z can take. Thereafter, they separately estimate P_i^* and W_i^* in each sub-dataset. The covariance matrix of (P^*, W^*) also needs to be estimated conditional on Z. Again, we recommend that researchers simply calculate this object separately for observations in each category of Z.

Analogous to the model given by Eq. 19, the feasible regression model involves adding the estimated copula terms to the original regression of interest:

$$y_i = g\left(P_i, W_i, Z_i\right) + \sum_{z=z_1}^{z_J} \sum_{k=1}^{d_P} \gamma_k^z \widehat{C}_k^z(P_i, W_i) I\left(Z_i = z\right) + u_i, \tag{20}$$

where Z can take all possible joint values in $(z_1, \ldots, z_J)$, $C_k^z(P_i, W_i)$ is the k^{th} copula term for category $Z_i = z$, the regression parameters γ_k^z need to be estimated together with the parameters in $g(P_i, W_i, Z_i)$, while the indicator function $I(Z_i = z)$ ensures that y_i is only related to the copula term corresponding to Z_i.[6] Online Appendix OA.7 describes the general model and how $C_k^z(P_i, W_i)$ act as control functions that remove the endogeneity problem.

If researchers have sufficient theoretical support to impose the constraint that the copula structure is independent of the value of Z_i, the copula terms collapse back to those given in Eq. 18. It is possible to test whether the copula structure varies between Z's categories by testing whether the copula terms, related to each endogenous variable, differ across Z.[7] If there are no continuous exogenous variables W and no discrete variables Z, the copula terms are simply given by $C\left(P_i\right)$ (i.e., as in the simple bivariate model).

We briefly outline a three-step procedure, explaining how this could be implemented in practice for models with continuous and discrete regressors (see also the flowchart in Fig. 3 in the later empirical example). For each value $z = z_1, z_2, \ldots, z_J$ that Z can hold, take the following steps:

1. Estimate the CDFs of P_i and W_i, conditional on $Z_i = z$. This could be done using the adjusted ECDF, and only using the data $\left(P_i, W_i, Z_i\right)$ for which $Z_i = z$.
2. Use the estimated CDFs to create the copula control function via

$$\widehat{C}^z\left(P_i, W_i\right) = \left(\widehat{C}^z\left(P_i\right)' \;\; \widehat{C}^z\left(W_i\right)'\right) \widehat{\Sigma}^{-1}_{\widehat{C}^z(P), \widehat{C}^z(W)} \begin{pmatrix} I_{d_P} \\ 0_{d_W \times d_P} \end{pmatrix}, \tag{21}$$

where $\widehat{\Sigma}_{\widehat{C}^z(P), \widehat{C}^z(W)}$ is the estimated variance-covariance matrix of $\widehat{C}^z(P)$ and $\widehat{C}^z(W)$, calculated by only using the

[5] If the number of categories is large relative to the sample size, it may be necessary to merge multiple Z categories or introduce other functional form restrictions. A notable scenario involves panel data models e.g. of the form $Y_{it} = Z_i + \beta P_{it} + u_{it}$, where Z_i represents fixed effects. In such cases, within transformations or first-differences can be applied to eliminate Z_i entirely. The control function can then be obtained using either several copulas for the marginal errors over time or a copula for their transformations or differences as in Haschka (2022).

[6] Note that the superscript notation in Eq. 20, such as γ_k^z, is not an exponent but distinguishes that there is a different γ_k for every category of Z.

[7] Although these tests could inform subsequent research, it is advisable not to incorporate their results into modifications of the currently investigated model to avoid multiple testing issues.

data (P_i, W_i, Z_i) for which $Z_i = z$. If we do not have any discrete exogenous variables Z in the model, Eq. 21 is calculated as given in Eq. 17.

These two steps yield estimated control functions $\widehat{C}_k^z(P_i, W_i)$ for each value z of Z_i (or $\widehat{C}_k(P_i, W_i)$ if no Z_i is present in the model). They are used to augment the original regression (Eq. 13) in the third step:

3. Run the OLS regression in Eq. 20. If no Z_i is present, the OLS regression should instead be given as in Eq. 19.

The following simulation Study 2 demonstrates that augmenting the original regression equation in this way gives consistent estimates, and that the bias depends on the non-normality level. Next, we provide a theoretical expression explaining how the bias is related to the copula terms' estimation error, and to the multicollinearity level in the copula model.

Analysis of bias

When the model contains multiple endogenous and exogenous regressors, the bias sources are similar to those in the single endogenous case as shown in Eq. 4. In particular, let β denote the coefficient vector of the parameters of interest to the researcher, which can include both endogenous and exogenous variables. In Online Appendix OA.8, we show that

$$bias\left(\widehat{\beta}|P, W, Z\right) \approx constant \cdot \left[estimation\ error\ of\ \widehat{C}^Z\left(P_i, W_i\right)\right] \cdot \frac{\gamma^Z}{\omega}, \tag{22}$$

where $\gamma^Z = \left(\gamma_1^Z, \ldots, \gamma_{d_p}^Z\right)$ and $\widehat{C}^Z(P, W) = (\widehat{C}_1^Z(P, W), \ldots, \widehat{C}_{d_p}^Z(P, W))$. The bias inflation factor ω is a number derived from a matrix determinant, which can be calculated from the data and measures the degree of multicollinearity between the regressors in Eq. 20. Crucially, ω will decrease towards zero as the multicollinearity level increases in the model with $\omega = 0$ corresponding to perfect multicollinearity (full details are given in Online Appendix OA.8). Multicollinearity will typically be an issue in the Gaussian copula setup when P_i is very similar to $\widehat{C}\left(P_i\right)$, that is, when the distribution of P_i is close to normal.

In the case where $g\left(P_i, W_i, Z_i\right) = \alpha + \beta P_i$, ω reduces to $1 - \widehat{Corr}\left(P_i, \widehat{C}\left(P_i\right)\right)^2$, which is the bias inflation factor in the single endogenous case that Eq. 4 provides. Like the single endogenous variable case, the bias also depends on the copula terms' estimation error. However, we now have a copula term, $\widehat{C}_k^z\left(P_i, W_i\right)$, for each value z of Z_i, and for each endogenous variable $k = 1, \ldots, d_p$. The estimation error of each of the copula terms $\widehat{C}_k^z\left(P_i, W_i\right)$ is weighted by the vector γ_k^Z in Eq. 22, which in turn depends on the k^{th} endogenous regressor's degree of endogeneity. In sum, Eq. 22 shows that the bias is inflated for a certain level of the copula terms' estimation error, because the regressors of interest are more related to the other regressors and to the copula terms. Since the copula terms' estimation error could be decreased by increasing the sample size, and the multicollinearity level could be estimated consistently, future research can use extensive simulations to provide guidelines for the sample size and values of the bias inflation factor, ω, that offer low levels of bias.

Simulation Study 2: Investigating the generalized Gaussian copula approach

Goal and design In this simulation study, we assess our generalized framework's finite sample properties for the Gaussian copula estimation developed previously, and show that the bias converges asymptotically to zero in a variety of situations.

Consequently, this simulation study's focus is not the different CDF estimators, but whether we are able to retrieve unbiased coefficients for the general framework. Specifically, we will investigate the bias in the coefficients from a regression with the following variables: multiple endogenous, a continuous exogenous (correlated with the endogenous), a binary exogenous, an interaction between two endogenous, an interaction between an endogenous and the continuous exogenous, and an interaction between an endogenous and the binary exogenous. Specifically, we consider the following DGP

$$\begin{aligned} y_i = {} & 1 - 1P_{1,i} + 1P_{2,i} - 1W_i - 2Z_i \\ & - 1P_{1,i}P_{2,i} - 1P_{1,i}W_i - 2P_{2,i}Z_i + \epsilon_i, \end{aligned} \tag{23}$$

where $P_{1,i}$ and $P_{2,i}$ are continuous endogenous variables, W_i is a continuous exogenous variable, and Z_i is a binary exogenous variable. To illustrate that the proposed method can cope with the copula distribution changes between the Z categories, we specify the following two copula distributions:

$$\begin{aligned} \begin{pmatrix} \epsilon_i \\ P_{1,i}^* \\ P_{2,i}^* \\ W_i^* \end{pmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.4 & 0.4 & 0 \\ 0.4 & 1 & 0.4 & 0.4 \\ 0.4 & 0.4 & 1 & 0.4 \\ 0 & 0.4 & 0.4 & 1 \end{bmatrix}\right) \text{ if } Z = 0, \\ \begin{pmatrix} \epsilon_i \\ P_{1,i}^* \\ P_{2,i}^* \\ W_i^* \end{pmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.6 & 0.6 & 0 \\ 0.6 & 1 & 0.6 & 0.6 \\ 0.6 & 0.6 & 1 & 0.6 \\ 0 & 0.6 & 0.6 & 1 \end{bmatrix}\right) \text{ if } Z = 1. \end{aligned} \tag{24}$$

Using Eq. 20, the model is estimated with the following regression specification

$$y_i = \alpha - \beta_1 P_{1,i} + \beta_2 P_{2,i} - \delta_1 W_i - \delta_2 Z_i - \delta_3 P_{1,i} P_{2,i} - \delta_4 P_{1,i} W_i - \delta_5 P_{2,i} Z_i + \sum_{z=0}^{1} \sum_{k=1}^{2} \gamma_k^z \widehat{C}_k^z \left(P_{1,i} P_{2,i}, W_i\right) I\left(Z_i = z\right) + u_i, \tag{25}$$

where we will investigate the bias of $\widehat{\beta}$ and $\widehat{\delta}$ after adding the copula terms. Note that we estimate the copula terms conditional on Z. As described above, this is done by splitting the data into those observations $(P_{1,i}, P_{2,i}, W_i, Z_i)$ for which $Z_i = 0$ and those observations for which $Z_i = 1$ and estimating the copula terms separately within these two subsamples. In Online Appendix OA.10, we also show how ignoring the copula structure's conditional dependence affects the results negatively.

In this simulation study, we again consider nine different sample sizes (100, 200, 400, 800, 1,600, 3,200, 6,400, 12,800, and 25,600), but now also consider different distributions (Beta, Gamma, Chi-squared, and lognormal) with different distribution values, not only a uniform distribution. We systematically vary the endogenous variables' non-normality by varying the distribution parameters so that they range from very non-normal to very close to normal.[8] We set the distribution of the continuous exogenous variable W to a slightly non-normal distribution.

As our adjusted ECDF estimator $\widehat{F}_4$ had the lowest bias in the single endogenous variable case (see simulation Study 1 and Online Appendix OA.4), we focus on this CDF estimator in the remainder of the section. In Online Appendix OA.9, we also present the results of the other CDF estimators and a discussion of their relative performance in this more general framework.

Results Fig. 2 shows the bias of all seven parameters (i.e., $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\delta}_{1-5}$). The bias converges monotonically toward zero for all the model's parameters, reaching levels of unbiasedness between 1600 and 3200 observations. Note that the results shown in Fig. 2 average across a wide range of different distributions with different degrees of non-normality, including some distributions that are very close to normal. The bias converges to zero even faster for distributions with high non-normality.

A closer investigation of the distribution's effects (Online Appendix OA.9) shows that previous findings about the endogenous variable's non-normality (and now also that of the exogenous variables) are still applicable. The approach requires more observations to reach unbiasedness when the distributions are more normal, but the bias converges very quickly to zero when they are substantially non-normal. Additional simulations not reported in this article suggest that our approach shares the same capabilities as that of Yang et al. (2022) regarding leveraging the non-normality of the exogenous W, when an endogenous regressor correlated with W is not (very) non-normal. Consequently, the system's total non-normality is more important when there are additional correlated covariates and not merely the non-normality of each endogenous regressor. This is also supported by our derivations in the previous section, where we show that the model's determinant influences the bias. The latter is due to the determinant reflecting the level of multicollinearity in the copula augmented regression and not, as in the bivariate model, only the correlations between the endogenous variable and the copula term.

Overall, these simulations substantiate that our more general framework permits almost unbiased estimation across a wide range of distributions, provided that the assumptions underlying the copula approach are fulfilled. However, the same finite sample bias revealed in previous research on the Gaussian copula method also affects the extended framework. This finite sample bias is stronger in more normal distributions and depends on the CDF estimators' quality, which our derivations and simulation Study 1 have shown. Online Appendix OA.9 shows that our proposed CDF estimator is also the overall best choice in this more general framework.

Empirical example

To guide researchers in applying the proposed more general Gaussian copula framework and the adjusted ECDF estimator, we revisit Park and Gupta's (2012) empirical example, using the data they obtain from the IRI (Information Resources, Inc.) marketing dataset (Bronnenberg et al., 2008).[9] The dataset includes store-level sales of paper towels (category sales) for the two largest independent stores (store 1 and store 2) in Eau Claire, Wisconsin, from 2001 to 2005 (i.e., $N = 260$ weeks). More specifically, similar to Park and Gupta (2012), we also consider the following regression model:

$$\begin{aligned} log\left(Sales_t\right) = \alpha + \beta_1 \cdot log\left(Price_t\right) + \beta_2 \cdot Promotion_t + \beta_3 \cdot Display_t \\ + \beta_4 \cdot I(Z_t = q_2) + \beta_5 \cdot I(Z_t = q_3) + \beta_6 \cdot I(Z_t = q_4) + \epsilon_t, \end{aligned} \tag{26}$$

[8] We use the correlation between an endogenous regressor and its copula term in a bivariate model as a measure of non-normality. This measure is motivated by $Corr\left(C\left(P_i\right), P_i\right) = 1$ only when P_i has a normal distribution — as, then, $C\left(P_i\right) = \Phi^{-1}\left(\mathrm{F_P}\left(P_i\right)\right) = P_i$). Furthermore, this measure of non-normality aligns with the extensive simulations in Online Appendix OA.6, which shows how a bivariate model's bias depends on the degree of non-normality quantified by $Corr\left(C\left(P_i\right), P_i\right)$.

[9] We thank Sungho Park and Sachin Gupta for providing the dataset used in their article.

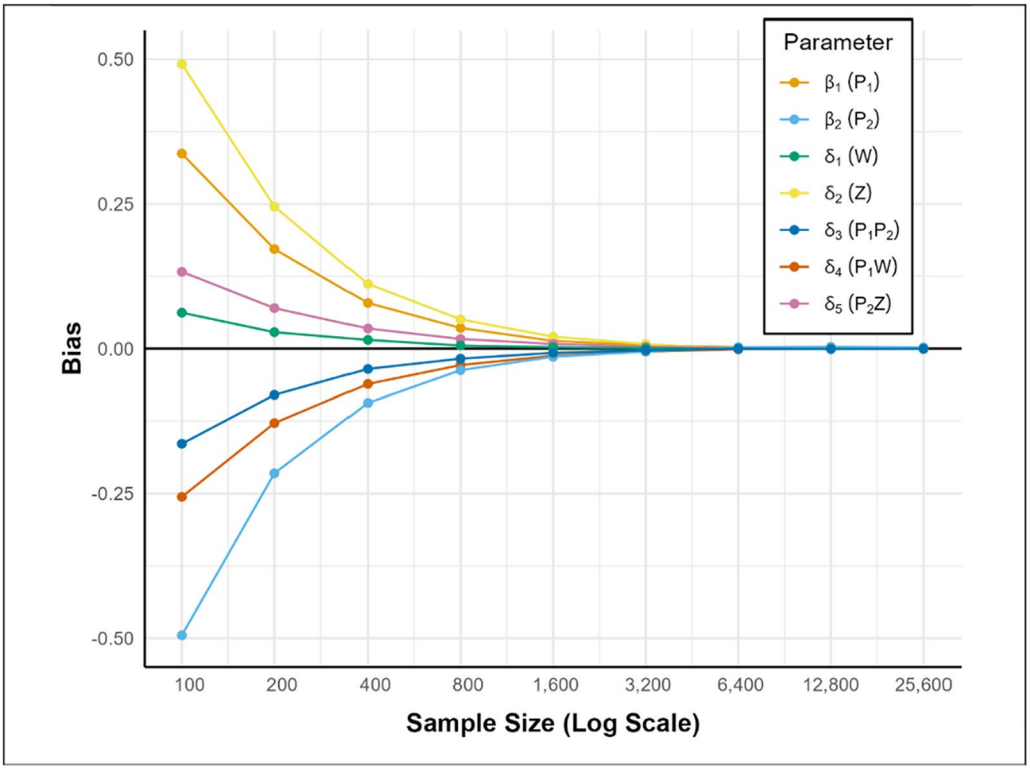

Fig. 2 Bias of F_4's different parameters. Note: The bias shown here averages across a wide range of different distributions with different degrees of non-normality. The pattern of the coefficients' positive and negative bias that we observe, corresponds to the uncorrected (endogenous) model's bias, suggesting that the bias in this copula model is a "remaining" bias that has not been corrected

where the logarithm of store sales across different points in time t is explained by the logarithm of retail price, promotion (e.g., feature advertising at the category level), display (e.g., shelf space allocation or special in-store display)[10] and a categorical variable Z_t, representing the quarter of the year ($I\left(Z_t = q_2\right) = 1$ if t is in the second quarter, $I(Z_t = q_3)$=1 if t is in the third quarter, and so on) with q_1 being the reference quarter,[11] and ϵ_t the error term. For the model estimation, we use our own code[12] using the statistical software R (R Core Team, 2022) and follow four analytical steps when applying the Gaussian copula approach with the adjusted ECDF:

Step 1: Theoretically justify the endogeneity problem.
Step 2: Check for non-normality in the regression system.
Step 3: Estimate a copula model:

- Estimate the CDFs of P_i and W_i, conditional on $Z_i = z$, or unconditionally if Z_i is not present.
- Use the estimated CDFs to create the copula control function via Eq. 21 if Z_i is present, or via Eq. 17 if Z_i is not present.
- Run the OLS regression in Eq. 20 if Z_i is present, or in Eq. 19 if Z_i is not present.

Step 4: Interpret the results and check for significance.

To begin with, Step 1 requires researchers to justify the presence of endogeneity theoretically. Consequently, they need to a priori carefully determine which of the variables are potentially endogenous, and which could be considered exogenous. This step usually also involves the careful consideration of alternative methods, such as the IV approach, if theoretically justifiable IVs are available in the dataset,

[10] More specifically, Park and Gupta (2012, p. 582) provide the following information about the variables Price, Promotion, and Display: "Retail price is defined on a per-roll basis. Retail price, in-store display, and feature advertising at the category level are computed as market share-weighted averages of UPC-level variables." While Price is a continuous variable, Promotion and Display are (at the store level) typically dummies. However, their computation as market share-weighted averages of UPC-level variables allows Park and Gupta (2012) to also consider Promotion and Display as continuous variables.

[11] In accordance with the operationalization of Z in the general framework, we use $\left(Q_{1,t}, Q_{2,t}, Q_{3,t}, Q_{4,t}\right) = Z_t \in \left\{q_1, ., q_4\right\}$ = {(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1)}, where Z can only hold four labels as the year quarters are mutually exclusive.

[12] The R code for this example is available at https://github.com/ECONshare/gaussian_copula.

or by providing reasons why such approaches would not be feasible.

These theoretical considerations protect researchers from an ill-considered application of the Gaussian copula approach to handle endogeneity problems. In this empirical example, we follow Park and Gupta's (2012) theoretical considerations in Step 1, and assume that unmeasured product characteristics or demand shocks influence both retailer pricing decisions and consumer decisions, resulting in price endogeneity. Likewise, we initially assume that Promotion and Display are exogenous variables. While Haschka (2022) shows that correlations between endogenous variables and exogenous variables could lead to bias when using the traditional Park and Gupta (2012) approach, the general framework we developed, allows correlations between all variables. It is also reasonable to allow the Gaussian copula's correlation structure to vary with each year quarter, as different unobserved variables might affect the demand and the retailers' pricing decisions differently during a year's quarters. This is likely to apply to many settings, and we suggest that researchers should assume that the copula structure varies between the categories of discrete exogenous variables Z, unless theoretical arguments clearly indicate otherwise.

In Step 2, we address non-normality, which traditionally focuses on the endogenous regressor (Park & Gupta, 2012). However, in our suggested framework, non-normality is only needed in the regression system as a whole (see the "Analysis of bias" section for our general framework).[13] If, for instance, the endogenous regressor log(Price) is normal, we need a non-normal exogenous regressor (Promotion or Display). Conversely, if the Promotion or Display variables are normal, we need a non-normal log(Price) variable. First, we visually check the continuous regressors' non-normality in the model (i.e., log(Price), Promotion, and Display). The estimated density plots of these three variables appear to be non-normal (Fig. OA.11-1 in the Online Appendix). This finding is supported by pronounced skewness of the log(Price), Promotion, and Display of store 1 (or store 2), which have values of -3.89 (-4.35), 1.89 (1.56), and 0.79 (1.00), respectively. To further substantiate our non-normality assessment, we apply the Anderson-Darling and Cramér-von Mises tests (as suggested by Becker et al., 2022 for the Gaussian copula approach), both of which reject the null hypothesis of normality with respect to each of the variables (p-values below 0.0001).

Next, in Step 3, we need to add the copula terms for the endogenous regressors to the original regression model. Assuming that only log(Price) is endogenous (as in Park & Gupta, 2012), we only need to add a single copula term to the model. However, in accordance with our general framework, the copula term needs to take the continuous exogenous variables (W) into account (i.e., Promotion and Display). Online Appendix OA.11.1 shows how to build the copula term using Step 3.1 and Step 3.2. Furthermore, by allowing the copula term to interact with the year quarters, the copula structure is allowed to vary across these quarters and provides the following regression model:

$$\begin{aligned}\log(Sales_t) = {} & \alpha + \beta_1 \cdot \log(Price_t) + \beta_2 \cdot Promotion_t + \beta_3 \cdot Display_t \\ & + \beta_4 \cdot I(Z_t = q_2) + \beta_5 \cdot I(Z_t = q_3) + \beta_6 \cdot I(Z_t = q_4) \\ & + \gamma_{Price}^{q_1} \widehat{C}_{Price,t}^{q_1} I(Z_t = q_1) + \gamma_{Price}^{q_2} \widehat{C}_{Price,t}^{q_2} I(Z_t = q_2) \\ & + \gamma_{Price}^{q_3} \widehat{C}_{Price,t}^{q_3} I(Z_t = q_3) + \gamma_{Price}^{q_4} \widehat{C}_{Price,t}^{q_4} I(Z_t = q_4) + u_t. \end{aligned} \tag{27}$$

The flowchart in Fig. 3 provides researchers with a tool for deciding which copula terms to include in their analyses. Following this flowchart, we start by categorizing the independent variables based on Step 1's theoretical considerations: P is the endogenous regressor log(Price); W is the non-normal exogenous regressors Promotion and Display, and Z is the discrete categorical variable denoting the year quarter.

We estimate the CDF of the log(Price), Promotion, and Display variables by using our adjusted ECDF (F_4) for each of the four year quarters. In accordance with Eq. 21, the copula term's computation of the endogenous log(Price) regressor then uses all three variables' estimated CDFs. We add the copula term to the original regression and estimate the new model via OLS. Table 2 shows the model estimation results for store 1 and store 2 (i.e., in the "Proposed: Price copula" columns 4 and 8). We also estimate a simple OLS regression that ignores endogeneity, and a model that replicates Park and Gupta's (2012) model directly.[14] In this model, the included copula

[13] In relation to our findings when analyzing the bias in the general framework, we could assess the magnitude of ω as an estimate of the degree of multicollinearity in the model, which tends to increase when the independent variables become more normally distributed. While we have analytically established this relationship, we refrain from reporting ω in our assessment as clear thresholds for problematic values have yet to be determined. We leave this for future research.

[14] We replicated the coefficient estimates of Park and Gupta (2012) as follows: The Epanechnikov kernel for the non-parametric density estimation with a bandwidth of 0.0225 allowed us to obtain the copula term for both store 1 and store 2. Note: Park and Gupta (2012) report using bandwidths of 0.036 and 0.039 for store 1 and store 2, respectively. However, these bandwidths yielded coefficient estimates that differed from the ones they reported in Table 11 of their research. Furthermore, the ρ parameter with an estimated value of 0.197 (0.080) for store 1 (store 2) that Park and Gupta (2012) report in their Table 11, is not an estimate of the log(Price) copula's slope, but an estimate of the correlation between the error term and the copula term. Based on Eq. 10 in Park and Gupta (2012), we replicate their estimated ρ by multiplying the copula's estimated slope with the estimated standard deviation from an OLS regression without any copula terms.

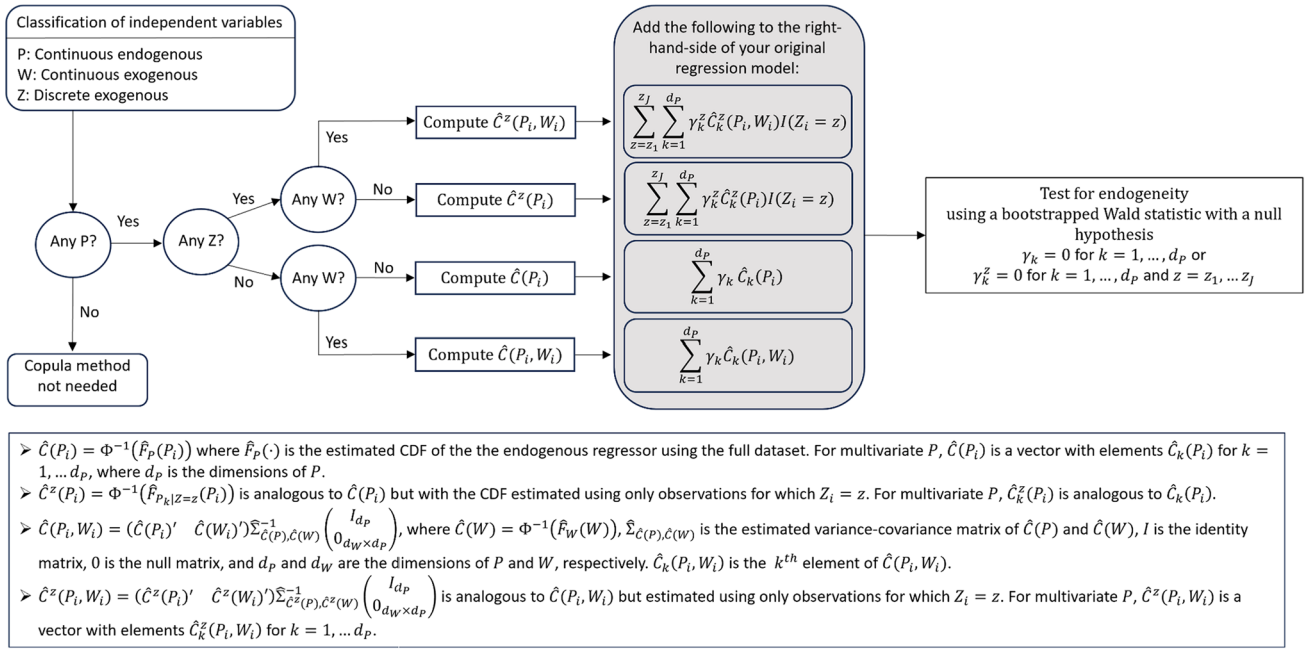

Fig. 3 Flowchart to apply to the Gaussian copula approach

term ignores the dependence structure between the variables in the regression system, and is only based on the log(Price).

In comparison to the OLS model without the copula terms, the Gaussian copula approach's results when using the adjusted ECDF show a larger (more negative) coefficient for Price and a smaller coefficient for Display in both store 1 and store 2; importantly, the coefficient on Display is not statistically significant in either store. Although most of the coefficients seem similar to those of Park and Gupta (2012; see their Table 11), which we also replicate in our research (Table 2; column "P&G Price copula"), our approach specifically results in differences in the log(Price)'s slope coefficient (Δ 0.126) and in the Display (Δ 0.063) in store 1. These differences are less pronounced for store 2. Based on our methodological considerations and on our simulation studies' findings, we anticipated differences between the original outcomes that Park and Gupta (2012) present and our framework, which includes using the adjusted ECDF that we propose.

Finally, researchers often want to test for endogeneity. In Step 4, we follow Park and Gupta (2012) and assess the statistical significance of the copula terms' coefficients using bootstrapping. However, we consider a larger number of bootstrap subsamples (i.e., 5,000) to obtain the standard errors, confidence intervals, and p-values.[15] In the case of a single endogenous variable, Park and Gupta (2012) demonstrate that scaling the copula term's parameter by the error term's variance in a regression without copula correction represents the correlation between the error term and the endogenous variable. A test of this parameter could, therefore, act as a test of endogeneity like the Hausman test (Hausman, 1978) as suggested by Papies et al. (2017). Becker et al. (2022) show that this test has considerable power issues with smaller sample sizes, and advise researchers not to over-rely on such tests. However, our more general framework requires a different test to determine the copula terms' joint significance. We recommend using a bootstrapped version of the Wald test (Wald, 1943),[16] which we apply to our example with the null hypothesis that $\gamma_{Price}^{q_1} = \gamma_{Price}^{q_2} = \gamma_{Price}^{q_3} = \gamma_{Price}^{q_4} = 0$ to determine if endogeneity is present in the proposed model. Specifically, if the null hypothesis is rejected, this indicates that at least one copula term's coefficient is different from zero, providing evidence of endogeneity. The p-value of 0.494 (0.773) determined for store 1 (store 2) does not support significant

[15] Owing to space constraints, Table 2 only reports the bootstrapped standard errors and uses asterisks for typical p-value cut-offs. In Table OA.11-1 in the Online Appendix, we provide the bootstrapped confidence intervals.

[16] Alternatively, one could also test for endogeneity using a Hausman test (Hausmann, 1978). Researchers should use the residual bootstrap approach that Wong (1996) describes for this test. However, we suggest using the Wald test (Wald, 1943), because it is relatively easy to apply and could be generalized to extensions with heteroscedastic error terms.

price endogeneity.[17] Consequently, we find no support for the claim that unmeasured product characteristics or demand shocks influence both consumer decisions and retailer pricing decisions.

Nevertheless, we do not advise researchers to take this result as an indication to use the uncorrected OLS model for interpretation. First, theoretical considerations strongly suggest the presence of endogeneity. Second, from previous research (Becker et al., 2022) we know that such a test's power could be low, resulting in false negatives. Third, our simulation studies have shown that the Gaussian copula approach with the adjusted ECDF estimator does not worsen the endogeneity bias in the model. Consequently, and to avoid pre-testing issues, researchers should consider the copula model results for their theoretically assumed endogenous variables, even if the copula terms are not significant.

Despite Park and Gupta (2012) only providing theoretical arguments for retail price endogeneity, we further extend the empirical example to simultaneously illustrate the analysis of multiple, theoretically assumed, endogenous variables. The unmeasured product characteristics or the demand shocks influencing both consumer and retailer decisions may not only affect the retail price, but also the retailer's promotion and display decisions. This should result in strong correlations between log(Price), Promotion, and Display. In this empirical example, we indeed see that log(Price) is negatively correlated with Promotion (-0.57 store 1; -0.50 store 2) and Display (-0.65 store 1; -0.55 store 2), while Promotion and Display show a strong positive correlation (0.64 store 1; 0.58 store 2). This indicates a higher Display and Promotion intensity at lower prices and vice versa. It therefore seems reasonable to assume that the unobserved demand shocks captured in the error term are not only correlated with log(Price), but also with the advertising variables Promotion and Display, thereby making all three variables endogenous (Step 1).

With respect to Step 2, we already demonstrated all three regressors' non-normality, which also allows us to apply the proposed framework in this extension of the empirical example. Regarding Step 3, when following the flowchart in Fig. 3, we designate log(Price), Promotion, and Display as endogenous regressors P, and the year quarters represented by the categorical exogenous regressors Z, while we have no continuous exogenous regressors W. We therefore need to include copula terms for all three endogenous regressors in the model. We interact each of the three copula terms with the indicator variables representing the year's quarters, giving the following regression model:[18]

$$\begin{aligned}\log\left(Sales_t\right) = {} & \alpha + \beta_1 \cdot \log\left(Price_t\right) + \beta_2 \cdot Promotion_t + \beta_3 \cdot Display_t \\ & + \beta_4 \cdot I(Z_t = q_2) + \beta_5 \cdot I\left(Z_t = q_3\right) + \beta_6 \cdot I(Z_t = q_4) \\ & + \sum_{j=1}^{4} \gamma_{Price}^{q_j} \widehat{C}_{Price,t}^{q_j} I\left(Z_t = q_j\right) + \sum_{j=1}^{4} \gamma_{Promotion}^{q_j} \widehat{C}_{Promotion,t}^{q_j} I\left(Z_t = q_j\right) \\ & + \sum_{j=1}^{4} \gamma_{Display}^{q_j} \widehat{C}_{Display,t}^{q_j} I\left(Z_t = q_j\right) + u_t. \end{aligned} \tag{28}$$

We estimate the copula terms separately for observations belonging to each of the four quarters by using the adjusted ECDF that we propose. As there are no W variables in the model, we only use the endogenous regressor's estimated CDF to compute each copula term. Table 2 shows store 1 and store 2's results (i.e., in the "Proposed: Price, Promotion, and Display copulas" columns 5 and 9).

Finally, Step 4 includes the significance test and interpretation of the results. Despite certain individual copula terms being significant, the results suggest that endogeneity is not a critical issue since a bootstrapped Wald test for the copula terms' joint significance returns a p-value of 0.211 (0.304) for store 1 (store 2). Nevertheless, as previously explained, we use the copula model's results rather than the naïve OLS. We focus on a comparison between the estimates resulting from our proposed framework (i.e., with respect to store 1,

[17] In both our replication of Park and Gupta's (2012) results, and in the analysis using the adjusted ECDF that we propose, we obtain the outcome that there is no critical price endogeneity problem for store 1 or store 2 (since the copula terms are non-significant). At first glance, this result may seem different from that of Park and Gupta (2012), who show that their modified copula term (i.e., the ρ-parameter) for log(Price) is statistically significant in store 1 (but not in store 2). They therefore concluded that there is a significant endogeneity issue in store 1's regression model, which the Gaussian copula approach revealed and treated by providing bias-corrected model estimations. However, Park and Gupta (2012) consider the ρ-parameter's empirical t-value of 1.8 for store 1 as an indication of a statistically significant outcome. This finding suggests that Park and Gupta (2012) use a two-sided test with a 10% probability of error or a one-sided test with a 5% probability of error to determine the significance tests' critical t-value. Since both copula terms and Park and Gupta's (2012) ρ-parameter can have positive and negative values, and we could not find an a priori strong hypothesis for the error correlation's direction, we used a two-sided test, for which we chose a stricter 5% probability of error. With such a test, Park and Gupta's (2012) result would also have been non-significant. This means that their results are in line with the results and conclusions that we present.

[18] If theoretically justified, a researcher could easily introduce non-linearities into the model, such as an interaction between log(Price) and Promotion, and/or a squared Display variable, without altering the calculation of the copula terms or the part of the model that incorporates these terms. This is because: (1) the calculation of the copula terms is only affected by the classification of the continuous variables as either endogenous or exogenous, and (2) the copula's interactions with discrete variables are intended to allow the copula structure to vary across categories of the discrete variables, rather than being based on the functional form of the structural model.

Table 2 Paper towel sales example

Parameters	Store 1				Store 2			
	OLS model	*P*&G:	Proposed:		OLS model	*P*&G:	Proposed:	
		Price copula	Price copula	Price, Promotion, and Display copulas		Price copula	Price copula	Price, Promotion, and Display copulas
Intercept	6.607*** (0.037)	6.600*** (0.037)	6.619*** (0.039)	6.561*** (0.136)	6.549*** (0.023)	6.549*** (0.024)	6.558*** (0.026)	6.452*** (0.067)
log(Price)	-0.676*** (0.148)	-0.932** (0.332)	-0.806*** (0.267)	-0.713*** (0.265)	-0.780*** (0.131)	-0.882*** (0.478)	-0.912*** (0.294)	-0.876** (0.383)
Promotion	0.407*** (0.046)	0.395*** (0.047)	0.396*** (0.048)	0.577*** (0.188)	0.433*** (0.035)	0.430*** (0.035)	0.423*** (0.034)	0.300 (0.221)
Display	0.173** (0.084)	0.209** (0.088)	0.146 (0.093)	0.243 (0.459)	0.158** (0.073)	0.163** (0.076)	0.136 (0.084)	0.678*** (0.257)
$I(Z = q_2)$	0.094** (0.039)	0.099** (0.039)	0.086** (0.038)	0.128 (0.080)	0.089*** (0.026)	0.090*** (0.026)	0.083*** (0.029)	0.127*** (0.038)
$I(Z = q_3)$	0.055 (0.034)	0.056 (0.034)	0.045 (0.035)	0.043 (0.066)	0.116*** (0.024)	0.115*** (0.024)	0.112*** (0.026)	0.097 (0.067)
$I(Z = q_4)$	-0.067** (0.034)	-0.066* (0.034)	-0.079** (0.036)	-0.016 (0.073)	0.060** (0.029)	0.059** (0.029)	0.051 (0.033)	0.084 (0.123)
c: log(Price)		0.037 (0.031)				0.012 (0.034)		
c: log(Price)*$I(Z = q_1)$			-0.011 (0.020)	0.035 (0.085)			0.009 (0.018)	0.044 (0.041)
c: log(Price)*$I(Z = q_2)$			0.022 (0.030)	0.055 (0.056)			0.008 (0.016)	0.015 (0.061)
c: log(Price)*$I(Z = q_3)$			0.029 (0.022)	0.036 (0.045)			0.009 (0.020)	0.032 (0.048)
c: log(Price)*$I(Z = q_4)$			0.017 (0.016)	0.060 (0.063)			0.021 (0.020)	0.068* (0.039)
c: Promotion*$I(Z = q_1)$				-0.099* (0.060)				-0.003 (0.051)
c: Promotion*$I(Z = q_2)$				-0.046 (0.045)				0.007 (0.067)
c: Promotion*$I(Z = q_3)$				-0.012 (0.052)				-0.009 (0.058)
c: Promotion*$I(Z = q_4)$				-0.047 (0.045)				-0.020 (0.033)
c: Display*$I(Z = q_1)$				-0.063 (0.099)				-0.054 (0.051)
c: Display*$I(Z = q_2)$				-0.056 (0.074)				-0.099** (0.048)
c: Display*$I(Z = q_3)$				-0.023 (0.080)				-0.064 (0.073)
c: Display*$I(Z = q_4)$				-0.049 (0.080)				-0.068 (0.052)

The table shows the regression coefficients with bootstrapped standard errors in parentheses; P&G = Replication of Park and Gupta's (2012, see their Table 11) results using our own code and the Epanechnikov kernel with a 0.0225 bandwidth; OLS = ordinary least squares; Proposed = the adjusted ECDF's results regarding the copula terms' estimator and calculations; c = copula

***$p < 0.01$; **$p < 0.05$; *$p < 0.10$

Table 3 Comparison of methodological approaches for endogeneity correction using Gaussian copula estimation and related methods

Approach	Key features/contributions	When to use?
Park and Gupta (2012)	• Introduces the Gaussian copula approach to the literature. • Suggests a maximum likelihood and control function approach. • Applicable for linear regression models and non-linear models, such as the random coefficient logit models. • Proposes estimation of the empirical distribution of the regressor using non-parametric kernel-based methods.	• If the model has a single or multiple endogenous regressors. • If exogenous regressors are not correlated with the endogenous regressors. In this case, discrete exogenous, non-linearities and interactions may also be used.
Haschka (2022)	• Shows that the original approach requires uncorrelated endogenous and exogenous regressors and extends the model to include correlated continuous endogenous and exogenous regressors. • Modifies maximum likelihood approach to account for correlations between endogenous and exogenous regressors. • Suggests within transformation or first differences for panel models and then assumes a joint distribution on the transformed variables.	• If using a panel data model with a single or multiple continuous endogenous regressors • If it is credible that the first differenced or within transformed variables have a joint distribution. • If endogenous regressors should be allowed to correlate with exogenous regressors.
Yang et al. (2022)	• Two-stage control function approach. • Provides theoretical justification for the copula methods. • Accounts for correlations between continuous endogenous and exogenous regressors. • Relaxes normality assumption of endogenous regressors if sufficient non-normality is available in correlated exogenous regressors.	• If endogenous regressors should be allowed to correlate with exogenous regressors. • If the model has normally distributed continuous endogenous regressors, but also includes non-normal exogenous regressors that are correlated with them.
This research	• Shows that the estimation quality of the CDF matters for performance and suggests an updated ECDF estimator. • Extends control function approach. • Allows continuous endogenous, and discrete and continuous exogenous variables that are allowed to correlate as well as non-linearities such as interactions for all independent variables. • Allows the copula structure to vary over categories of discrete exogenous variables.	• If the model requires extending Yang et al. (2022) to accommodate all types of non-linearities in the independent variables. • If the analysis includes discrete exogenous variables. • If the copula structure can vary across exogeneous categories (e.g. seasonally varying endogeneity between sales and price). • If an intercept is included, use the proposed ECDF estimator (applies to all previous Gaussian copula-based approaches)
Breitung et al. (2024)	• Suggests a nonparametric control function approach and studies its asymptotic and small sample properties. • Relaxes the assumption of the error term normality and Gaussian copula correlation structure.	• If the researcher suspects that the error term is not normally distributed and/or the Gaussian copula assumption does not hold. • If the researcher can assume a linear relationship between endogenous and all exogeneous variables and that the part of the endogenous regressor that is correlated with the error term is non-normal.
Qian and Xie (2024)	• Proposes a semiparametric odds ratio (SOR) model that nests the copula approach of Park and Gupta (2012) and allows for non-copula dependence between the error term and the endogenous regressor. • The proposed one-step estimation procedure enhances efficiency and simplifies inferential processes.	• If the model has discrete endogenous regressors* and there is dependence between regressors. • If high computational power is available and the researcher can implement complex model setups.

*Methods based on the Gaussian copula correlation structure, such as those by Park and Gupta (2012) and the approach proposed in this paper, can theoretically accommodate discrete endogenous regressors with more than two categories. However, including discrete endogenous variables in the model increases multicollinearity and results in poor finite sample performance. Therefore, we do not recommend using these methods when the endogenous regressor has only a few discrete categories

we compare the "Proposed: Price copula" column 4 and the "Price, Promotion, and Display copulas" column 5; regarding store 2, we compare the "Proposed: Price copula" column 8 and the "Proposed: Price, Promotion, and Display copulas" column 9). Regarding store 1, the coefficients on Promotion and Display increase substantially when they are considered endogenous together with log(Price) (Table 2, column 5) compared to the situation when only log(Price) is considered endogenous (Table 2, column 4). Regarding store 2, the coefficient on Promotion shows a substantial increase, and Display a substantial decrease when these variables are considered endogenous together with log(Price) (Table 2, column 9) compared to the situation when only log(Price) is considered endogenous (Table 2, column 8). However, in this extended example the sample size could be a limiting factor. The 260 observations may not be enough to reliably estimate such a complex model with multiple copulas and using our extended framework's full flexibility.

Conclusions and future research

Endogeneity is a common problem in regression analyses of non-experimental data that leads to biased results in marketing research and other disciplines (e.g., Rutz & Watson, 2019; Shugan, 2004). A typical issue of the popular IV approach that researchers often encounter is a lack of suitable IVs that fulfill both the relevance and exclusion restriction condition (e.g., Rossi, 2014; Sande & Ghosh, 2018). It is also unclear whether the IV approach will ultimately improve or worsen the situation (i.e., the cure can be worse than the disease; e.g., Bound et al., 1993). To overcome these concerns, IV-free methods have been proposed to help researchers detect and address endogeneity problems without the need for additional instruments. Among the IV-free methods to deal with endogeneity issues, Park and Gupta's (2012) Gaussian copula approach is increasingly popular (see the review presented by Becker et al., 2022). However, Becker et al. (2022) reveal the Gaussian copula approach's limited applicability when regression models consider an intercept, which other studies confirm (e.g., Eckert & Hohberger, 2023; Falkenström et al., 2023).

In this research, we use theoretical derivations to show that the bias depends on the CDF estimator's quality. Based on these insights, we propose an adjusted ECDF estimator for the Gaussian copula approach that mitigates the endogeneity bias problem in regression models with and without an intercept. Simulation studies highlight this new estimator's strong performance in numerous data situations that researchers are likely to encounter in empirical applications. At the same time, we compare the new estimator's results with those of other established estimators. We thereby further substantiate that the choice of CDF estimator matters, and that the adjusted ECDF performs effectively in the Gaussian copula approach, consistently outperforming other established estimators in terms of a lower finite sample bias (i.e., when sample sizes are relatively small). Additionally, our analytical derivations identify another key factor influencing the bias, namely model collinearity. This refers to the correlation between the endogenous variable and the copula term in simple models, or the determinant ω in multivariate models. We show that this correlation (and ω) is also connected to the non-normality level in the independent variables.

Another critical methodological research question concerns the Gaussian copula approach's efficacy when handling multiple endogenous and exogenous regressors with different scales (i.e., discrete or continuous scales) and non-linearities (e.g., interaction terms). Our additional methodological advances enable the creation of a more general Gaussian copula framework that enables researchers to more effectively address endogeneity problems in a large variety of regression models. We illustrate the practical application of our advancements by means of an empirical marketing example with real-world data. This empirical example takes researchers step-by-step through the application of the more general framework we propose and highlights the core decisions that researchers must make in the process. In addition, the example provides researchers and practitioners with valuable guidance in their quest for empirical rigor by reducing the bias resulting from endogeneity in their regression models.

Our research unveils novel insights and contributes to prior knowledge about the Gaussian copula approach for dealing with endogeneity issues in regression models by (1) developing an adjusted estimator and (2) presenting a new comprehensive and adaptable framework capable of accommodating multiple correlated endogenous and exogenous variables in regression models with and without intercept. This framework is designed to handle exogenous variables with discrete (e.g., binary) and continuous scales and can capture non-linearities, including interaction terms in endogenous and exogenous variables. Consequently, this research contributs to a valid application of the Gaussian copula approach. Table 3 provides a summary of our research's contribution relative to other recent advances on the Gaussian copula approach and related methods (see also Park & Gupta, 2024).

However, these research results also have limitations. Our general Gaussian copula framework requires an a priori specification of the exogenous and the endogenous variables from the set of available regressors. Consequently, researchers still need to assess their variables' causal structure carefully. For any variable for which we cannot argue for exogeneity, we must include additional copula terms. Otherwise, the results will suffer from endogeneity bias.

However, including more copula terms increases model complexity and data requirements, which can also increase the likelihood that some of the bias will remain and that tests will have low power. Future research can further investigate the trade-off between increasing the model complexity by including additional copula terms and misspecifying variables as exogenous. Moreover, including multiple copula terms into a model also increases the chance of one being significant (i.e., a multiple testing problem). In our empirical example, we propose a bootstrap-based Wald test to assess the copula terms' joint significance. Future research can assess this test's power, and compare it to other potential approaches, such as a bootstrapped Hausman test. In addition, regarding our most general model, which also includes discrete regressors, we suggest estimating the copulas per discrete level of the categorical variable. Our simulation studies show that if the copula structure varies with the discrete variable's categories, only the conditional estimation yields unbiased results. Future research can develop tests that might help researchers decide whether this additional model complexity is necessary.

Our general framework's nature should also make it easily applicable to panel data models by using within transformations or the dummy variable approach. Future research could deepen this contributions' relevance by using additional simulation studies to show our extended framework's performance with respect to models using panel data, and by using empirical examples to develop related guidelines for applications. Similarly, future research could investigate extensions of the more general framework to limited dependent variables (e.g., binary outcomes, such as in binary choice models).

Finally, like the original approach, the proposed more general Gaussian copula framework requires compliance with central assumptions concerning (1) a normal error distribution, (2) non-normal endogenous regressor(s), and (3) fulfillment of a Gaussian copula structure. Our findings do not alleviate general concerns about violating these central assumptions, such as the fulfillment of a Gaussian copula structure or a normal error distribution, which, as Becker et al. (2022) show, could have a substantial impact.[19] While this research relaxes some of the requirements regarding non-normality (e.g., through higher quality CDF estimation by using our adjusted ECDF estimator that has a lower finite sample bias), sufficient non-normality is still important. Moreover, in the general framework that we introduce, it is no longer the non-normality of a single variable that matters, but rather the presence of sufficient non-normality in at least some of the variables that we leverage to identify the parameters. We also show that a potential way of quantifying this non-normality is through a determinant based on the predictor matrix that we denote as ω, which is a key component of the approaches' bias. These findings could guide future research efforts in terms of establishing guidelines for the Gaussian copula method's viability, based on the sample size and ω. Future research may provide guidelines by using empirically relevant simulations to explore the ω values' boundary conditions where the Gaussian copula approach is likely to yield low bias. Such investigations could provide better thresholds for characterizing the necessary level of non-normality than simple univariate tests of each variable's non-normality.

Any currently available method aimed at correcting endogeneity bias depends on certain assumptions. Some of these are untestable, like the Gaussian copula structure. Regarding, for example, the instrumental variables approach, the exclusion restriction is the most problematic assumption. However, researchers can use marketing theory to provide careful arguments for their instruments' exclusion restriction, and the quality of the instruments can be judged by other researchers based on these arguments. Conversely, there is a lack of guidance on how to justify the unobservable Gaussian copula structure. We therefore recommend that Gaussian copula methods should only be applied after a careful search for suitable IV's. In any case, it will be a pressing concern of future research to find ways to allow researchers to better assess whether the assumptions of the Gaussian copula are met. To address these considerations in practice, researchers could triangulate the causal effect of interest by means of multiple methods, such as IV and frugal IV-free approaches like the Gaussian copula approach (see Ebbes et al., 2009, for other frugal IV-free methods).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11747-024-01055-4.

Authors' contributions All authors equally contributed to the manuscript. However, Benjamin D. Liengaard particularly focused on the methods development and designing the simulations and deriving conclusions from the data analysis, Jan-Michael Becker particularly focused on designing the simulations and deriving conclusions from the data analysis as well as the positioning of the paper, Mikkel Bennedsen, Phillip Heiler, and Luke N. Taylor particularly focused on the methods development, and Christian M. Ringle particularly focused on the positioning of the paper and the empirical example. All authors reviewed the results and approved the final version of the manuscript.

[19] Breitung et al. (2024) use a semiparametric approach that does not require a full distributional model (i.e., a Gaussian copula) for the dependence structure but instead uses a linear relationship between endogenous and all exogeneous variables. In contrast, our method is essentially unrestricted with respect to discrete exogeneous variables and models any dependence between continuous exogenous and endogenous directly via a copula. The latter implies a linear relationship between the respective quantile-CDF transformations that enter as control functions.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Not applicable.

Declarations

Ethical approval Not applicable.

Competing interests Although this research does not use the statistical software SmartPLS (https://www.smartpls.com), Jan-Michael Becker and Christian M. Ringle acknowledge that they are co-developers and co-founders of SmartPLS and, as such, have a financial interest in SmartPLS.

References

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

Bound, J., Jaeger, D. A., & Baker, R. M. (1993). The cure can be worse than the disease: A cautionary tale regarding instrumental variables. *NBER Technical Paper Series*, Technical Paper No. 137.

Breitung, J., Mayer, A., & Wied, D. (2024). Asymptotic properties of endogeneity corrections using nonlinear transformations. *The Econometrics Journal*, forthcoming.

Bronnenberg, B. J., Kruger, M. W., & Mela, C. F. (2008). Database paper—the IRI marketing data set. *Marketing Science, 27*(4), 745–748.

Ebbes, P., Wedel, M., & Böckenholt, U. (2009). Frugal IV alternatives to identify the parameter for an endogenous regressor. *Journal of Applied Econometrics, 24*(3), 446–468.

Eckert, C., & Hohberger, J. (2023). Addressing endogeneity without instrumental variables: An evaluation of the Gaussian copula approach for management research. *Journal of Management, 49*(4), 1460–1495.

Falkenström, F., Park, S., & McIntosh, C. N. (2023). Using copulas to enable causal inference from nonexperimental data: Tutorial and simulation studies. *Psychological Methods, 28*(2), 301–321.

Gui, R., Meierer, M., Algesheimer, R., & Schilter, P. (2022). *R package REndo: Fitting linear models with endogenous regressors using latent instrumental variables*, version 2.4.7.

Haschka, R. E. (2022). Handling endogenous regressors using copulas: A generalization to linear panel models with fixed effects and correlated regressors. *Journal of Marketing Research, 59*(4), 860–881.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*(6), 1251–1271.

Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software, 27*(5), 1–32.

Hill, T. P., & Mann, J. (2000). Alternative empirical distributions based on weighted linear combinations of order statistics. *Stochastic Analysis and Applications, 18*(1), 87–99.

Jean, R. J. B., Deng, Z., Kim, D., & Yuan, X. (2016). Assessing endogeneity issues in international marketing research. *International Marketing Review, 33*(3), 483–512.

Li, C., Li, H., & Racine, J. S. (2017). Cross-validated mixed-datatype bandwidth selection for nonparametric cumulative distribution/survivor functions. *Econometric Reviews, 36*(6–9), 970–987.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, *25*(1), 221–247.

Papies, D., Ebbes, P., & van Heerde, H. J. (2017). Addressing endogeneity in marketing models. In P. S. H. Leeflang, J. E. Wieringa, T. H. A. Bijmolt, & K. H. Pauwels (Eds.), *Advanced methods in modeling markets* (pp. 581–627). Springer.

Park, S., & Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science, 31*(4), 567–586.

Park, S., & Gupta, S. (2024). A review of copula correction methods to address regressor–error correlation. *Impact at JMR* (May 15, 2024).

Qian, Y., & Xie, H. (2024). Correcting regressor-endogeneity bias via instrument-free joint estimation using semiparametric odds ratio models. *Journal of Marketing Research*, *61*(5), 916–936.

Qian, Y., Koschmann, A., & Xie, H. (2024). A practical guide to endogeneity correction using copulas. *NBER Working Paper Series*, Working Paper 32231.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

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

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

Sande, J. B., & Ghosh, M. (2018). Endogeneity in survey research. *International Journal of Research in Marketing, 35*(2), 185–204.

Shugan, S. M. (2004). Endogeneity in marketing decision models. *Marketing Science*, *23*(1), 1–3.

Silverman, B. W. (1998). *Density estimation for statistics and data analysis*. Chapman & Hall.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54*, 426–482.

Wong, K. (1996). Bootstrapping Hausman's exogeneity test. *Economics Letters*, *53*(2), 139–143.

Yang, F., Qian, Y., & Xie, H. (2022). Addressing endogeneity using a two-stage copula generated regressor approach. *NBER Working Series*, Working Paper 29708.

Zaefarian, G., Kadile, V., Henneberg, S. C., & Leischnig, A. (2017). Endogeneity bias in marketing research: Problem, causes and remedies. *Industrial Marketing Management, 65*, 39–46.