Gaussian Copula Regression in the Presence of Thresholds

Christine Eckert, University of Technology, Sydney
Jan Hohberger, Esade Law and Business School, Barcelona
Philip Hans Franses, Erasmus School of Economics, Rotterdam

Econometric Institute Report 2022-02

Abstract:

Park and Gupta's (2012) introduction of the Gaussian Copula (GC) approach to deal with endogeneity has made a significant impact on empirical marketing research with many papers using this approach. Recent studies have however started to explore and examine the approach and its underlying assumptions more closely, resulting in a more critical picture of it. A particular challenge is the non-testable assumption that the dependency structure between the endogenous regressor and the error term should be described by a Gaussian copula. In general, there exists a limited understanding of what this assumption implies, what causes its violation, and potential remedies. Our study addresses this explicitly. We provide a detailed discussion of the dependency structure assumption and how thresholds in the data can lead to its violation and biased results. We use real and simulated data to show how threshold detection *before* applying the GC approach can overcome this problem and thereby provide researchers with a useful tool to increase the likelihood of the GC approach's assumptions being met.

Keywords: Gaussian Copula, Endogeneity, Threshold Regression, Research Methods

JEL codes: C13, C24, C26

February 2022

1 Introduction

Empirical marketing research frequently relies on observational data to inform the theoretical understanding of marketing phenomena and aid managerial decision-making. Providing accurate and unbiased inferences for observational data can however be challenging. In regression analysis, endogeneity – a situation where the explanatory variable is correlated with the unobserved error term - can lead to biased coefficients and thus to misleading interpretations of the data (Rutz & Watson, 2019; Wooldridge, 2010). Endogeneity issue have become a central concern of empirical marketing research during the last two decades and often features on top of the list of reviewers' comments (Hult et al., 2018; Rutz & Watson, 2019; Zaefarian, Kadile, Henneberg, & Leischnig, 2017).[1]

One relatively recent approach dealing with endogeneity is the Gaussian Copula (GC) approach by Park and Gupta (2012). Reflective of the big obstacle researchers often face when confronted with endogeneity concerns, as well as thanks to its flexibility and ease of implementation, the GC approach experienced an enthusiastic uptake within the marketing area (e.g. Datta, Ailawadi, & van Heerde, 2017; Guitart & Stremersch, 2021; Lenz, Wetzel, & Hammerschmidt, 2017; Vomberg, Homburg, & Gwinner, 2020) and is becoming also more popular in related management fields (e.g. Becerra & Markarian, 2021; Haschka & Herwartz, 2020; Reck, Fliaster, & Kolloch, 2021). However, like with any statistical approach, the performance of the GC approach hinges on assumptions, which in this case are: 1) the error term should be normally distributed 2) the endogenous regressor must be sufficiently non-normally distributed 3) the regressor needs sufficient support and 4) the dependency structure between the endogenous regressor and the error term can be described by a Gaussian copula.

[1] We refer the reader to the comprehensive overview by Rutz and Watson (2012) on methods used in marketing research to control for endogeneity.

In their initial paper, Park and Gupta (2012) provide an extensive set of simulations to test the sensitivity of the method to these assumptions and conclude that robustness to violations of these assumptions is mostly given. Two very recent studies by Becker, Proksch, and Ringle (2021) and Eckert and Hohberger (2022) however paint a less rosy picture of the GC approach. Both studies explore the different assumptions of the GC approach in more detail. Using a comprehensive set of simulations they show that under more realistic conditions (e.g., the inclusion of a constant in the estimation equation, non-normal errors), the GC approach may provide estimates that are far from the true values. Both studies particularly stress that meeting the fourth assumption, namely that a Gaussian copula describes the dependency structure between error term and endogenous regressor, is crucial for the performance of the approach. Yet, marketing researchers so far lack guidance regarding what this assumption implies. For example, Becker et al (2021) note that the dependency structure "is inherently unobservable and therefore only subject to untestable theoretical considerations". Rutz and Watson (2019) equate this assumption with the requirement of constant correlation between error and endogenous regressor, yet their statement is too conservative and only holds under additional assumptions regarding the error term and the endogenous regressor. Danaher and Smith (2011, p. 9) state explicitly that the fact that "the copula function is based on an elliptical distribution, such as the multivariate t or Gaussian, … should not be confused with using an elliptical distribution for the data itself." In fact, Danaher and Smith (2011) instead report that the Gaussian copula is a robust copula for most applications.

Similarly, Eckert and Hohberger (2022) show that the GC approach's dependency structure assumption indeed may hold in many settings of relevance as it can be derived – under specific assumptions - from the different patterns (i.e., measurement error, omitted variable bias, reverse causality) that induce endogeneity. For example, Falkenström, Park,

and McIntosh (2021) use a Gaussian copula data generating mechanism in their simulations to explore the GC approach's performance in the context of omitted variable bias.

One situation that is however not covered amongst these settings is the case of endogeneity thresholds in the data. Thresholds are shifts in economic relationships that lead to regression functions revealing a jump or a kink at a threshold point (Yang, Zhang, Lee, & Chen, 2021). Thresholds are a quite common phenomenon in many economic and organizational settings (Hansen, 2000) and also appear frequently within a marketing context (Biyalogorsky, Gerstner, & Libai, 2001; Han, Gupta, & Lehmann, 2001; Jia, Yang, Lu, & Park, 2018; Pauwels, Srinivasan, & Franses, 2007; Terui & Dahana, 2006; Vastani & Monroe, 2019). For example, marketing academics, as well as practitioners, have long acknowledged that consumers' response to price changes is frequently not monotonic and symmetric but instead subject to thresholds (Pauwels et al., 2007). The fact that small discounts might lead to under-proportional effects on market performance compared to larger discounts has thereby been described in various studies (Gutenberg 1976, Hruschka 2000, Gilbride and Allenby 2004).

Usually, thresholds in marketing are discussed in the context of changes in slopes, such as price elasticities changing if a reference price has been crossed (Pauwels et al., 2007). Thresholds can however also apply to the underlying endogeneity structure of the data, i.e. the degree of endogeneity changes if the endogenous regressor passes a specific endogeneity threshold. For example, in case of measurement error, larger retailers with lower prices may have automated systems to record prices and sales, whereas smaller retailers with higher prices may manually record prices and sales. In these cases, the measurement error and thus induced endogeneity could be lower for lower prices. If the model does not explicitly account for these thresholds in endogeneity, these thresholds will resemble discrete omitted variables that are correlated with both endogenous regressor and error term. In these cases,

the dependency structure between error and endogenous regressor can no longer be described by the Gaussian copula, which in turn might lead to the GC approach providing biased estimates.

The objective of the present paper is therefore to shed more light on the dependency structure between endogenous regressor and error term and focus particularly on the situation of thresholds. We provide a detailed discussion of the Gaussian copula dependency structure assumption, how it is violated in the context of endogeneity thresholds in the model, and how this violation can lead to biased results when using the GC approach. We use real and simulated data to show how threshold detection *before* applying the GC approach can overcome this problem. If endogeneity thresholds are explicitly modeled, the GC approach's assumption regarding the error-regressor dependency structure is more likely to be met and the GC approach is thus more suitable to apply.

Our study contributes to the ongoing discourse on endogeneity as well as the still-nascent discussion about the use of copulas and other IV-free methods in marketing and related fields (Hult et al., 2018; Lewbel, 2019; Papies, Ebbes, & Van Heerde, 2016; Park & Gupta, 2012; Rutz & Watson, 2019; Zaefarian et al., 2017). More specifically it addresses recently discovered challenges of the GC approach (Becker et al., 2021; Eckert & Hohberger, 2022) in regards to the assumption that the dependency structure between error and endogenous regressor can be described using a Gaussian Copula. We provide a tool for researchers to gain more confidence that this assumption is met and offer a solution for a particular, but widely encountered violation.

The paper starts with a review of the GC approach and its assumptions. We then zoom into the assumption regarding the dependency structure between endogenous regressor and error term. We explain its implications, discuss how endogeneity thresholds in the data may violate this assumption, and show the detrimental impact of endogeneity thresholds in the

context of a case study as well as simulations. We then describe our proposed data preparation step of using threshold regression to test for thresholds in the data before applying the GC approach and demonstrate its viability via simulations and two empirical examples. The paper finishes with a discussion of the limitations of the threshold regression approach as well as guidance for its applications.

2 Gaussian Copula Approach to Deal with Endogeneity Revisited

2.1 Reception of the GC Approach

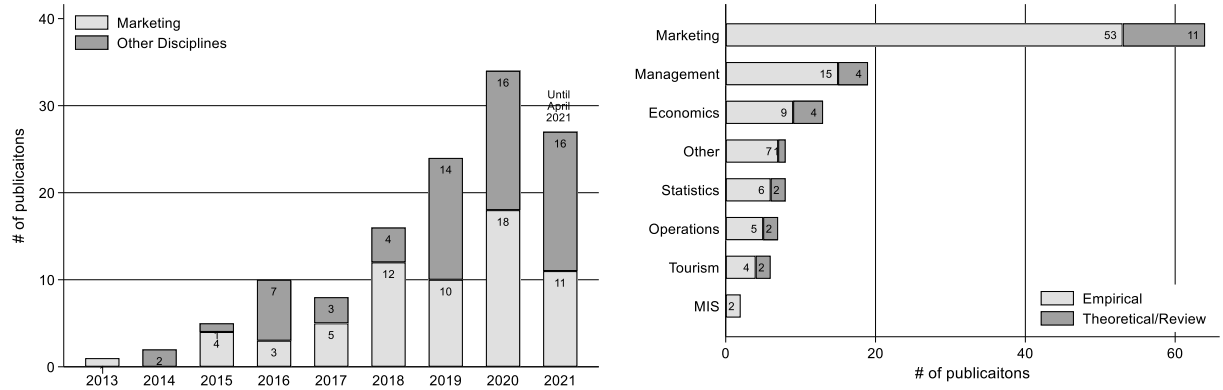
The Gaussian Copula approach was introduced by Park in Gupta in 2012 to address the practical difficulty in applied research to find suitable instruments and to deal with the limitations or complexity of alternative methods. It is part of the family of instrument-free approaches to address endogeneity concerns, which identifies model parameters through statistical assumptions and functional form (Ebbes, Wedel, & Böckenholt, 2009; Lewbel, 2012; Rigobon, 2003).

The GC approach has seen a fast uptake in marketing and related fields. Top marketing journals now frequently publish papers that use the GC approach (e.g. Datta, Ailawadi & van Heerde, 2017; Guitart & Stremersch, 2021; Lenz, Wetzel & Hammerschmidt, 2017; Vomberg, Homburg & Gwinner, 2020) and it features in several review articles on endogeneity (Papies et al., 2016; Rutz & Watson, 2019; Sarstedt et al., 2020). Figure 1 shows the increasing impact of the GC approach in marketing and related fields by the number of references to the original paper by Park and Gupta (2012).[2] Although the majority of studies are within the marketing area, a substantial number of references are

[2] We identified articles by collecting all publications citing the seminal work of Park and Gupta (2012) identified in the Scopus database. Due to its sized and coverage the Scopus database is frequently used in bibliometric studies to analysis the semination of scholarly work (Zupic and Čater, 2015; Wilden et al., 2019). An additional literature search did not find any application of the GC approach who did not cite Park and Guprta (2012).

from related fields (e.g. management, economics, and statistics). Furthermore, although several of the publications are reviews, which discuss and propagate the use of the GC approach (e.g. Papies et al., 2016; Rutz & Watson, 2019), Figure 1 also shows that the majority of publications referring to Park and Gupta (2012) use the GC approach in their empirical application and often as the main analysis method.

Figure 1: Application and discussion of the Gaussian Copula approach



2.2 Conceptual underpinning and assumptions of the GC Approach

The starting point of the GC approach is a simple linear regression model with an endogenous regressor:

$$Y_t = W_t'\beta + X_t\alpha + \varepsilon_t, \quad (1)$$

where *t* indexes observations, Y_t is a (1x1) dependent variable, W_t is a (kx1) vector of exogenous regressors, X_t is a (1x1) endogenous regressor[3], ε_t is a structural error term, and α and β are model parameters. The situation where X_t is correlated with ε_t creates the endogeneity problem and this may result in biased estimates of α (Wooldridge, 2010) as well as β if X_t is also correlated with W_t.

[3] We focus on continuous regressors since the GC approach requires the endogenous regressor to have sufficient support.

The GC approach addresses potential endogeneity issues by acknowledging that if the *joint* distribution $f(X, \varepsilon)$ of the regressor X_t and the error term ε_t was known, then it would be possible to obtain consistent estimates of the model parameters by maximizing the likelihood function derived from this joint distribution. Building on the earlier work by Danaher and Smith (2011), Park and Gupta (2012) propose a unique copula model to infer this joint distribution. Specifically, they make use of Sklar's theorem (Chu, Tian, & Wang, 2018) which states that there exists a copula function C such that the joint cumulative distribution function (CDF) can be derived from the marginal CDFs of X_t and ε_t. Furthermore, once the joint CDF of X_t and ε_t is known, the joint probability distribution function $f(X_t, \varepsilon_t)$ can be calculated and used to obtain the parameters of interest.

Park and Gupta (2012) assume the marginal distribution of ε_t to be normal (e.g. Kleibergen & Zivot, 2003; Park & Gupta, 2012; Wooldridge, 2010). They derive the marginal CDF of P_t based on observed realizations of X_t using nonparametric density estimation. Finally, they assume that the link between the marginal distributions of X_t and ε_t can be described by a Gaussian copula. Park and Gupta (2012) show that the probability distribution function $f(X, \varepsilon)$ leads to the following regression equation, which provides another way – and so far the most popular (see Becker et al., 2021) – to estimate the model:

$$Y_t = W_t'\beta + X_t\alpha + X_t^*\alpha^* + \varepsilon_t^*, \tag{2}$$

with α^* another parameter to be estimated, ε_t^* normally distributed white noise, and the additional regressor defined as $X_t^*\alpha^* = \Phi^{-1}\left(\widehat{F_P}(X)\right)$, where Φ denotes the standard normal CDF and $\widehat{F_P}$ the observed empirical CDF of X. Least squares regression then provides consistent estimates of the parameters and bootstrapping can be applied to obtain correct standard errors[4]. As such, the GC approach simply adds the additional regressor X_t^* to the

[4] Bootstrapping is required due to the method estimating a quantity in step 1, and using this quantity as regressor in step 2.

model, which makes it easy also in the case of multiple endogenous regressors or interactions of endogenous variables (Papies et al., 2016).

Park and Gupta (2012) note the different conditions for their proposed approach to work. First, the support of the endogenous regressor *X* has to be "rich enough for the proposed method to handle the endogeneity problem properly" (Park & Gupta, 2012, p. 547). Park and Gupta (2012) show that the GC approach performs badly when this condition is not met, such as in the case of binary regressors, or in the case of other discrete regressors with limited support. This makes the GC approach not suitable to deal with endogeneity that is due to self-selection into a limited number of outcomes (Feinberg, Salisbury, & Ying, 2016).

Second, Park and Gupta (2012) note that for the GC approach to be identified, the regressor must be "sufficiently" non-normally distributed. As the distribution of the endogenous regressor approaches the normal distribution, the correlation between *X* and X_t^* increases, resulting in a multicollinearity problem, that – in severe cases – can lead to insufficient identification of parameters. Third, for the GC approach to be identified, Park and Gupta (2012) assume that the error term is normal distributed. Finally, the GC approach assumes that the dependency between the endogenous regressor and the error term can be described by a Gaussian copula.

Park and Gupta (2012) provide several simulations to demonstrate that their approach is robust to violations of these assumptions. Yet, recent work by Becker et al. (2021) and Eckert and Hohberger (2022) highlighted several limitations of the approach and argued that the method provides less useful and more biased estimates if the different assumptions are not met or the number of observations is too low. Further, Becker et al. (2021) challenged the implementation of the simulations by Park and Gupta and noted that the omission of a constant in both the simulation and estimation equations by Park and Gupta artificially boosted the performance of the GC approach. Importantly, both studies argue that the GC

approach produces highly biased estimates in cases where the dependency structure assumption regarding error term and endogenous regressor is not met. They note that this issue is particularly troubling as this dependency structure is not observable and it is very difficult to provide empirical or theoretical justification for it.

3 Dependency Structure Between Error and Endogenous Regressor

In the previous section, we introduced the GC approach and discussed prior work examining the robustness of the method to violations of its assumptions. Thereby we highlighted the initial challenges and discussion of the assumption that the dependency structure between error and endogenous regressor can be described using a Gaussian copula (Becker et al., 2021; Eckert & Hohberger, 2022). In this section, we will provide a more focused and detailed discussion of this assumption. We provide an illustration of its implication, discuss its relevance for marketing research and demonstrate how endogeneity thresholds violate the assumption using an empirical example as well as in simulations.

3.1 The Gaussian Copula Assumption

The GC approach is based on the idea that if the researcher knows the joint distribution of the endogenous regressor and the error term $f(X, \varepsilon)$, then they can obtain consistent estimates of the model parameters by maximizing the regression model's likelihood function derived from this joint distribution. In reality, this joint distribution is unknown and the researcher only observes realizations X_t and ε_t of it. Park and Gupta's approach is now to assume that the dependency structure of the error and the endogenous regressor can be modeled via a Gaussian copula. The implications of this assumption are revealed in the following.

Let $h(X)$ and $g(\varepsilon)$ be the *marginal densities* and $U_X=H(X)$ and $U_\varepsilon=G(\varepsilon)$ the *marginal distribution* functions of X and ε (remember that ε is assumed normal distributed),

respectively. If the dependence between X and ε can be modeled via the Gaussian copula, then this implies that the *joint bivariate distribution function F(X, ε)* can be written as $F(X,\varepsilon) = C(U_X, U_\varepsilon) = \Psi_X\big(\Phi^{-1}(U_X), \Phi^{-1}(U_\varepsilon)\big): [0,1]^2 \rightarrow [0,1]$, where Φ is the univariate standard normal and Ψ_ρ the bivariate standard normal distribution function with correlation coefficient ρ. Let $c(U_X, U_\varepsilon) = \partial^2 C/(\partial X \partial \varepsilon)$ be the derivative of $C(U_X, U_\varepsilon)$, then the *joint density* of X and ε can be derived as

$$f(X,\varepsilon) = \mathrm{c}(U_X, U_\varepsilon)h(X)g(\varepsilon) = \mathrm{c}\big(H(X), G(\varepsilon)\big)h(X)g(\varepsilon).$$

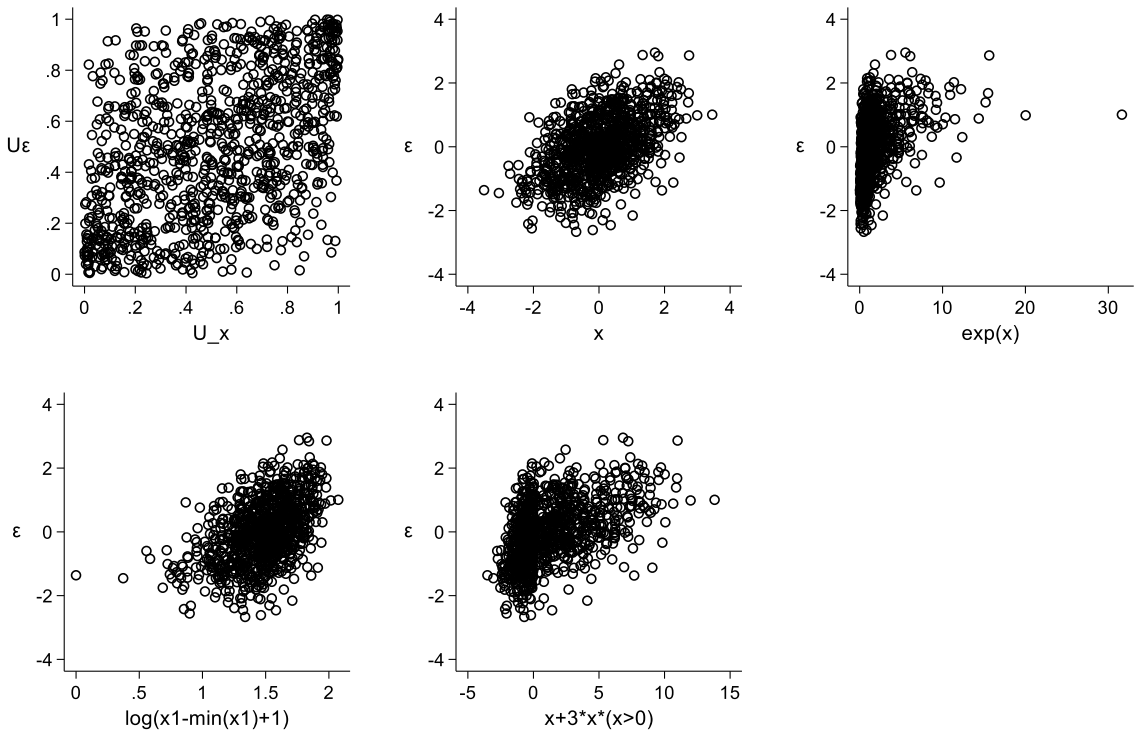
These equations show that the correlation ρ captures dependence among the elements $f(X,\varepsilon)$, but it is not equal to the correlation between these elements.

To see this, it is useful to recall the property of copulas that they are invariant under strictly monotonic increasing transformations (Nelsen, 2007, p. 25). This property tells us that for any pair of continuous random variables V and Z that can be described by a copula C_{VZ}, and any pair of strictly monotone increasing functions $f_V: \mathbb{R} \rightarrow \mathbb{R}$, $f_Z: \mathbb{R} \rightarrow \mathbb{R}$, the equality $C_{f_V(V) f_Z(Z)} = C_{VZ}$ holds, i.e. the dependency structure between V and Z can be described by the same copula as the dependency structure between $f_V(V)$ and $f_V(Z)$. Strictly increasing transformations can however also be non-linear transformation, and as such, they do not necessarily preserve the (linear) Pearson correlation. In fact, the Pearson correlation may be different for different parts of the support of these functions, which in turn implies that the fact that two variables are being described by a Gaussian copula does not necessarily imply that their correlation is constant.

Figure 2 reports scatterplots of the pairs (U_{Xt}, $U_{\varepsilon t}$) from the Gaussian copula (ρ=0.5) reported in Park and Gupta (2012) and the resulting pairs of (X_t, ε_t) from the joint density f generated via $X_t = \Phi^{-1}(U_{Xt})$and $\varepsilon_t = \Phi^{-1}(U_{\varepsilon t})$, as well as some exemplary monotonic transformations of X_t. Although the first two scatterplots show that the Gaussian copula imposes a constant correlation between the marginal distribution functions as well as the joint

distribution for Gaussian marginals, it is also evident that constant correlation does not need to hold for arbitrary subsets of the joint distribution.

Figure 2: Scatterplots of Marginal and Joint Distributions



The dependency structure between the regressor and the error term is unobservable and thus testing this assumption is impossible. Yet, Eckert and Hohberger (2022) show that under certain assumptions, the prevalent endogeneity types (measurement error, omitted variable and reverse causality) can naturally lead to the Gaussian Copula dependency structure. More specifically, the authors show that for all sources of endogeneity, the endogenous regressor and the error term can be written as a linear combination of independent random variables (see their Web Appendix B). If these random variables are all normal distributed, it follows that the Gaussian Copula describes the dependency structure between the endogenous regressor and the error term and any strictly increasing monotonic transformations of these variables.

3.2 Threshold and the Impact on the Dependency Structure

Thresholds are shifts in economic relationships that lead to regression functions that reveal a jump or a kink at a threshold point (Yang et al., 2021). Thresholds can be frequently observed in a marketing context (Biyalogorsky et al., 2001; Han et al., 2001; Jia et al., 2018; Pauwels et al., 2007; Terui & Dahana, 2006; Vastani & Monroe, 2019). Researchers usually discuss them in the context of changes in slopes, such as changes in price sensitivity above versus below a reference price (Han et al., 2001) or changes in advertising responsiveness (Bemmaor, 1984). Thresholds can however also apply to endogeneity levels as the discussion of measurement error in the introduction revealed. Similarly, the omission of feature and display in market response models may lead to endogeneity thresholds as these two marketing mix elements are usually used when the price is low: Although the price is correlated with the unobserved display and feature when considering the whole range of prices, it is uncorrelated when focusing only on low prices or when focusing only on high prices as in these cases feature and display are (if dummy coded) constant 1 or constant 0, respectively.

Thresholds in the correlation patterns between error and endogenous regressor can also be due to asymmetric effects of the marketing mix on performance like sales or market shares, which are frequently found in marketing practice. Unobserved changes in (discrete) market conditions, as well as (unobserved discrete) reference levels of advertising (or price levels), can become omitted variables that lead to a threshold pattern in the relationship between already endogenous regressors and error terms (for example, Gijsenberg, 2017; Pauwels et al., 2007; Sotgiu & Gielens, 2015).

It is important to note that the derivation by Eckert and Hohberger (2022) mentioned in the previous section does not cover situations in which the data exhibits threshold

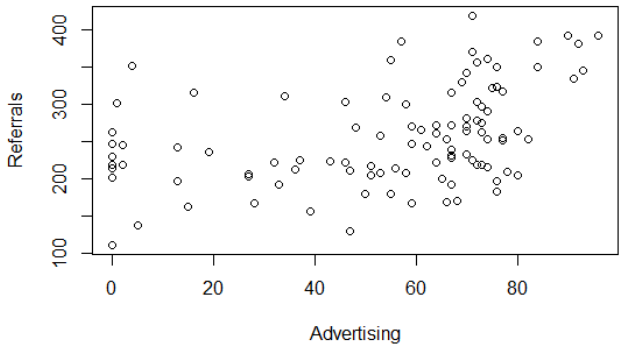
relationships, as such data patterns cannot always be generated by strictly increasing monotonic transformations of underlying normal variables. As such, the GC assumption regarding the dependency structure between endogenous regressor and the error term is less likely to be met and biased estimates may result if the GC approach is nonetheless applied.

3.3 Empirical example: Advertising and Referrals

To illustrate the consequences of thresholds in the joint distribution of endogenous regressor and error we draw on an advertising study by Tellis, Chandy, and Thaivanich (2000). This study examines advertising spending and referral data from a television advertising firm that provides a service for consumers seeking a medical service (for more details please see Tellis et al., 2000). A referral service can optimally allocate advertising within a certain time interval, and hence, based on past referral levels, firms can decide to adjust advertising schedules to optimize responses (e.g. Calli, Weverbergh, & Franses, 2012). When this happens at a higher frequency than the week, like days, then the weekly advertising variable in a model for referrals can become endogenous. Moreover, the degree of endogeneity may change when the firm considers different advertising strategies based on different thresholds in referral levels.

In our analysis, we focus on 104 weeks of observations for the Miami area aggregated at a weekly level (Tellis & Franses, 2006). Figure 3 depicts the number of advertisements and referrals in a scatterplot. This scatterplot clearly shows that sales are rather unaffected by sales for advertising levels below around 70, but that above this threshold advertising has a positive impact on sales. The scatterplot however also suggests that the base level of referrals changes for levels of advertising below and above this threshold.

Figure 3: Scatterplot Advertising vs Referrals (Miami)



In **Error! Reference source not found.** we report the simple regression results obtained with OLS and GC estimation, respectively. Before applying the GC approach, we tested for normality of the endogenous regressor (advertising) using the Shapiro-Wilk test. In all cases, normality was strongly rejected. The results for the OLS estimation show a significant and positive effect of advertising on referrals, while the GC estimates are negative, albeit insignificant. This GC estimate contradicts marketing theory as well as the results from the visual inspection of the data.

Table 1: Regression Advertising and Referrals (Miami area)

	OLS		GC	
	β	S.E.	β	S.E.
Intercept	200.802	13.664	291.281	38.821
GC term			50.649	19.797
Advertising	1.033	0.225	-0.687	0.717

Note: Standard Errors in GC are bootstrapped

3.4 Simulations

To provide a more general examination we now simulate the performance of the GC approach in three different settings. In the first scenario, we assume that the dependency structure changes over the support of the endogenous regressor. In the second scenario, we change the constant of the regression depending on the value of the regressor – this reflects the omitted variable example discussed above. Finally, we also discuss a third scenario, in

which the slope of the regression depends on the value of the regressor. As this third scenario is merely a violation of the linearity in parameters assumption of OLS and does not affect the correlation between the regressor and the error term, we expect the GC approach to work well in this last scenario.

Change in dependency structure: In the first scenario, we violate the GC assumption that the dependency structure between the endogenous regressor *x* and error term *ε* can be characterized by *one* Gaussian Copula. Instead, we assume that for the lower half of the regressor the dependency structure is captured by a Gaussian Copula with a lower correlation than for the upper half. To do so, we draw (*T/2*) tuples

$$\begin{pmatrix} \varepsilon_{t,L}^* \\ x_{t,L}^* \end{pmatrix} \sim N \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$$

and (*T/2*) tuples

$$\begin{pmatrix} \varepsilon_{t,H}^* \\ x_{t,H}^* \end{pmatrix} \sim N \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

Let $\varepsilon_L\ , \varepsilon_H^*, x_L^*, x_H^*$ denote the respective (T/2) row vectors. We then apply the following transformations and vector concatenations to obtain the Tx1 vectors of error terms and regressors: $\varepsilon = \left[\varepsilon_L\ , \varepsilon_H^* + max\left(\varepsilon_L\ \right)\right]$ and $x = [\Phi(x_L^*), \Phi(x_H^*) + max\ (\Phi(x_L^*))]$ where Φ denotes the univariate standard normal distribution function. We then standardize the error and the regressor to have a mean of 0 and a standard deviation of 1. Finally, we generate the dependent variable *y* according to the following equation: $y = 1 + x + \varepsilon$

Change in regression constant for threshold in x: In the second scenario, we follow the GC assumption that the regressor and error term are jointly distributed according to one Gaussian Copula. We allow for a break in the constant of the regression that occurs at the mean of the regressor. More specifically, we draw *T* tuples

$$\begin{pmatrix} \varepsilon_t \\ x_t^* \end{pmatrix} \sim N \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

define ε and x^* as the respective Tx1 vectors, and obtain the regressor *x* via $x = \Phi(x^*)$. We again standardize the error and the regressor and then generate the dependent variable *y* according to the following equation:

$$y = \begin{cases} 0 + x + \varepsilon, for\ x \leq mean(x) \\ 1 + x + \varepsilon, for\ x > mean(x) \end{cases}$$

Change in regression slope for threshold in x: In the third scenario, we again follow the GC assumption that the regressor and the error term are characterized by a Gaussian Copula, but assume a correlation of zero between them. We allow for a break in the slope of the regression that occurs at the mean of the regressor. Again, we draw *T* tuples

$$\begin{pmatrix} \varepsilon_t \\ x_t^* \end{pmatrix} \sim N \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

define ε and x^* as the respective Tx1 vectors, and obtain the regressor *x* via $x = \Phi(x^*)$. We standardize the error and the regressor and then generate the dependent variable *y* according to the following equation:

$$y = \begin{cases} 1 + \varepsilon, for\ x \leq mean(x) \\ 1 + x + \varepsilon, for\ x > mean(x) \end{cases}.$$

Our simulations use a sample size of T=500. We generate 500 data sets per scenario as replicates, and OLS and GC are used on each data set to obtain the empirical sampling distribution of the parameter estimates. We use 500 bootstraps for each data set to obtain the GC standard errors.

Figure 4: Visual inspection of simulated data

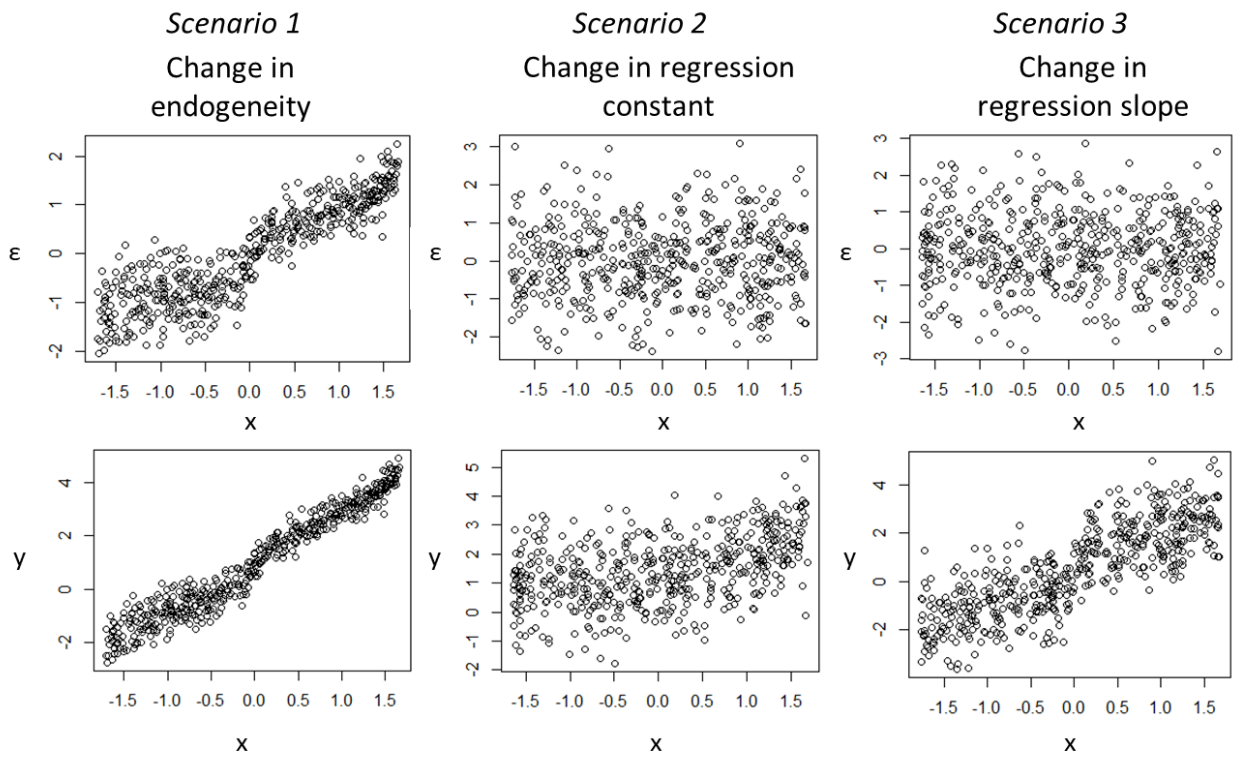


Figure 4 illustrates the distribution of x, ε, and y for a typical data set for each of these scenarios, and the estimation results are summarized in **Error! Reference source not found.**2. To assess the bias of the estimates we calculate t-bias, which is the absolute deviations of the slope estimates from the true parameter value expressed relative to their standard errors (see also Park & Gupta, 2012).

The t-bias is a measurement of the Type 1 error and an unbiased estimator will have a t-bias that is below two. The results show that – as expected – OLS results are biased in all three scenarios. While in scenario 1 the OLS is misspecified due to endogeneity, in scenarios 2 and 3 OLS is misspecified as it does not account for changes in the constant or the regression slope, respectively. The GC approach is also highly biased – again this is to be expected: The GC approach does not account for changes in constant and slope in scenarios 2 and 3, and although it does account for endogeneity, the specification of the dependency structure between the endogenous regressor and the error term in scenario 1 violates its fourth assumption. Particularly worrying is the fact that the median estimate in scenarios 1 and 2 is

even farther away from the true parameter for GC than for OLS. And for scenario 3, OLS reports a slope closer to the average of the two underlying true slopes than GC.

Table 2: Simulation results

	Scenario 1 Change in endogeneity		*Scenario 2* Change in regression constant		*Scenario 3* Change in regression slope	
	OLS	GC	OLS	GC	OLS	GC
Median β	1.888	2.061	1.433	1.969	0.503	0.466
SE of β	0.021	0.111	0.046	0.223	0.046	0.233
t_{bias}	42.019	9.595	9.370	4.323	-	-

Note: SE of β is based on bootstrapped standard errors. We do not report t_{bias} in scenario 3 as it relies on two parameters.

4. Overcoming limitations of the GC Approach

4.1 Threshold Detection

The examples and simulations discussed above all have in common, that the data exhibits a form of threshold in the data. Fortunately, econometric literature indicates that thresholds can be detected simply by using estimation methods based on Ordinary Least Squares, *even when a regressor is potentially endogenous* (Boldea, Hall, & Han, 2012; Hall, Han, & Boldea, 2012; Perron & Yamamoto, 2014, 2015). The OLS-based estimation methods thereby typically involve a search over different partitions of the data, and the use of test statistics based on F or t-tests. Due to the search, these tests have no standard distributions, and these have been derived in the relevant literature (Andrews, 1993; Andrews & Ploberger, 1994). Tables with critical values are also available and packages like Eviews or R provide these per case.

OLS to detect thresholds is considered better than methods based on instrumental variables (IV) in terms of the efficiency of the estimators and the power of the tests (Perron & Yamamoto, 2014, 2015). Additionally, OLS does not suffer from weak identification

issues that can involve IV (Perron & Yamamoto, 2014, 2015). As such, we suggest first determining thresholds in the error-regressor distribution via OLS-based threshold detection, before then applying the GC approach to the resulting parts of the data.

4.2 Simulations

In this section, we return to the simulation scenarios discussed in section 3.4, but this time split the data by threshold first and then apply the methods on the resulting subsamples of the data. We report the results for two different types of thresholds. First, we implement the algorithm described in Bai and Perron (2003) using the endogenous regressor as the breakpoint variable. This approach is for example implemented in the R package *strucchange.* Secondly, we report results when splitting the data by the true threshold as per simulation specification. While the first approach is agnostic and does not require any prior knowledge on behalf of the researcher, the second condition reflects situations in which researchers do have some prior insights into the data generating process.

Table 3: Simulation Results Threshold Application

	Scenario 1 Change in endogeneity		*Scenario 2* Change in regression constant		*Scenario 3* Change in regression slope	
	OLS	GC	OLS	GC	OLS	GC
Panel A: Threshold detection via Bai and Perron (2003)						
A1: Above estimated threshold						
Median β	1.830	1.104	1.012	0.892	0.915	0.849
SE of β	0.042	0.230	0.126	0.641	0.122	0.619
t_{bias}	19.588	0.743	0.732	0.662	1.014	0.521
A2: Below estimated threshold						
Median β	1.408	1.101	0.995	1.102	0.098	0.059
SE of β	0.062	0.292	0.127	0.618	0.130	0.636
t_{bias}	6.585	0.704	0.736	0.636	1.121	0.620
Panel B: Threshold based on true threshold						
B1: Above true threshold						
Median β	1.848	1.036	1.002	1.034	1.002	0.995
SE of β	0.043	0.226	0.127	0.631	0.127	0.609
t_{bias}	19.857	0.659	0.673	0.716	0.612	0.670
B2: Below true threshold						
Median β	1.416	1.043	0.999	1.029	0.011	-0.028
SE of β	0.045	0.209	0.127	0.627	0.127	0.623
t_{bias}	9.405	0.979	0.712	0.653	0.668	0.687

Note: SE of β is based on bootstrapped standard errors.

Error! Reference source not found. reports the results of the simulations. For scenario 1 – change in dependency structure between error and regressor over the support of the regressor – we see that the GC approach performs particularly well when the threshold is known and provides unbiased parameters ($t_{bias}<2$) for the sample above and below the threshold (see panels B1 and B2, respectively). However, even when the threshold is unknown apriori and has to be estimated, the GC approach yields unbiased results close to the true values. This is in contrast to OLS, which continues to provide biased results for the subsamples due to the endogeneity still present in the subsamples.

For scenarios 2 and 3, we also see that applying threshold detection methods or using known information about the threshold improves the performance of the GC approach. All panels reveal that the GC approach provides unbiased results ($t_{bias}<2$), but we do notice that OLS estimates are closer to the true values and that standard errors of OLS are much smaller. This result is to be expected as the OLS estimator is the BLUE estimator for these subsamples of the data as all OLS assumptions (including no correlation between error and regressor) are met for these subsamples. It also reflects the findings by Becker et al (2021) that document that the GC approach is not as unbiased as initially hoped for when including a constant in the estimation and simulation.

Summarizing our simulations in this section show that the GC approach provides unbiased results even for changes in correlation between the error term and regressor once thresholds in the data are accounted for.

4.3 Empirical Examples Accounting for Thresholds in Advertising and Referrals

Returning to the advertising example of section 3.3 we now compare the initial GC estimation with the proposed approach incorporating the threshold detection by Bai and Perron (2003). The results in table 4 show that in contrast to the results in section 3.2 the GC approach incorporating the threshold provides face valid results. While for values below a threshold of 71, neither OLS nor GC shows a significant impact of advertising on referrals, for values larger than 71, both approaches find a positive and significant impact of advertising on referrals, with the GC approach however detecting a much higher slope. Explanations for this threshold effect may hereby be that advertisers possibly switched to another advertising strategy above the threshold, or that word of mouth may have only picked up at a higher advertising level, resulting in an omitted variable that dampens estimated advertising effectiveness in regular OLS (Papies et al., 2016).

Table 4: Empirical Data - Accounting for Threshold

	OLS		GC	
	β	S.E.	β	S.E.
Panel A: Below threshold				
Intercept	217.975	13.457	251.363	28.236
Advertising	0.432	0.275	-0.418	0.703
GC term			23.393	16.592
Panel B: Above threshold				
Intercept	18.393	117.587	-611.13	330.774
Advertising	3.585	1.508	11.831	4.331
GC term			-69.659	36.263

Note: Estimated threshold using Bai and Perron (2003) at advertising=71.

Discussion and Conclusion

The objective of the present paper was to explore the GC approach in the context of thresholds. Although the GC approach to correct for endogeneity received a very positive reception in marketing over the last years and has gained rapidly in popularity, recent work painted a more sobering picture of the approach (Becker et al., 2021; Eckert & Hohberger, 2022). In this paper, we focus on the assumption regarding the Gaussian copula dependency structure between the error term and the regressor and relate it to the long-standing discussion of thresholds in marketing (Pauwels et al., 2007; Vastani & Monroe, 2019). We show that the assumption regarding the dependency structure is less likely to hold in the presence of thresholds. Both our simulation studies and our empirical analyses show that the GC method fails when thresholds in the data are ignored. However, we show that when thresholds are detected and integrated into the model, the GC method works well. We thus recommend to always test for thresholds before applying the GC methodology. Although our discussion and examples are marketing specific, the suggested approach is not limited to this context. The GC approach has received increasing attention in related fields which also frequently face

thresholds situations, e.g. innovation, management, finance and economics and management (Castellacci & Archibugi, 2008; Gupta, Crilly, & Greckhamer, 2020; Heeley & Jacobson, 2008; Law & Singh, 2014)

It is important to note that our paper deals with a specific violation of the GC approach's assumption regarding the dependency structure between endogenous regressor and error term. As such, even if researchers account for thresholds in their data, this does not necessarily imply that the assumption is automatically met. Although our research provides one way to increase the likelihood of this assumption being met, we echo the conclusions by Eckert and Hohberger (2021) who suggest always using alternative approaches such as IV regressions or other IV free methods to ensure the robustness of findings. We also warn readers that threshold detection itself is not without its caveats. There is no single best method to detect thresholds. Various approaches exist, and when there are various methods, the application of these may result in different empirical results (Yu & Phillips, 2018).
We conclude with the observation that after the introduction of a new method it is not uncommon that research emerges that cautions a more prudent application (Becker et al., 2021; Eckert & Hohberger, 2022). For example, this was the case for instrumental regression and the use of appropriate instrumental variables (Murray, 2006; Rossi, 2014). Yet, these more critical pictures should not lead researchers to abandon the new method altogether. Instead, they should carefully check the assumptions and explore the boundaries of its performance. With this paper, we aimed at exactly this. Our results thereby show that the GC method can be useful in practice when it is properly applied and obvious violations of its assumptions remedied.

References

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, 821-856.

Andrews, D. W., & Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica: Journal of the Econometric Society*, 1383-1414.

Bai, J., & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics, 18*(1), 1-22.

Becerra, M., & Markarian, G. (2021). Why Are Firms with Lower Performance More Volatile and Unpredictable? A Vulnerability Explanation of the Bowman Paradox. *Organization science, 32*(5), 1149-1390.

Becker, J.-M., Proksch, D., & Ringle, C. M. (2021). Revisiting Gaussian copulas to handle endogenous regressors. *Journal of the Academy of Marketing Science, forthcoming*.

Bemmaor, A. C. (1984). Testing alternative econometric models on the existence of advertising threshold effect. *Journal of Marketing Research, 21*(3), 298-308.

Biyalogorsky, E., Gerstner, E., & Libai, B. (2001). Customer referral management: Optimal reward programs. *Marketing Science, 20*(1), 82-95.

Boldea, O., Hall, A. R., & Han, S. (2012). Asymptotic distribution theory for break point estimators in models estimated via 2SLS. *Econometric Reviews, 31*(1), 1-33.

Calli, M. K., Weverbergh, M., & Franses, P. H. (2012). The effectiveness of high-frequency direct-response commercials. *International Journal of Research in Marketing, 29*(1), 98-109.

Castellacci, F., & Archibugi, D. (2008). The technology clubs: The distribution of knowledge across nations. *Research Policy, 37*(10), 1659-1673.

Chu, Y., Tian, X., & Wang, W. (2018). Corporate innovation along the supply chain. *Management science*.

Danaher, P. J., & Smith, M. S. (2011). Modeling multivariate distributions using copulas: applications in marketing. *Marketing Science, 30*(1), 4-21.

Datta, H., Ailawadi, K. L., & van Heerde, H. J. (2017). How Well Does Consumer-Based Brand Equity Align with Sales-Based Brand Equity and Marketing-Mix Response? *Journal of Marketing, 81*(3), 1-20.

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. (2022). Addressing Endogeneity without Instrumental Variables: An Exploration of the Gaussian Copula Approach for Management Research. *Journal of Management, forthcoming*.

Falkenström, F., Park, S., & McIntosh, C. N. (2021). Using copulas to enable causal inference from non-experimental data: Tutorial and simulation studies. *Psychological methods, Advance online publication.* https://doi.org/10.1037/met0000414.

Feinberg, F. M., Salisbury, L. C., & Ying, Y. (2016). When random assignment is not enough: Accounting for item selectivity in experimental research. *Marketing Science, 35*(6), 976-994.

Gijsenberg, M. J. (2017). Riding the waves: revealing the impact of intrayear category demand cycles on advertising and pricing effectiveness. *Journal of Marketing Research, 54*(2), 171-186.

Guitart, I. A., & Stremersch, S. (2021). The impact of informational and emotional television ad content on online search and sales. *Journal of Marketing Research, 58*(2), 299-320.

Gupta, K., Crilly, D., & Greckhamer, T. (2020). Stakeholder engagement strategies, national institutions, and firm performance: A configurational perspective. *Strategic Management Journal, 41*(10), 1869-1900.
Hall, A. R., Han, S., & Boldea, O. (2012). Inference regarding multiple structural changes in linear models with endogenous regressors. *Journal of econometrics, 170*(2), 281-302.
Han, S., Gupta, S., & Lehmann, D. R. (2001). Consumer price sensitivity and price thresholds. *Journal of Retailing, 77*(4), 435-456.
Hansen, B. E. (2000). Sample splitting and threshold estimation. *econometrica, 68*(3), 575-603.
Haschka, R. E., & Herwartz, H. (2020). Innovation efficiency in European high-tech industries: Evidence from a Bayesian stochastic frontier approach. *Research Policy, 49*(8), 104054.
Heeley, M. B., & Jacobson, R. (2008). The recency of technological inputs and financial performance. *Strategic Management Journal, 29*(7), 723-744.
Hult, G. T. M., Hair Jr, J. F., Proksch, D., Sarstedt, M., Pinkwart, A., & Ringle, C. M. (2018). Addressing endogeneity in international marketing applications of partial least squares structural equation modeling. *Journal of International Marketing, 26*(3), 1-21.
Jia, H., Yang, S., Lu, X., & Park, C. W. (2018). Do consumers always spend more when coupon face value is larger? The inverted U-shaped effect of coupon face value on consumer spending level. *Journal of Marketing, 82*(4), 70-85.
Kleibergen, F., & Zivot, E. (2003). Bayesian and classical approaches to instrumental variable regression. *Journal of econometrics, 114*(1), 29-72.
Law, S. H., & Singh, N. (2014). Does too much finance harm economic growth? *Journal of Banking & Finance, 41*, 36-44.
Lenz, I., Wetzel, H. A., & Hammerschmidt, M. (2017). Can doing good lead to doing poorly? Firm value implications of CSR in the face of CSI. *Journal of the Academy of Marketing Science, 45*(5), 677-697.
Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics, 30*(1), 67-80.
Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of economic literature, 57*(4), 835-903.
Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *The journal of economic perspectives, 20*(4), 111-132.
Nelsen, R. B. (2007). *An introduction to copulas*: Springer Science & Business Media.
Papies, D., Ebbes, P., & Van Heerde, H. J. (2016). Adressing Endogeneity in Marketing Models. In P. S. H. Leeflang, J. E. Wieringa, T. H. A. Bijmolt, & K. H. Pauwels (Eds.), *Advanced Methods for Modeling Markets*: Springer.
Park, S., & Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science, 31*(4), 567-586.
Pauwels, K., Srinivasan, S., & Franses, P. H. (2007). When do price thresholds matter in retail categories? *Marketing Science, 26*(1), 83-100.
Perron, P., & Yamamoto, Y. (2014). A note on estimating and testing for multiple structural changes in models with endogenous regressors via 2SLS. *Econometric Theory*, 491-507.
Perron, P., & Yamamoto, Y. (2015). Using OLS to estimate and test for structural changes in models with endogenous regressors. *Journal of Applied Econometrics, 30*(1), 119-144.

Reck, F., Fliaster, A., & Kolloch, M. (2021). How to Build a Network that Facilitates Firm-level Innovation: An Integration of Structural and Managerial Perspectives. *Journal of Management studies, Advanced online* https://doi.org/10.1111/joms.12768
Rigobon, R. (2003). Identification through heteroskedasticity. *Review of Economics and statistics, 85*(4), 777-792.
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.
Sarstedt, M., Ringle, C. M., Cheah, J.-H., Ting, H., Moisescu, O. I., & Radomir, L. (2020). Structural model robustness checks in PLS-SEM. *Tourism Economics, 26*(4), 531-554.
Sotgiu, F., & Gielens, K. (2015). Suppliers caught in supermarket price wars: victims or victors? Insights from a Dutch price war. *Journal of Marketing Research, 52*(6), 784-800.
Tellis, G. J., Chandy, R. K., & Thaivanich, P. (2000). Which ad works, when, where, and how often? Modeling the effects of direct television advertising. *Journal of Marketing Research, 37*(1), 32-46.
Tellis, G. J., & Franses, P. H. (2006). Optimal data interval for estimating advertising response. *Marketing Science, 25*(3), 217-229.
Terui, N., & Dahana, W. D. (2006). Research note—estimating heterogeneous price thresholds. *Marketing Science, 25*(4), 384-391.
Vastani, S. F., & Monroe, K. B. (2019). Role of customer attributes on absolute price thresholds. *Journal of Services Marketing*.
Vomberg, A., Homburg, C., & Gwinner, O. (2020). Tolerating and managing failure: An organizational perspective on customer reacquisition management. *Journal of Marketing, 84*(5), 117-136.
Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Mass.: MIT Press.
Yang, L., Zhang, C., Lee, C., & Chen, I.-P. (2021). Panel kink threshold regression model with a covariate-dependent threshold. *The Econometrics Journal, 24*(3), 462-481.
Yu, P., & Phillips, P. C. (2018). Threshold regression with endogeneity. *Journal of econometrics, 203*(1), 50-68.
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.