Essays of Instrument-free Causal Inference, Machine Learning and Marketing Applications

by

Fan Yang

B.S., Sun Yat-sen University, 2013

Mphil, Hong Kong University of Science and Technology, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Business Administration - Marketing)

The University of British Columbia
(Vancouver)

July 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Essays of Instrument-free Causal Inference, Machine Learning and Marketing Applications

submitted by **Fan Yang** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Business Administration - Marketing**.

Examining Committee:

Professor Yi Qian, Sauder School of Business, The University of British Columbia
Supervisor

Professor Joey Hoegg, Sauder School of Business, The University of British Columbia
Supervisory Committee Member

Professor Ralph Winter, Sauder School of Business, The University of British Columbia
Supervisory Committee Member

Professor Hui Xie, Health Sciences, Simon Fraser University
Supervisory Committee Member

Professor Charles Weinberg, Sauder School of Business, The University of British Columbia
University Examiner

Professor Kevin Song, Vancouver School of Economics, The University of British Columbia
University Examiner

Abstract

A prominent challenge when drawing causal inference using observational data is the ubiquitous presence of endogenous regressors. This dissertation investigates causal inference and endogeneity correction, both in methodology development and empirical analysis.

The first essay (in Chapter 2) develops a new instrument-free method using copula to address the endogeneity problem. The classical econometric method to handle regressor endogeneity requires instrumental variables that must satisfy the stringent condition of exclusion restriction. We use the statistical tool copula to directly model the dependence among the regressors and the error term, and abstract information from existing regressors as a generated regressor added to the outcome regression. Our proposed 2sCOPE method extends the existing copula method to a more general setting by allowing (nearly) normal endogenous regressors and correlated exogenous regressors, and is straightforward to use and broadly applicable. We theoretically prove the consistency and efficiency of 2sCOPE, and demonstrate the performance of 2sCOPE via simulation studies and an empirical application.

The second essay (in Chapter 3) further studies the causal inference and endogeneity correction methods for high-dimensional data. The more and more common high-dimensional data in the current big data era make the classical causal inference methods suffer finite-sample bias, inefficiency, or even fail to work when the dimension is larger than the sample

size. In this essay, we extend the 2sCOPE method developed in Chapter 2 to the high-dimensional setting, and propose a lasso-based 2sCOPE method. We demonstrate the performance of the proposed method via simulation studies and an empirical application.

The third essay (in Chapter 4) empirically studies vertical differentiation in two-sided markets, where network size plays the central role. Vertical differentiation is a common strategy in one-sided markets, but whether it is profitable for two-sided platforms is hard to say because of the network effect. In this essay, I take advantage of a unique data set from a leading ride-hailing platform, and develop a structural simultaneous demand and supply model to quantify network externalities. The result shows that besides the product intrinsic value, network value is crucial in determining the degree of product differentiation in two-sided markets.

Lay Summary

Causal inference and endogeneity correction are central to social science research. For observational data, there are mainly three streams of approaches to correct endogeneity: instrument variable approach, structural model approach, and instrument-free approach. Recently, there is a growing interest in instrument-free methods because of its simplicity that no instruments are needed.

This thesis investigates different endogeneity-correction methods in both methodology development and empirical applications. In Essay 1, I develop a new instrument-free method, called 2sCOPE, which is straightforward to implement and can greatly broaden the applicability of the instrument-free methods for dealing with endogeneity issues in practice. In Essay 2, I further adapt the 2sCOPE method for high-dimensional data by combining it with machine learning techniques. Finally, I empirically study consumer behavior in the growing two-sided platforms, and bring insights and suggestions for two-sided platforms based on the estimation results using classical methods for causal inference in Essay 3.

Preface

I am the primary author of the work presented in this Ph.D. thesis. I was responsible for identifying the research questions, conducting the literature review, collecting and managing data, analyzing the data, modeling and coding the estimation procedures, and preparing the manuscript. Specific contributions for each chapter are described below.

Chapter 1: Introduction

I am the primary author of this chapter, with intellectual contributions from Yi Qian and Hui Xie.

Chapter 2 (Essay 1): Addressing Endogeneity using a Two-stage Copula Generated Regressor Approach

I was primarily responsible for identifying the research question, reviewing the literature, conducting simulation studies, cleaning the data, developing and estimating the model, and preparing the manuscript. Yi and Hui contributed to the model development, manuscript editing, and identifying and positioning the research question.

Chapter 3 (Essay 2): Lasso-based Instrument-free Casual Inference

I was primarily responsible for identifying the research question, reviewing the liter-

ature, conducting simulation studies, organizing the data, developing and estimating the model, and preparing the manuscript. Yi and Hui contributed to the model development, manuscript editing and identifying and positioning the research question.

Chapter 4 (Essay 3): Vertical Differentiation in two-sided markets

I was responsible for identifying the research question, organizing the literature, managing the data, developing and estimating the model, conducting counterfactual analysis, and preparing the manuscript.

Chapter 4: Conclusion

I am the primary author of this chapter, with intellectual contributions from Yi Qian and Hui Xie.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I am deeply grateful to my advisor Dr. Yi Qian, who has taken a tremendous amount of time and efforts in advising me and provided generous support for my doctoral study, and has great expertise, patience and kindness in guiding my research. I am deeply honored to have been her student. I would also like to thank my committee member Dr. Hui Xie for his advice and helpful discussions of my research projects. I would like to express my sincere gratitude to my supervisory committee members Dr. Joey Hoegg, and Dr. Ralph Winter for their constructive and insightful suggestions and advice on my research, for their kindness and support, and for spending their valuable time serving on my committee. I would also like to express my sincere thankfulness to Dr. Charles Weinberg in our division for all his enlightening advice and suggestions on my research.

I extend my gratitude to faculty in the Marketing division and all other faculty members who have been very generous in providing support and guidance. I am also delighted that I had such an incredible team of Ph.D. students with whom I can talk to and spend time with. And I want to thank Ms. Elaine Cho and Ms. Florence Yen for their extraordinary responsiveness and approachability.

Finally, I am deeply grateful to my family. They have always been supporting my pursuit of education all the way unconditionally. This long journey of graduate study would not have been possible without their love and support.

Dedication

This thesis is dedicated to my parents Fuping Yang and Liumei Wang, and to my sisters Liu Yang and Xiu Wang.

Chapter 1

Introduction

Causal inference is central to social science research and business practitioners. People are always interested in causal effect instead of just simple descriptive correlation. Nowadays, the rapidly available observational data and technological innovations in this digital era make causal inference even more important to provide real-world evidence for better decision makings. However, one challenge for empirical researchers to draw valid causal inference from observational data is the presence of regressor endogeneity problem, in which regressors are correlated with the (unobserved) structural error.

In the literature, there are mainly three streams of approaches for endogeneity correction using observational data: instrumental variable (IV) approach, structural model approach, and instrument-free approach. The first approach, the IV approach, is to find a good IV that satisfies two conditions: it is strongly correlated with the endogenous regressor via an explainable and validated relationship (i.e., relevance restriction), yet uncorrelated with the structural error (i.e., exclusion restriction). Once a good IV can be found, it can abstract the exogenous part from the endogenous regressor and draw correct causal inference using the two-stage least square method. But empirical researchers usually face the challenge

of finding good IVs in practice. When using the IV approach to correct endogenous, researchers had better explain the validity of the IV, and show evidence of high relevance with the endogenous regressor.

The second stream of approach, the structural model approach, is to specify the economic structure that generates the observational data, including the endogenous regressors. One example is to build up the simultaneous structural supply-demand model that generates the endogenous marketing-mix variables. Doing so allows researchers not only to recover parameters of interest and make causal inferences, but also to perform counterfactual analysis and provide managerial insights. One concern with this approach is that the performance highly depends on model assumptions of supply side. When using this approach, the more information of the supply side, the better for model identification.

The third stream of approach, the instrument-free approach, has aroused increasing interest recently. This approach addresses the endogeneity problem by using information in the existing variables, and the advantage is that no extra instruments are needed. One example is the copula method proposed by Park and Gupta (2012). They use the statistical tool copula to directly model the dependence between the endogenous regressor and the error term, and construct a generated regressor from the existing regressors to correct endogeneity. Intuitively, they use copula to split the error term into an endogenous part and an exogenous part, and correct endogeneity by directly controlling for the endogenous part.

In this thesis, I investigate different endogeneity-correction methods to draw causal inference. Specifically, I conduct three independent studies, two of which are about methodology development of causal inference methods, and one is an empirical application applying the classical causal inference methods. On the one hand, the increasingly available observational data (with more complicated structure e.g., high-dimension) in this digital big data era bring new challenges for traditional methods to draw causal inference and demand the

development of methodologies. Meanwhile, the interaction among different disciplines and knowledge combination provide researchers with fertile soil (new opportunities) for new creative method development to address economic problems.

In the first essay, I develop a new instrument-free approach using the statistical tool copula. I extend the existing copula methods (Park and Gupta 2012, Haschka 2021) to a more general setting with much weaker assumptions. Specifically, the proposed two-stage copula endogeneity correction (2sCOPE) method allows close-to-normal or even normal endogenous regressors, and can include exogenous regressors that are correlated with the endogenous regressor to the model. Simulation result shows that the existing copula correction methods can yield biased estimates under the setting of endogenous regressors that have insufficient non-normality or are correlated with exogenous regressors. The proposed 2sCOPE can provide unbiased estimates, and can even improve estimation efficiency by up to 50%. Moreover, the 2sCOPE method uses the control function approach by adding a generated regressor derived from existing regressors to control for endogeneity, and thus is straightforward to use and broadly applicable. Overall, 2sCOPE method can greatly increase the ease of and broaden the applicability of instrument-free methods for dealing with regressor endogeneity in practice.

In the second essay, I further adapt the instrument-free method, 2sCOPE method, proposed in the first essay for high-dimensional data by combining the causal inference method with machine learning techniques. Nowadays, with the fast growth of internet and technological innovations, high-dimensional data are available everywhere (e.g., census and survey data, complicated network data in online platforms, text, pictures, videos, genetic data etc.). I propose a lasso-based 2sCOPE method by combining the 2sCOPE method with lasso-based feature selection methods to deal with high-dimensional data for causal inference. The simulation result shows that there is a large bias of 2sCOPE method alone without feature selection in high-dimensional setting, while the proposed lasso-based 2sCOPE method

gains substantial improvement in both estimation accuracy and efficiency (around 50%). I also apply the proposed method to an interesting data application. Specifically, I examine how the government's policy strictness during COVID-19 period affects citizens' happiness. By using the lasso-based 2sCOPE method, I get a robust and more significantly negative effect of policy stringency on people's happiness, compared with the estimate using regular 2sCOPE without feature selection.

On the other hand, the recent technological innovations and fast growth of online and mobile platforms bring new business models that can fundamentally change consumer behavior. These changes bring new opportunities and perspectives in understanding consumers. In the third essay, I empirically apply the traditional causal inference methods to study consumer behavior in the emerging two-sided markets (e.g., Taobao, Amazon, Airbnb, Uber, Lyft, Meituan, DoorDash). Specifically, I use structural model and IV approach to correct the endogeneity of network size in examining vertical differentiation in two-sided markets. Network externalities, the distinct feature in two-sided markets, make the conventional vertical differentiation strategy more complicated. It on the one hand makes firms better off in market expansion, but on the other hand further segments the market and limits the positive network effects, weakening the benefit of market expansion. Thus, understanding and quantifying the economic impact of network externalities in two-sided markets is of great importance. In this essay, I take advantage of a distinct data set from a leading ride-hailing platform in New York City, and develop a simultaneous structural demand and supply model to accommodate both the riders' and drivers' decisions and quantify network externalities. The result shows that both the product intrinsic value and the network value are important in determining the degree of product differentiation for two-sided platforms.

The chapters proceed as follows. In the first essay, we address endogeneity using a two-stage generated regressor approach in Chapter 2. We provide the study overview in Section 2.1, summarize the relevant literature in Section 2.2, develop a method in Section

2.3, conduct simulation studies in Section 2.4 and empirical application in Section 2.5. We further discuss the economic intuition and scope in Section 2.6 and conclude this chapter in Section 2.7. In the second essay, we adapt the proposed method in the first essay to the high-dimensional setting by combining with machine learning methods in Chapter 3. We discuss the study motivation and overall results in Section 3.1 and review related literature in Section 3.2. We develop the method in Section 3.3, and conduct simulation study in Section 3.4 and data application in Section 3.5. We further conclude the chapter in Section 3.6. In the third essay, we investigate vertical differentiation in two-sided markets in Chapter 4. We provide the study overview in Section 4.1, review relevant literature in Section 4.2, and discuss used dataset and preliminary data evidence in Section 4.3. We further develop the structural model in Section 4.4, show the estimation results in Section 4.5, conduct counterfactual analysis in Section 4.6 and conclude the chapter in Section 4.6. Chapter 5 gives a brief conclusion to the thesis.

Chapter 2

Addressing Endogeneity Using a Two-stage Generated Regressor Approach

2.1 Introduction

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

regressor endogeneity issue because both the supply and the demand are influenced by unobserved omitted common shocks.

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

To address the lack of suitable IVs, there has been a growing interest in developing and applying IV-free endogeneity-correction methods. Several instrument-free approaches have been developed, including identification via higher moments (Lewbel 1997, Erickson and Whited 2002), heteroscedasticity (Rigobon 2003, Hogan and Rigobon 2003), and latent instrumental variables (Ebbes et al. 2005). All three IV-free methods decompose the endogenous regressor into an exogenous part and an endogenous part. The assumption of the endogenous regressor containing an exogenous component is akin to the stringent condition of exclusion restriction for IVs, and thus can be difficult to justify.

Park and Gupta (2012) proposed an alternative instrument-free method that uses the copula model (Danaher and Smith, 2011) to capture the regressor-error dependence. [1] Compared with the three IV-free methods above, their copula method does not impose the exo-

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

geneity assumption as it directly models the association between the structural error and the endogenous regressor via copula. Furthermore, the copula method can handle discrete endogenous regressors better than other IV-free methods. These features considerably increase the feasibility of endogeneity correction, as evidenced by the rapidly increasing use of the copula correction method (see examples of recent applications in the next section of the literature review). However, similar to other IV-free methods, the copula correction method also requires the distinctiveness between the distributions of the endogenous regressor and the structural error (Park and Gupta 2012). This means that the endogenous regressor is required to have a non-normal distribution for model identification with the commonly assumed normal structural error distribution. Furthermore, we show that the existing copula correction method implicitly requires all exogenous regressors to be uncorrelated with the linear combination of copula transformations of endogenous regressors (henceforth referred to as copula control function (CCF)) used to control for endogeneity, and may yield significant bias when there are noticeable correlations between the CCF and exogenous regressors.

In practical applications, both requirements of sufficient regressor non-normality and no correlation between CCF and exogenous regressors can be too strong, and pose significant challenges and limitations for applying the copula correction method. We often encounter endogenous regressors or include transformations of endogenous variables as regressors that have close-to-normality distributions. Examples of such regressors in economics and marketing management studies include stock market returns (Sorescu et al., 2017), corporate social responsibility (Eckert and Hohberger, 2022), the organizational intelligent quotient (Mendelson, 2000), and the logarithm of price (see Figure 2.3 in Section 5). Theoretically, the endogenous regressor and the structural error can contain a common set of unobservables that collectively have a normal distribution, which can lead to a close-to-normal distribution of the endogenous regressor. In these situations, even if the model is identified asymptotically, close-to-normality of endogenous regressors can cause substan-

tial estimation bias even in moderate sample size and can require a very large sample size (> 2000) to mitigate the finite-sample bias (Becker et al., 2021). Correlations between the CCF and exogenous regressors are also quite common in practical applications, especially when the exogenous regressors are included to control for observed confounders. Examples of such exogenous control variables abound in marketing and management studies, such as customer-specific variables (age, household size, income, past purchase behaviors, etc.) when estimating the returns of consumer targeting strategies on product sales (Papies et al., 2017) and firms' similarity when estimating the effect of competition on innovation (Aghion et al., 2005). Although regressor normality or insufficient regressor non-normality leads to more severe identification issues, including model non-identifiability and poor finite sample performance (Table 2.1), correlations between CCF and exogenous regressors may occur more frequently than close-to-normality of endogenous regressors. Thus, we consider the two requirements of sufficient regressor non-normality and no correlation between CCF and exogenous regressors as being equally stringent, which call for more general and flexible copula correction methods that relax both requirements.

In this paper, we develop a generalized two-stage copula endogeneity correction method, denoted as 2sCOPE, that relaxes the above two requirements. Similar to the existing copula method (Park and Gupta 2012, denoted as $\text{Copula}_{\text{Origin}}$), 2sCOPE requires neither IVs nor the assumption of exclusion restriction. It corrects endogeneity by adding residuals obtained from regressing the latent copula data for each endogenous regressor on the latent copula data for the exogenous regressors as generated regressors in the structural regression model. To demonstrate the benefits of 2sCOPE, we also consider as a benchmark method, denoted as COPE, which is a direct extension of $\text{Copula}_{\text{Origin}}$ and corrects endogeneity by adding latent copula data themselves as generated regressors. The proposed 2sCOPE method is straightforward to use, and it overcomes the above two key limitations of $\text{Copula}_{\text{Origin}}$ as shown in Table 2.1. $\text{Copula}_{\text{Origin}}$ can be viewed as a special case of the

2sCOPE. Importantly, we prove that the 2sCOPE can identify causal effects under much weaker assumptions than $\text{Copula}_{\text{Origin}}$, as summarized in Table 2.1.

The contributions of this work are three folds. *First*, to the best of our knowledge, this work is the first in the literature to provide formal proofs of the theoretical properties of copula correction methods, along with clearly defined assumptions required for causal effect identification (Table 2.1). These theoretical results are much needed because model identifiability is central to addressing the endogeneity issue, and timely given the rapid adoption of copula correction methods in marketing research and elsewhere. Recent methodological work notes lacking rigorous proofs of required model identification conditions and estimation properties (consistency and efficiency) as one main weakness of existing copula correction methods (Haschka, 2021) [2], and calls for further studies of their theoretical properties (Becker et al., 2021). The theoretical results presented in this work fill in this important knowledge gap, and contribute to a better understanding of the properties of the copula correction methods and guidance of their practical use.

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

[2] Owing to the complex form of the estimation method, Haschka (2021) notes the lack of theoretical proofs of required model identification conditions and estimation consistency as one limitation of the copula correction method developed there, and thus has to rely solely on simulation studies to evaluate its empirical properties.

Features	$\text{Copula}_{\text{Origin}}$	Haschka (2021)	2sCOPE
Nonnormality of Endogenous Regressors[1]	Required	Required	Not Required[2]
No Correlated Exogenous Regressors[3]	Required (implicit)	Not Required	Not Required
Intercept Included[4]	YES	NO[5]	YES
Theoretical Proof	YES	NO	YES
Estimation Method	Control Function & MLE	MLE	Control Function
Structural Model	Linear Regression RCL Slope Endogeneity	LPM-FE	Linear Regression LPM-FE, LPM-RE, LPM-ME RCL, Slope Endogeneity

Table 2.1: A Comparison of Copula Methods

Note: [1]: When required, normality of any endogenous regressor leads to non-identifiable models. Insufficient non-normality of endogenous regressors can also cause poor finite sample performance (finite sample bias and large standard errors) and require extremely large sample size to perform well.
[2]: Non-normality of endogenous regressors is not required as long as at least one correlated exogenous regressor is not normally distributed.
[3]: In our paper, correlated exogenous regressors refer to those exogenous regressors correlated with the CCF (copula control function) used to control for endogeneity.
[4]: Becker et al. (2021) shows the significance of including intercept in marketing applications, and the problem of adding intercept using the copula method $\text{Copula}_{\text{Origin}}$ (Park and Gupta 2012).
[5]: The approach cannot estimate the intercept term, which is removed from the panel model prior to estimation using first-difference or fix-effects transformation (Web Appendix A.8 of Haschka (2021)).
LPM: Linear Panel Model; FE: Fixed Effects for individual-specific intercepts with common slope coefficients; RE: Random Effects; ME: Mixed-Effects (including both fixed-effects and random coefficients); RCL: Random Coefficient Logit

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

The second novel finding of our theoretical investigation is as follows. Although the exogenous regressors being correlated with the CCF require special handling for consistent

[3] Although Haschka (2021) explains why correlated regressors can cause potential bias for $\text{Copula}_{\text{Origin}}$, no condition of when bias can occur is given. Specifically, it is possible that with multiple endogenous regressors, the CCF is uncorrelated with exogenous regressors when pairwise correlations between endogenous and exogenous regressors are non-zeros. Even if there is only one endogenous regressor and CCF reduces to be proportional to the copula transformation of the endogenous regressor, the correlation coefficient is not invariant to nonlinear transformations and thus changes after the copula transformation of the endogenous regressor.

causal-effect estimation, they can be beneficial as well by providing additional information to help relax model identification requirements. They could help address the problem of insufficient regressor non-normality, and sharpen model estimates. Furthermore, we prove that when both COPE and 2sCOPE methods yield consistent estimates, 2sCOPE improves the efficiency (i.e., precision) of the structural model estimation by exploiting the correlations between the endogenous and exogenous regressors. The efficiency gain is substantial and can be up to $\sim$50% in our empirical evaluation, meaning that the sample size can be reduced by $\sim$50% to achieve the same estimation efficiency as compared with the COPE method, which does not exploit the correlations between endogenous and exogenous regressors.

Second, the proposed 2sCOPE method is the first copula-correction method that simultaneously relaxes the non-normality assumption of endogenous regressors and handles correlated endogenous and exogenous regressors (Table 2.1). Except for Haschka (2021), existing copula correction methods do not consider correlated endogenous and exogenous regressors, and are subject to potential bias if this correlation is present. Haschka (2021) generalized Park and Gupta (2012) to fixed-effects linear panel models with correlated regressors by jointly modeling the structural error, endogenous and exogenous regressors using copulas and maximum likelihood estimation (MLE). However, as noted in Haschka (2021), Haschka's approach still requires the non-normality of endogenous regressors. Thus, all existing copula correction methods require sufficient non-normality assumption of endogenous regressors for model identification (Park and Gupta, 2012; Haschka, 2021; Becker et al., 2021; Eckert and Hohberger, 2022); even when the model is identified, insufficient regressor non-normality can cause significant finite sample bias in the sample size of less than 2,000 (Haschka, 2021; Becker et al., 2021; Eckert and Hohberger, 2022). Becker et al. (2021) suggested a minimum absolute skewness of 2 for an endogenous regressor in order to ensure good performance of Gaussian copula correction methods in sample size as small

as 200 (Figure 8 in Becker et al. 2021). These strong requirements can significantly limit the use of copula correction methods in practical applications, despite the rapid adoption of these methods recently. Our proposed 2sCOPE method overcomes these important limitations of existing copula correction methods. First, we prove that the structural model with normally distributed endogenous regressors can be identified using the 2sCOPE method as long as one of the exogenous regressors correlated with endogenous ones is nonnormally distributed, which is considerably more feasible in many practical applications. Second, consistent with the above theoretical result, our evaluation in Section 2.4.3 demonstrates superior finite-sample performance of 2sCOPE and shows that 2sCOPE eliminates or substantially reduces the significant finite sample bias problem due to insufficient regressor non-normality raised in Becker et al. (2021) and Eckert and Hohberger (2022). Overall, the proposed 2sCOPE method can greatly broaden the applicability of the instrument-free methods for dealing with endogeneity issues in practice, as seen in Table 2.1.

Finally, 2sCOPE employs generated regressors to address endogeneity. By including generated regressors in the structural model to control endogeneity, 2sCOPE enjoys several benefits associated with using a control function (Heckman and Robb Jr 1985, Blundell and Powell 2003, Blundell and Powell 2004, Petrin and Train 2010, Blundell and Matzkin 2014, Wooldridge 2015) to address endogeneity, such as incurring little extra computational and modeling burden to address endogeneity, broader applicability with much weaker assumptions, and increased robustness to model misspecifications [4]. The vast majority of applications of the existing copula correction method have used the generated-regressor approach (Becker et al., 2021). We demonstrate that 2sCOPE retains these desirable properties of the control function approach for a range of commonly used models in marketing studies, as shown in Table 2.1, while relaxing the two key assumptions of $\text{Copula}_{\text{Origin}}$: regressor non-normality and uncorrelatedness between CCF and exogenous regressors. In many of these

[4] As shown in Becker et al. (2021), Gaussian copula control function approach is more robust against error term misspecifications than the Gaussian copula MLE approach. Note that when we compare MLE with control function approach, we only consider the estimation method dimension.

models, the MLE approach becomes much more difficult or computationally infeasible, but our 2sCOPE approach is straightforward. Section 2.3.4 presents an example in which extending the MLE approach of Haschka (2021) to random coefficient linear panel models with correlated endogenous and exogenous regressors requires numerically evaluating potentially high-dimensional integrals of complicated functions of the random effect distributions, whereas 2sCOPE eliminates the need to evaluate these high-dimensional integrals and can be implemented using standard software programs for random coefficient linear panel models assuming all regressors are exogenous. Furthermore, the generated-regressor approach facilitates studying the theoretical properties of the proposed 2sCOPE procedure and the comparison of these procedures. In this work, we provide theoretical proofs for the implicit assumption needed to ensure consistency of $\text{Copula}_{\text{Origin}}$, and the consistency and efficiency comparison for the proposed 2sCOPE under correlated regressors and normally distributed regressors.

The remainder of this paper unfolds as follows. Section 2.2 reviews the related literature on methods for causal inference with endogenous regressors. In Section 2.3, we show the implicit assumption of $\text{Copula}_{\text{Origin}}$. We then propose the 2sCOPE method, providing theoretical proofs for the consistency of the proposed 2sCOPE method as well as for efficiency gain and model identifiability with normally distributed regressors under the 2sCOPE method. We also summarize the estimation procedure of the proposed method. In Section 2.4, we evaluate the performance of our proposed 2sCOPE method using simulation studies and compare it with $\text{Copula}_{\text{Origin}}$ and its direct extension COPE under different scenarios. In Section 2.5, e apply the proposed 2sCOPE method to estimate price elasticity using store purchase databases in section 2.5. We discuss the economic intuition and scope of the proposed 2sCOPE method in section 2.6, and conclude the paper in Section 2.7.

2.2 Literature Review

The marketing, economic and statistics literature develops a rich set of methods to draw causal inferences. The gold standard to estimate causal effects is using randomized assignment2 such as controlled lab experiments and field experiments (Johnson et al. 2017, Anderson and Simester 2004, Godes and Mayzlin 2009). When controlled experiments are not feasible, quasi-experimental designs such as regression discontinuity and difference in differences are used to mimic randomized experiments and to enable the identification of causal effects with observational data (Hahn et al. 2001, Hartmann et al. 2011, Narayanan and Kalyanam 2015, Athey and Imbens 2006, Shi et al. 2017). However, these quasi-experimental designs have special data and design requirements, and cannot cope with the general issue of endogenous regressors when estimating causal effects using observational data.

There is a large literature focusing on approaches to addressing endogenous regressors when inferring causal effects. Rutz and Watson (2019), Papies et al. (2017) and Park and Gupta (2012) provided an overview of addressing endogeneity in marketing. Three broad classes of solutions are discussed. The most commonly used solution is to find observed instrumental variables to correct for endogeneity (Kleibergen and Zivot 2003, Qian 2008, Ataman et al. 2010, Van Heerde et al. 2013 and Novak and Stern 2009). Angrist and Krueger (2001) and Rossi (2014) provide a survey of literature that uses the instrumental variables approach. Rossi (2014) surveyed 10 years of publications in *Marketing Science* and *Quantitative Marketing and Economics*, which revealed that the most commonly used instrumental variables are lagged variables, costs, fixed effects and Hausman-style variables from other markets. However, the survey found that the strength of the instruments is rarely measured and reported, which is needed to detect the weak instrument problem. Moreover, one generally cannot test the exclusion restriction condition and verify the validity of instruments.

The survey also found that most papers lack a discussion of why the instruments used are valid. In a word, though the theory of instrumental variables is well-developed, good instruments are difficult to find, making the IV approach hard to implement in practice. Studies that identify good instruments are subsequently highly valued.

The second class of solutions to mitigate endogeneity is to specify the economic structure that generates the observational data including endogenous regressors (e.g., a supply-side model for marketing-mix variables). Doing so allows researchers not only to recover parameters of interest and make causal inferences, but also to perform counterfactual analysis (Chintagunta et al. 2006). Some other examples of this approach in the marketing literature are Berry (1994), Sudhir (2001), Dubé et al. (2002), Yang et al. (2003), Sun (2005), Dotson and Allenby (2010) and Otter et al. (2011). The key concern with this approach is that the performance highly depends on model assumptions of the supply side. Incorrect assumptions or insufficient information on the supply side can lead to biased estimates (Chintagunta et al. 2006, Hartmann et al. 2011)

The third class of solutions in the domain of endogeneity correction is instrument-free methods. This is a more recent stream of methodological development. Three extant instrument-free approaches are discussed in Ebbes et al. (2009): the higher moments (HM) approach (Lewbel 1997, Erickson and Whited 2002), the identification through heteroscedasticity (IH) estimator (Rigobon 2003, Hogan and Rigobon 2003), the latent instrumental variables (LIV) method (Ebbes et al. 2005). Recently Wang and Blei (2019) proposed a deconfounder approach that has some flavor to the LIV approach. All these approaches divide the endogenous regressor P into an endogenous and an exogenous part, $P = f(Z) + v$, where $f(Z)$ is treated as an exogenous random variable with unique structures imposed for model identification in different methods. However, the assumption of $f(Z)$ being exogenous is hard to guarantee. Park and Gupta (2012) introduced another instrument-free method that doesn't require the exogeneity of $f(Z)$. It directly models the

association between the structural error and the endogenous regressor via copula.

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

2.3 Methods

In this section, we develop a copula-based instrument-free method to handle endogenous regressors when there exist exogenous regressors that are correlated with endogenous regressors. We first review the $\text{Copula}_{\text{Origin}}$ method in Park and Gupta (2012). We show that $\text{Copula}_{\text{Origin}}$ implicitly assumes no correlations between the exogenous regressors and the CCF, as well as how the violation of the assumption can cause bias in the structural model parameter estimates for the current copula-based instrument-free method. Then we present a newly proposed method to deal with the problem and the detailed estimation procedure. We also show how exogenous regressors correlated with endogenous regressors can sharpen structural model parameter estimates and permit the identification of the structural model containing normally distributed endogenous regressors, known to cause the model non-identifiability issue for $\text{Copula}_{\text{Origin}}$.

2.3.1 Assumptions in the Existing Copula Endogeneity-Correlation Method ($\text{Copula}_{\text{Origin}}$)

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

$$Y_t = \mu + P_t\alpha + W_t'\beta + \xi_t, \tag{2.1}$$

where $t = 1, 2, ..., T$ indexes either time or cross-sectional units, Y_t is a (1×1) dependent variable, P_t is a (1×1) endogenous regressor, W_t is a $(\text{k}\times1)$ vector of exogenous regressors, ξ_t is the structural error term, and (μ, α, β) are model parameters. P_t is correlated with ξ_t, and this correlation generates the endogeneity problem. W_t is exogenous, which means it is

[5] Unlike Park and Gupta (2012), our model includes the intercept term. As shown in Becker et al. (2021), it is important to include the intercept term when evaluating the copula correction method.

not correlated with ξ_t, but can be correlated with the endogenous variable P_t.

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

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

$$F(P,\xi) = C(H(P),G(\xi)) = C(U_p,U_\xi), \tag{2.2}$$

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

To capture the association between the endogenous regressor P and the error ξ, Park and Gupta (2012) used the following Gaussian copula for its many desirable properties (Danaher

2007; Danaher and Smith 2011):

$$F(P,\xi) = C(U_p, U_\xi) = \Psi_\rho(\Phi^{-1}(U_p), \Phi^{-1}(U_\xi)) \\ = \frac{1}{2\pi(1-\rho^2)^{1/2}} \int_{-\infty}^{\Phi^{-1}(U_p)} \int_{-\infty}^{\Phi^{-1}(U_\xi)} \exp\left[\frac{-(s^2 - 2\rho \cdot s \cdot t + t^2)}{2(1-\rho^2)}\right] dsdt, \tag{2.3}$$

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

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

$$\begin{pmatrix} P_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \tag{2.4}$$

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

$$\xi_t = \sigma_\xi \xi_t^* = \sigma_\xi \rho P_t^* + \sigma_\xi \sqrt{1-\rho^2}\omega_t, \tag{2.5}$$

where the first part $\sigma_\xi \rho P_t^*$ captures the correlation between ξ_t and the endogenous regressor,

and the other part $\sigma_\xi \cdot \sqrt{1-\rho^2}\omega_t$ is an independent new error term. Equation (2.1) can be rewritten as follows:

$$Y_t = \mu + P_t\alpha + W_t\beta + \sigma_\xi \cdot \rho \cdot P_t^* + \sigma_\xi \cdot \sqrt{1-\rho^2} \cdot \omega_t. \tag{2.6}$$

Based on the above representation, Park and Gupta (2012) suggested the following generated regressor approach to correcting for the endogeneity of P_t: the ordinary least square (OLS) estimation of Equation (2.6) with $P_t^* = \Phi^{-1}(U_p)$ included as an additional regressor will yield consistent model estimates. Park and Gupta (2012) also pointed out that in order for the above approach to work, P_t needs to have a non-normal distribution. Suppose P_t is normally distributed, $P_t = P_t^* \cdot \sigma_p$, resulting in perfect collinearity between P_t and P_t^* and violating the full rank assumption required for identifying the linear regression model in Equation (2.6). Thus, P_t should follow a different distribution from the normal error term so that the causal effect of P that is independent of all other regressors can be identified.

However, we show here that an additional and implicit assumption for the above generated-regressor approach to yield consistent model estimates is the uncorrelatedness between P_t^* and W_t. For the OLS estimation to yield consistent estimation, the error term ω_t in Equation (2.6) is required to be uncorrelated with all the regressors on the right-hand side of the equation: P_t, W_t, P_t^*. The theorem below shows that W_t becomes endogenous in Equation (2.6) when W_t and P_t^* are correlated.

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

Proof: See the Appendix, Proof of Theorem 1.

To summarize, the generated regressor procedure based on Equation (2.6) makes the

following set of assumptions.

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

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

Assumption 3. *Nonnormality of the endogenous regressor P_t;*

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

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

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

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

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

In Park and Gupta (2012), all the above assumptions except Assumption (4) have been made explicit. Among the first three assumptions, Park and Gupta (2012) have shown reasonable robustness of their copula method to non-normal distributions of the error term (Assumption 1) and alternative forms of copula functions (Assumption 2), although it is not surprising to observe the sensitivity of $\text{Copula}_{\text{Origin}}$ to gross violations of these assumptions, such as highly skewed error distributions (Becker et al., 2021). By contrast, the assumption that the endogenous regressor P_t follows a non-normal distribution (Assumption 3) is critical. An endogenous regressor following a normal distribution can cause the structural model to be unidentifiable regardless of sample size; a nearly normally distributed endogenous regressor may require a very large sample size for the method to perform well and may cause the method to have poor performance for a finite sample size. Moreover, we have shown above that for their method to work, there should be no exogenous regressors that are correlated with P_t^* (Assumption 4). Both the Assumptions (3 and 4) can be too strong and substantially limit the applicability of the instrument-free copula method in practice.

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

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

follows the multivariate normal distribution:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{p\xi} \\ \rho_{pw} & 1 & 0 \\ \rho_{p\xi} & 0 & 1 \end{bmatrix} \right), \tag{2.7}$$

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

Under the above Gaussian copula model in Equation (3.4), one can develop a direct extension of $\text{Copula}_{\text{Origin}}$, which adds generated regressors P_t^* and W_t^* into the structural regression model to correct for endogeneity bias (Appendix). The resulting method, denoted as COPE, is shown to yield consistent causal effect estimates (Theorem A1 in the Appendix) without requiring **Assumption 4** needed for $\text{Copula}_{\text{Origin}}$. However, COPE requires endogenous regressors P_t and exogenous regressors W_t to be both non-normally distributed (Theorem A1 in the Appendix). To overcome the limitations of COPE, below we derive the 2sCOPE method that relaxes both assumptions and is shown to be more efficient than COPE.

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

$$Y_t = \mu + P_t\alpha + W_t\beta + \xi_t \tag{2.8}$$

$$P_t^* = W_t^*\gamma + \varepsilon_t, \tag{2.9}$$

where the two error terms ε_t and ξ_t are correlated because of the endogeneity of P_t. Under the assumption that ξ_t follows a normal distribution, ε_t and ξ_t follow a bivariate normal distribution, since they are a linear combination of tri-normal variate (ξ_t^*, P_t^*, W_t^*) under the

Gaussian copula assumption.

The main idea of 2sCOPE is to make use of the fact that by conditioning on ε_t, the structural error term ξ_t becomes independent of both P_t and W_t. That is, by conditioning on the component of P_t causing the endogeneity of P_t (i.e, ε_t here), the structural error is not correlated with both P_t and W_t, thereby ensuring the consistency of standard estimation methods. In this sense, ε_t serves as a (scaled) control function to address the endogeneity bias. To demonstrate this point, note that the Gaussian copula model in Equation (3.4) can be rewritten as follows:

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

$$\begin{pmatrix} \omega_{1,t} \\ \omega_{2,t} \\ \omega_{3,t} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right). \tag{2.10}$$

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

$$P_t^* = \rho_{pw} W_t^* + \sqrt{(1-\rho_{pw}^2)} \cdot \omega_{2,t} = \rho_{pw} W_t^* + \varepsilon_t, \tag{2.11}$$

which shows γ in Equation (2.9) is ρ_{pw} and $\varepsilon_t = \sqrt{(1-\rho_{pw}^2)} \cdot \omega_{2,t}$, and

$$
\begin{aligned}
Y_t &= \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* + \frac{-\sigma_\xi \rho_{pw}\rho_{p\xi}}{1-\rho_{pw}^2} W_t^* + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t} \\
&= \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2}(P_t^* - \rho_{pw} W_t^*) + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}, \\
&= \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2}\varepsilon_t + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}. \qquad (2.12)
\end{aligned}
$$

Equation (2.12) suggests adding the estimate of the error term ε_t from the first stage regression as a generated regressor to the outcome regression instead of using P_t^* and W_t^*. As shown in Theorem 2, the new error term $\omega_{3,t}$ is uncorrelated with all the regressors in Equation (2.12), ensuring the consistency of model estimates. This two-step procedure, named 2sCOPE, adds the first-stage residual term $\widehat{\varepsilon}_t$ to control for endogeneity and in this aspect is similar to the control function approach of Petrin and Train (2010). However, unlike Petrin and Train (2010), 2sCOPE requires no use of instrumental variables.

Theorem 2. ***Estimation Consistency.*** *Assuming (1) the error term is normal, (2) the endogenous variable P_t or correlated regressors W_t is nonnormal, and (3) a Gaussian Copula for the error term, P_t and W_t, $Cov(\omega_{3,t}, W_t) = Cov(\omega_{3,t}, P_t) = Cov(\omega_{3,t}, \varepsilon_t) = 0$ in Equation (3.8).*

Proof: See the Appendix, Proof of Theorem 2.

According to Theorem 2, the proposed method 2sCOPE can yield consistent estimates when assumptions are met. Specifically, Assumption 4 is relaxed because 2sCOPE can handle the case when extra exogenous regressors that can be correlated with the endogenous regressor are included in the model. Theorem 3 below further shows that 2sCOPE relaxes Assumption 3 (the non-normality assumption on endogenous regressors), a critical model

identification condition required in all other copula correction methods.

Theorem 3. ***Nonnormality Assumption Relaxed.*** *Assuming (1) the error term is normal, (2) one of the correlated exogenous regressors W_t is nonnormal, and (3) a Gaussian Copula for the error term, P_t and W_t, 2sCOPE estimator $\widehat{\theta_2}$ is consistent when P_t follows a normal distribution while the COPE estimator $\widehat{\theta_1}$ is not consistent.*

Proof: See the Appendix, Proof of Theorem 3.

Theorem 3 shows that as long as one of the exogenous regressors that are correlated with the endogenous regressor P_t is nonnormally distributed, 2sCOPE can correct for endogeneity for a normal regressor P_t while COPE cannot. Intuitively, when one of P_t and W_t is normal, P_t^* (or W_t^*) becomes a linear function of P_t (or W_t) under the Gaussian copula assumption, rendering the second stage model in COPE to fail the full rank assumption and become unidentified. Thus, COPE cannot deal with normal endogenous regressors. For the proposed 2sCOPE method in Equation (2.12), adding the first stage residual $\widehat{\varepsilon_t}$ as the generated regressor helps model identification. As long as not all W_ts are normal, ε_t would not be a linear function of P_t and W_t and thus the second stage model (Equation 2.12) in 2sCOPE would satisfy the full rank requirement for model identification. Thus, our proposed method 2sCOPE can relax the nonnormality assumption on the endogenous regressor required in Park and Gupta (2012) as long as one of W_t is nonnormally distributed.

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

Theorem 4. ***Variance Reduction.*** *Assuming (1) the error term is normal, (2) the endogenous variable P_t and correlated regressors W_t are nonnormal, and (3) a Gaussian Copula for the error term, P_t and W_t, $\mathbf{Var}(\widehat{\theta_2}) \leq \mathbf{Var}(\widehat{\theta_1})$, where $\widehat{\theta}_1$ and $\widehat{\theta}_2$ denote parameter estimates from COPE and 2sCOPE, respectively.*

Proof: See the Appendix, Proof of Theorem 4.

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

2.3.3 Multiple Endogenous Regressors

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

$$Y_t = \mu + P_{1,t} \cdot \alpha_1 + P_{2,t} \cdot \alpha_2 + W_t\beta + \xi_t. \tag{2.13}$$

Under the multivariate Gaussian distribution assumption on $(\xi_t, P^*_{1,t}, P^*_{2,t}, W^*_t)$, the system equations of 2sCOPE method in Equation (2.8, 2.9) are readily extended to the case with two endogenous regressors as

$$Y_t = \mu + P_{1,t}\alpha_1 + P_{2,t}\alpha_2 + W_t\beta + \xi_t, \tag{2.14}$$

$$P^*_{1,t} = \rho_{wp1}W^*_t + \varepsilon_{1,t}, \tag{2.15}$$

$$P^*_{2,t} = \rho_{wp2}W^*_t + \varepsilon_{2,t}, \tag{2.16}$$

where Equations (2.15) and (2.16) can be directly derived from the Gaussian copula assumption; $(\xi_t, \varepsilon_{1,t}.\varepsilon_{2,t})$ are a linear transformation of $(\xi_t, P^*_{1,t}, P^*_{2,t}, W^*_t)$, and thus also follow a multivariate Gaussian distribution. As a result, we can decompose the structural error ξ_t

as additive terms for $\varepsilon_{1,t}$, $\varepsilon_{2,t}$ and a remaining independent error term $\omega_{4,t}$ as follows

$$Y_t = \mu + P_{1,t}\alpha_1 + P_{2,t}\alpha_2 + W_t\beta + \eta_1\varepsilon_{1,t} + \eta_2\varepsilon_{2,t} + \sigma_\xi \cdot m \cdot \omega_{4,t}, \tag{2.17}$$

where $\varepsilon_{1,t} = P^*_{1,t} - \rho_{wp1}W^*_t$ and $\varepsilon_{2,t} = P^*_{2,t} - \rho_{wp2}W^*_t$, m is a constant depending only on the correlation coefficients in the Gaussian copula, and η_1, η_2 and $\omega_{4,t}$ are the same as those defined in Equation (A.6) in the Appendix for describing COPE for multiple endogenous regressors and thus the new (scaled) error term $\omega_{4,t}$ is independent of latent copula data $(P^*_{1,t}, P^*_{2,t}, W^*_t)$ as well as all functions of these latent data including $P_{1,t}, P_{2,t}, W_t, \varepsilon_{1,t}, \varepsilon_{2,t}$. Because $\omega_{4,t}$ is independent of all regressors on the right side of Equation (2.17), the OLS estimation of Equation (2.17) yields a consistent estimation of structural model parameters. Note that Equation (2.17) can also be obtained from Equation (A.7) in Online Appendix A.3 for describing COPE for multiple endogenous regressors by noting that $\varepsilon_{1,t} = P^*_{1,t} - \rho_{wp1}W^*_t$ and $\varepsilon_{2,t} = P^*_{2,t} - \rho_{wp2}W^*_t$. However, 2sCOPE adds only two residual terms $(\varepsilon_{1,t}, \varepsilon_{2,t})$ as generated regressors instead of three copula transformations of regressors $(P^*_{1,t}, P^*_{2,t}, W^*_t)$ as generated regressors, as COPE does (Equation (A.7) in the Appendix). Thus, 2sCOPE adds a smaller number of generated regressors than COPE, and provides higher estimation efficiency. In addition, by adding residual terms as the generated regressors, 2sCOPE relaxes the assumption of regressor non-normality required to COPE as long as not all W_ts are normal. The proof for the estimation consistency of 2sCOPE, estimation efficiency gain and relaxation of the regressor-nonnormality assumption for 2sCOPE can be found in the Appendix under the related Theorems 2, 3, 4.

Table 2.2 summarizes the assumptions for the three methods: our proposed 2sCOPE method, the existing copula method $\text{Copula}_{\text{Origin}}$ and $\text{Copula}_{\text{Origin}}$'s direct extension, COPE. Our proposed 2sCOPE method can deal with the case when there are exogenous regressors that are correlated with the endogenous regressors. Moreover, 2sCOPE can further relax the regressor-nonnormality assumption. Table 2.3 summarizes and compares the estimation

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

Table 2.2: Summary of Assumptions for the Three Methods

procedure of COPE and 2sCOPE.

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

Table 2.3: Estimation Procedure

2.3.4 2sCOPE for Random Coefficient Linear Panel Models with Endogenous Regressors

We consider the following random coefficient model for linear panel data

$$Y_{it}|\mu_i,\alpha_i,\beta_i \quad = \quad \bar{\mu}+\mu_i+P_{it}'\alpha_i+W_{it}'\beta_i+\xi_{it}, \tag{2.18}$$

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

The linear panel data model as specified in Equation (2.18) is general and includes the linear panel model with only individual-specific intercepts considered in Haschka (2021) as a special case. Specifically, Haschka (2021) fixes (α_i,β_i) to be the same value (α,β) across all units, assuming all cross-sectional units have the same slope coefficients. In contrast, the model in Equation (2.18) relaxes this strong assumption and can generate unit-specific slope parameters, which can be used for targeting purposes.

A fully random coefficient model typically assumes that (μ_i,α_i,β_i) follows a multivariate normal distribution. When all regressors are exogenous, estimation algorithms for such random coefficient models are well-established and computationally feasible even for a high-dimensional vector of random effects (μ_i,α_i,β_i): with the normal conditional dis-

tribution for $Y_{it}|(\mu_i, \alpha_i, \beta_i)$ in Equation (2.18) and the multivariate normal prior distribution for random effects $(\mu_i, \alpha_i, \beta_i)$, marginally Y_{it} follows a normal distribution with a closed-form expression containing no integrals of random effects $(\mu_i, \alpha_i, \beta_i)$, leading to an easy-to-evaluate likelihood function (Greene, 2012). For instance, R function `lme()` can be used to obtain MLEs of population model estimates and empirical Bayes estimates of random effects. Alternatively, one can assume a mixed-effect model where μ_i is a fixed effect parameter with μ_i's allowed to be correlated with the regressors P_{it} and W_{it}. To avoid the potential incidental parameter problem associated with these fix-effect parameters, one often uses the first-difference or fixed-effects transformation to eliminate the incidental intercept parameters as follows

$$\tilde{y}_{it}|\alpha_i, \beta_i \quad = \quad \tilde{P}'_{it}\alpha_i + \tilde{W}'_{it}\beta_i + \tilde{\xi}_{it}, \tag{2.19}$$

where $\tilde{y}_{it}$, $\tilde{P}_{it}$, $\tilde{W}_{it}$ and $\tilde{\xi}_{it}$ denote new variables obtained from the first-difference or fixed-effect transformation. Haschka (2021) considered a special case of Equation (2.19) by fixing (α_i, β_i) to be constants.

It is straightforward to apply 2sCOPE to address regressor endogeneity in the general random coefficient model for linear panel data in Equation (2.18) and the transformed one without intercepts in Equation (2.19). [7] Assuming $(P_{it}, W_{it}, \xi_{it})$ follow a Gaussian copula, COPE adds the generated regressor $P^*_{it} = \Phi^{-1}(\widehat{H}(P_{it}))$ and $W^*_{it} = \Phi^{-1}(\widehat{L}(W_{it}))$ into Equation (2.18) to control for regressor endogeneity. The 2sCOPE procedure adds the residuals obtained from regressing P^*_{it} on W^*_{it}. Thus, 2sCOPE method can be implemented using standard software programs for random coefficient linear panel models assuming all regressors are exogenous (see Section 2.4.6 for an illustration using the R function `lme()`). By contrast, the MLE approach for copula correction in the random coefficients model accounting

[7] Similar to Haschka (2021), a GLS transformation can be applied to both sides of Equation (2.19), resulting in a pooled regression for which 2sCOPE can be directly applied.

for correlated endogenous and exogenous regressors is not available yet and would require constructing complicated joint likelihood on the error term, P_t and W_t, which involves newly appearing numerical integrals of random effects and cannot be maximized by standard estimation algorithms for random coefficient models. [8] Finally, current applications applying $\text{Copula}_{\text{Origin}}$ do not consider the role of exogenous regressors. Our analysis shows that this may yield bias if any exogenous regressor is correlated with the CCF added to control endogeneity, for which 2sCOPE should be used to address regressor endogeneity.

2.3.5 2sCOPE for Slope Endogeneity and Random Coefficient Logit Model

In the Appendix, we derive the 2sCOPE method to tackle the slope endogeneity problem and address endogeneity bias in random coefficient logit models with correlated and normally distributed regressors. In these two cases, we show how to apply 2sCOPE to correct for the endogenous bias, which can avoid the potential bias of $\text{Copula}_{\text{Origin}}$ due to the potential correlations between the exogenous regressors and CCF, as well as make use of the correlated exogenous regressors to relax the non-normality assumption of endogenous regressors, improve model identification and sharpen model estimates. As shown there, 2sCOPE can be implemented using standard estimation methods by adding generated regressors to control for endogenous regressors. By contrast, the maximum likelihood approach can require constructing a complicated joint likelihood that is not what the standard estimation method uses and thus requires separate development and significantly more computation involving numerical integration.

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

2.4 Simulation Study

In this section, we conduct Monte Carlo simulation studies for the following goals: (a) to assess the performance of the proposed method for correlated regressors, (b) to assess the performance of the proposed method under regressor normality and near normality, (c) to assess generalizability and restrictions of the distributional assumptions about the endogenous and exogenous regressors, and (d) to compare the performance of the proposed method with existing methods. Following Park and Gupta (2012), we measure the estimation bias using t_{bias} calculated as the ratio of the absolute difference between the mean of the sampling distribution and the true parameter value to the standard error of the parameter estimate. As defined above, t_{bias} represents the size of bias relative to the sampling error. The Appendix provides additional simulation results on the robustness of 2sCOPE to the misspecifications of the structural error distribution and the copula dependence structure.

2.4.1 Case 1: Non-normal Regressors

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

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{p\xi} \\ \rho_{pw} & 1 & 0 \\ \rho_{p\xi} & 0 & 1 \end{bmatrix} \right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right) \tag{2.20}$$

$$\xi_t = G^{-1}(U_{\xi,t}) = G^{-1}(\Phi(\xi_t^*)) = \Phi^{-1}(\Phi(\xi^*)) = 1 \cdot \xi_t^*, \tag{2.21}$$

$$P_t = H^{-1}(U_{P,t}) = H^{-1}(\Phi(P_t^*)), \quad W_t = L^{-1}(U_{W,t}) = L^{-1}(\Phi(W_t^*)), \tag{2.22}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + \xi_t = 1 + 1 \cdot P_t + (-1) \cdot W_t + \xi_t. \tag{2.23}$$

where ξ_t^* and P_t^* are correlated with the correlation coefficient $\rho_{p\xi} = 0.5$, and thus ξ_t and P_t are correlated, generating the endogeneity problem. W_t^* is exogenous and is not correlated with ξ_t^*. But W_t^* and P_t^* are correlated with the correlation coefficient $\rho_{pw} = 0.5$, and thus W_t and P_t are correlated. We consider four different estimation methods: (i) OLS, (ii) $\text{Copula}_{\text{Origin}}$ in the form of Equation (2.6), (iii) the extended method COPE in the form of Equation (A.3), and the proposed method 2sCOPE in the form of Equation (3.8). We set the sample size T = 1000, and generate 1000 data sets as replicates using the DGP above. In the simulation, we use the gamma distribution $Gamma(1,1)$ with shape and rate equal to 1 for P_t and the exponential distribution $Exp(1)$ with rate 1 for W_t. Models are estimated on all generated data sets, providing the empirical distributions of the parameter estimates.

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

Table 2.4: Results of the Simulation Study Case 1: Non-normal Regressors

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

Table 2.4 reports estimation results. As expected, OLS estimates of both α and β are

biased ($t_{bias} = 15.75/8.24$) as a result of the regressor endogeneity. The estimation result of $\text{Copula}_{\text{Origin}}$ reduces the bias, but still shows significant bias for both the coefficient estimates of P_t and W_t. The bias of $\text{Copula}_{\text{Origin}}$ depends on the strength of the correlation between W and P. Stronger correlations between P^* and W^* can cause a larger bias in $\text{Copula}_{\text{Origin}}$ estimates. For example, when the correlation between W^* and P^* increases from 0.5 to 0.7, the bias of estimated α increases by around five times (from 0.055 to 0.260 in Table 2.4 under the column "$\text{Copula}_{\text{Origin}}$"). The bias confirms our derivation in the model section, demonstrating that using the existing copula method may not solve the endogeneity problem completely with correlated regressors.

We next examine our proposed method. 2sCOPE provides consistent estimates without the use of instruments. The average estimate of $\rho_{p\xi}$ is close to the true value 0.5 and is significantly different from 0, implying a significant correlation between the endogeneity regressor and the error term. Moreover, the proposed method 2sCOPE shows larger efficiency. The standard error of $\alpha(\beta)$ in 2sCOPE is 0.070 (0.042), which is 2.78% (37.31%) smaller than the corresponding standard errors using COPE. We further calculate the estimation precision of COPE and 2sCOPE using the D-error measure $|\Sigma|^{1/K}$ (Arora and Huber 2001, Qian and Xie 2021), where Σ is the covariance matrix of the parameter estimates in the regression mean function, and K is the number of these parameters. A smaller value of D-error means greater estimation efficiency and improved estimation precision. When $\rho_{pw} = 0.5$, the D-error measure is 0.002613 for COPE and 0.001614 for 2sCOPE (Table 2.4), and thus 2sCOPE increases estimation precision by 38.2%, meaning that for 2sCOPE to achieve the same precision with COPE, sample size can be reduced by 38.2%. A 39.3% of efficiency gain for 2sCOPE is found for $\rho_{pw} = 0.7$ in Table 2.4.

We perform a further simulation study for a small sample size. Specifically, we use the same DGP as described above to generate synthetic data, except with the sample size T=200. Table A.1 in the Appendix reports the results and shows that OLS estimates have endogene-

ity bias and $\text{Copula}_{\text{Origin}}$ reduces the endogeneity bias but the significant bias remains. Our proposed method, 2sCOPE, performs well and has unbiased estimates for the small sample size of T=200. The efficiency gain of 2sCOPE relative to COPE appears to be greater when the sample size becomes smaller. When the correlation between P^* and W^* is 0.5, the D-error measures are 0.0166 and 0.0091 for COPE and 2sCOPE (Appendix Table A.1), respectively, meaning that 2sCOPE increases estimation precision by 1-0.0091/.0166=46% compared with COPE, and thus sample size can be reduced by almost a half ($\sim$50%) for 2sCOPE to achieve the same estimation precision as that achieved by COPE. A similar magnitude of efficiency gain for 2sCOPE relative to COPE ($\sim$50%) is observed when the correlation between P^* and W^* is 0.7 (Appendix Table A.1).

2.4.2 Case 2: Normal Regressors

Next, we examine the case when the endogenous regressor and (or) the correlated exogenous regressor are normally distributed. We pay special attention to this case because normality is not allowed for endogenous regressors in Park and Gupta (2012). We use the Gaussian copula as described in Equations (2.24) to Equations (3.17) for DGP to generate the data, except that the marginal CDFs for regressors ($H(\cdot)$ and $L(\cdot)$) are chosen according to the distributions listed in the first two columns in Table 2.5.

Table 2.5 summarizes the estimation results. As expected, OLS estimates are biased. $\text{Copula}_{\text{Origin}}$ produces biased estimates whenever the endogenous regressor P follows a normal distribution. The estimates of $\text{Copula}_{\text{Origin}}$ are biased when P follows a gamma distribution (first row of Table 2.5) for a different reason: P and W are correlated. Same with $\text{Copula}_{\text{Origin}}$, the results of COPE are biased in all three scenarios when at least one of P_t and W_t is normal. When W_t is normal, β is 0.323 away from the true value -1; when P_t is normally distributed, α is 0.684 away from the true value; when both P_t and W_t are normal,

α is 0.663 away from the true value 1 and β is 0.324 away from the true value -1. This is expected because COPE adds P_t^* and W_t^*, the copula transformation of regressors, as additional regressors, and will cause perfect co-linearity and model non-identification problems whenever at least one of these regressors is normally distributed.

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

Table 2.5: Results of Case 2: Normal Regressors

By contrast, the proposed 2sCOPE method provides consistent estimates as long as P_t and W_t are not both normally distributed. Both α and β are tightly distributed near the true value whenever P_t or W_t is nonnormally distributed. Unlike $\text{Copula}_{\text{Origin}}$ and COPE, 2sCOPE adds the residual term obtained from regressing P_t^* on W_t^* as the generated regressor. Thus, as long as P_t and W_t are not both normally distributed, the residual term is not perfectly co-linear with the original regressors, permitting model identification. Only when both P_t and W_t are normally distributed (the last scenario in Table 2.5), the residual term added into the structural regression model becomes a linear combination of P_t and W_t,

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

2.4.3 Case 3: Performance Under Insufficient Non-Normality of Endogenous Regressors

The above section shows that the proposed 2sCOPE can deal with normal endogenous regressors, while $\text{Copula}_{\text{Origin}}$ and COPE cannot. In this section, we examine the performance of these methods in the more common situation of close-to-normal regressors. Although models are identified asymptotically (i.e., infinite sample size), appreciable finite sample bias can occur with realistic sample size commonly seen in marketing studies, if the endogenous regressor is too close to a normal distribution (Becker et al., 2021; Haschka, 2021). Becker et al. (2021) suggest a minimum absolute skewness of 2 for an endogenous regressor in order for $\text{Copula}_{\text{Origin}}$ to have good performance in a sample size as small as 200. This requirement can significantly limit the use of copula correction methods in practical applications. Given that the proposed 2sCOPE can handle normal endogenous regressors, we expect that 2sCOPE can handle much better the finite sample bias caused by insufficient regressor non-normality than the existing copula correction methods. Thus, in this subsection, we further explore and compare the finite sample performance of those methods when the distribution of the endogenous regressor has various magnitudes of closeness to normality. Specifically, we show the performance using some commonly used distributions in practice.

We use the data-generating process (DGP) below:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{p\xi} \\ \rho_{pw} & 1 & 0 \\ \rho_{p\xi} & 0 & 1 \end{bmatrix} \right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right) \quad (2.24)$$

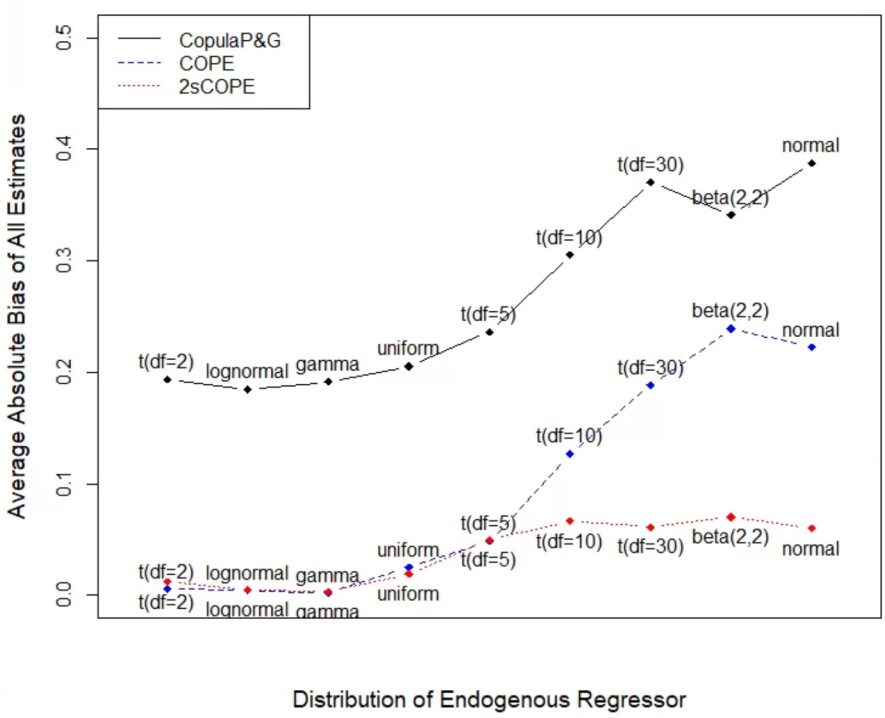
$$\xi_t = G^{-1}(U_{\xi,t}) = G^{-1}(\Phi(\xi_t^*)) = \Phi^{-1}(\Phi(\xi^*)) = 1 \cdot \xi_t^*, \tag{2.25}$$

$$P_t = H^{-1}(U_{P,t}) = H^{-1}(\Phi(P_t^*)), \quad W_t = L^{-1}(U_{W,t}) = L^{-1}(\Phi(W_t^*)), \tag{2.26}$$

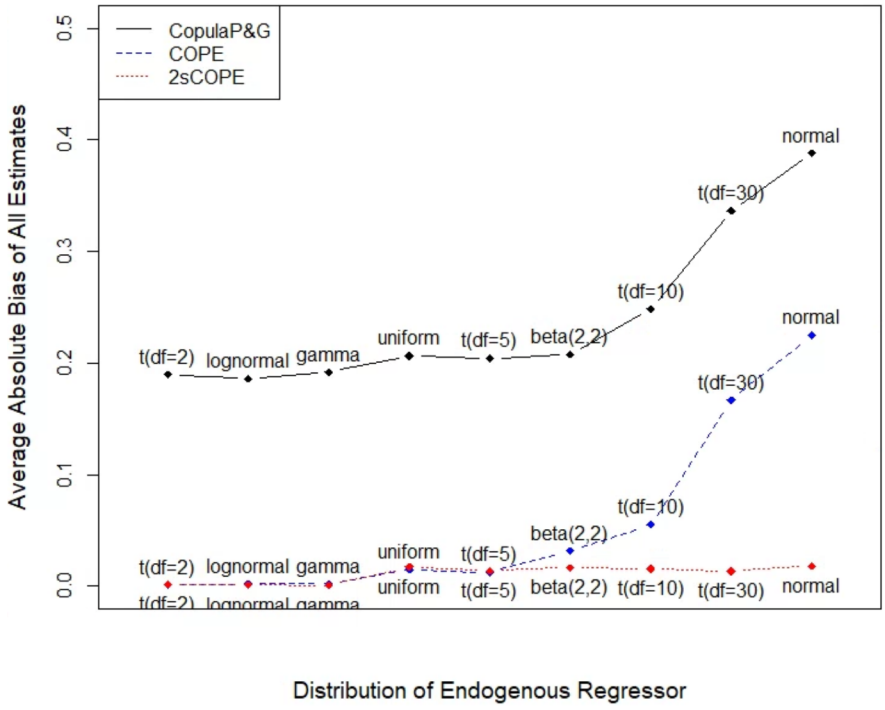
$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + \xi_t = 1 + 1 \cdot P_t + (-1) \cdot W_t + \xi_t. \tag{2.27}$$

where we still use $Exp(1)$ for W as an example. We compare the performance of $\text{Copula}_{\text{Origin}}$, COPE and 2sCOPE methods when the endogenous regressor P is from some commonly used distributions that are close to normal to different extents. Specifically, we use uniform, log normal, t, gamma, beta and normal distributions, and use the average absolute bias of all the estimated coefficients to measure the performance.

Figure 2.1 shows the estimation bias using different distributions of the endogenous regressor P. According to the figure, estimates of $\text{Copula}_{\text{Origin}}$ are biased for all distributions, consistent with our theoretical proof. Compare the estimates of COPE and 2sCOPE, first, COPE cannot handle normal endogenous regressors. This result is consistent with our theoretical proof and the simulation result in Case 2. Moreover, the bias is not even decreasing as the sample size increases. Second, COPE suffers from finite-sample bias when the endogenous regressor follows some insufficient non-normal distributions (e.g., beta distribution, t distribution with a large degree of freedom). Moreover, the smaller the sample size, the larger the estimation bias of COPE would be with insufficient non-normal endogenous regressor. This tells us that COPE doesn't perform well for close-to-normal distributions. Last but not least, the bias of COPE also indicates how close a distribution is to the normal distribution. T distribution with a degree of freedom of 30 is closer to normal than the same distribution with degrees of freedom of 10, 5 and 2, and beta distribution locates between t (df=5) and t(df=10). In contrast to COPE, our proposed 2sCOPE method can get consistent estimates for all normal and close-to-normal distributions.



(a) Sample Size N=200



(b) Sample Size N=1000

Figure 2.1: Estimation Bias for Different Distributions of Endogenous Regressor.
Note: 'lognormal' is lognormal(0,1), 'uniform' is U[0,1], and 'gamma' is $Gamma(1,1)$.

2.4.4 Case 4: Multiple Endogenous Regressors

In this case, we examine the performance of our proposed method when the model has multiple endogenous regressors. We use the data-generating process (DGP) with two endogenous regressors and one exogenous regressor that is correlated with the endogenous regressor below:

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

$$= N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 & 0.4 & 0.5 \\ 0.3 & 1 & 0.4 & 0.5 \\ 0.4 & 0.4 & 1 & 0 \\ 0.5 & 0.5 & 0 & 1 \end{bmatrix} \right), \tag{2.28}$$

$$\xi_t = G^{-1}(U_{\xi,t}) = G^{-1}(\Phi(\xi_t^*)) = \Phi^{-1}(\Phi(\xi^*)) = 1 \cdot \xi_t^*, \tag{2.29}$$

$$P_{1,t} = H_1^{-1}(U_{p1}) = H_1^{-1}(\Phi(P_{1,t}^*)),\ P_{2,t} = H_2^{-1}(\Phi(P_{2,t}^*)), \tag{2.30}$$

$$W_t = L^{-1}(U_{W,t}) = L^{-1}(\Phi(W_t^*)), \tag{2.31}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + \xi_t = 1 + 1 \cdot P_{1,t} + 1 \cdot P_{2,t} + (-1) \cdot W_t + \xi_t, \tag{2.32}$$

where $H_1^{-1}(\cdot)$ ($H_2^{-1}(\cdot)$) and $L^{-1}(\cdot)$ are the inverse distribution functions of the gamma and exponential distributions used to generate these regressors. Sample size T = 1000. We generate 1000 data sets, and use existing methods and our proposed method to estimate the model. Table 2.6 shows the estimation results. Both the OLS and $\text{Copula}_{\text{Origin}}$ estimates are biased, while our proposed method provides unbiased estimates for all parameters, indicating that our proposed method performs well with multiple endogenous regressors.

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

Table 2.6: Results of the Simulation Study Case 3: Multiple Endogenous Regressors

2.4.5 Case 5: Multiple Exogenous Control Covariates

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

$$\begin{aligned}\begin{pmatrix} P_t^* \\ W_{1,t}^* \\ W_{2,t}^* \\ \xi_t^* \end{pmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw1} & \rho_{pw2} & \rho_{\xi p} \\ \rho_{pw1} & 1 & \rho_w & 0 \\ \rho_{pw2} & \rho_w & 1 & 0 \\ \rho_{\xi p} & 0 & 0 & 1 \end{bmatrix}\right) \\ &= N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.3 & 0 \\ 0.5 & 0.3 & 1 & 0 \\ 0.5 & 0 & 0 & 1 \end{bmatrix}\right), \end{aligned} \tag{2.33}$$

$$\xi_t = G^{-1}(\Phi(\xi_t^*)) = \Phi^{-1}(\Phi(\xi^*)) = 1 \cdot \xi_t^*, \tag{2.34}$$

$$P_t = H^{-1}(\Phi(P_t^*)), \quad W_{1,t} = L^{-1}(\Phi(W_{1,t}^*)), \tag{2.35}$$

$$W_{2,t} = \begin{cases} 1, & \text{if} \quad \Phi(W_{2,t}^*) \geq 0.5 \\ 0. & \text{if} \quad \Phi(W_{2,t}^*) < 0.5 \end{cases}, \tag{2.36}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta_1 \cdot W_{1,t} + \beta_2 \cdot W_{2,t} + \xi_t = 1 + 1 \cdot P_t + (-1) \cdot W_{1,t} + (-1) \cdot W_{2,t} + \xi_t, \tag{2.37}$$

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

The estimation results for the multiple-exogenous-regressor case with both discrete and continuous ones are summarized in Table 2.7. The OLS and $\text{Copula}_{\text{Origin}}$ estimates are biased because of endogeneity and correlated exogenous regressors, respectively. The proposed 2sCOPE method performs well and provides consistent estimates for all parameters. This indicates that our proposed method performs well with multiple exogenous correlated regressors. Moreover, correcting for endogeneity using our proposed method does not require every exogenous correlated regressor to be informative (i.e., continuously distributed).

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

Table 2.7: Results of the Simulation Study Case 4: Multiple Exogenous Control Covariates

2.4.6 Case 6: Random Coefficient Linear Panel Model

We investigate the performance of our proposed 2sCOPE method in a random coefficient linear panel model. We use the copula and marginal distributions for $[P_{it}, W_{it}, \xi_{it}]$ as specified in Case 1 (Equations 2.24-3.16). We assign $\rho_{pw} = 0.7$ as an example. We then generate the outcome Y_{it} using the following standard random coefficient linear panel model:

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

where $[\mu_i, a_i, b_i] \sim N(0, I_3)$, $t = 1, ..., 50$ indexes occasions for repeated measurements, and $i = 1, ..., 500$ indexes the individual units. The above random coefficients model permits individual units to have heterogeneous baseline preferences (μ_i) and heterogeneous responses to regressors (a_i, b_i). Such random coefficient models are frequently used in marketing studies to capture individual heterogeneity and to profile and target individuals. The correlation between ξ_{it} and P_{it} creates the regressor endogeneity problem in the random coefficient model, which can cause biased estimates for standard linear random coefficient estimation methods ignoring the regressor-error correlation. We generate individual-level panel data

as described above 1000 times and use the data for estimation. Estimation results are in Table 2.8. LME is the standard estimation method for linear mixed models assuming all regressors are exogenous, as implemented in the R function $lme()$. LME and $\text{Copula}_{\text{Origin}}$ are biased because of endogeneity and correlated exogenous regressors, respectively. Our proposed method 2sCOPE provides unbiased estimates that are tightly distributed around the true values for all parameters.

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

Table 2.8: Results of the Simulation Study Case 5: Random Coefficient Linear Panel Model

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

2.4.7 Misspecification of the Error ξ_t

Similar to $\text{Copula}_{\text{Origin}}$, we assume the structural error ξ_t to be normally distributed. Though the normality of ξ_t is a reasonable and commonly used assumption in marketing and economics literature, the true distribution of ξ_t is often unknown, resulting in possible misspecifications. In this simulation study, we examine the robustness of the proposed method to the departures from the normality of ξ_t. We generate 1,000 data sets using the same multivariate

normal distribution as in Equation (2.24). The rest of DGP is:

$$\xi_t = G^{-1}(U_{\xi,t}) = G^{-1}(\Phi(\xi_t^*)), \tag{2.38}$$

$$P_t = H^{-1}(U_{p,t}) = H^{-1}(\Phi(P_t^*)), \qquad W_t = L^{-1}(U_{w,t}) = L^{-1}(\Phi(W_t^*)), \tag{2.39}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + \xi_t = 1 + 1 \cdot P_t + (-1) \cdot W_t + \xi_t. \tag{2.40}$$

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

Table 2.9 reports estimation results. OLS and $\text{Copula}_{\text{Origin}}$ estimates are still biased, consistent with the normal error case. Importantly, Becker et al. (2021) showed that $\text{Copula}_{\text{Origin}}$ has estimation bias for misspecification of ξ_t even when no W_ts are included, the case in which $\text{Copula}_{\text{Origin}}$ should be consistent. This indicates that $\text{Copula}_{\text{Origin}}$ is not robust in the misspecification of the error term. In contrast, 2sCOPE can recover the true parameter values despite the misspecification of ξ_t, demonstrating the robustness of the proposed 2sCOPE method to the normal error assumption. See more robustness check results for different distributions of the error term in the Appendix.

2.5 Empirical Application

In this section, we apply our method to a real marketing application. We illustrate the proposed method to address the price endogeneity issue using store-level sales data of the toothpaste category in Chicago over 373 weeks from 1989 to 1997 [9]. To control for product size, we select data of toothpaste with the most common size, which is 6.4 oz. Retail

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

			OLS			COPE			2sCOPE		
Distribution of ξ_t	Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
U[-0.5,0.5]	μ	1	0.912	0.013	6.808	1.004	0.025	0.144	1.002	0.017	0.105
	α	1	1.160	0.010	16.41	0.996	0.017	0.263	0.996	0.017	0.233
	β	-1	-1.072	0.009	8.033	-1.000	0.019	0.023	-0.998	0.011	0.147
	$\rho_{p\xi}$	0.5	-	-	-	0.497	0.049	0.054	0.495	0.035	0.155
	σ_ξ	0.289	0.251	0.004	9.018	0.291	0.009	0.287	0.290	0.008	0.197
Beta(0.5,0.5)	μ	1	0.896	0.016	6.461	1.005	0.031	0.166	1.003	0.020	0.145
	α	1	1.190	0.012	15.72	0.994	0.019	0.343	0.994	0.018	0.318
	β	-1	-1.086	0.011	7.763	-0.999	0.022	0.043	-0.998	0.014	0.183
	$\rho_{p\xi}$	0.5	-	-	-	0.483	0.050	0.339	0.481	0.033	0.593
	σ_ξ	0.354	0.311	0.005	9.046	0.357	0.009	0.355	0.356	0.009	0.258
t (df=3)	μ	1	0.504	0.082	6.071	0.972	0.198	0.142	0.983	0.127	0.135
	α	1	1.903	0.089	10.13	1.026	0.227	0.113	1.024	0.217	0.110
	β	-1	-1.410	0.064	6.448	-1.003	0.129	0.020	-1.012	0.109	0.111
	$\rho_{p\xi}$	0.5	-	-	-	0.449	0.088	0.577	0.454	0.069	0.676
	σ_ξ	1.732	1.503	0.231	0.992	1.701	0.246	0.124	1.698	0.244	0.141

Table 2.9: Results of the Simulation Study Case D1: Misspecification of ξ_t

price is usually considered endogenous. The endogeneity of retail price can come from unmeasured product characteristics or demand shocks that can influence both consumers' and retailers' decisions. Since these variables are unobserved by researchers, they are absorbed into the structural error, leading to the endogeneity problem. Prices of different stores are correlated and often used as an IV for each other. This allows us to test the performance of the proposed 2sCOPE method in an empirical setting where a good IV exists. Besides the endogenous price, two promotion-related variables, bonus promotion and direct price reduction, would also affect demand. Following Park and Gupta (2012), we treat the pro-

motion variables as exogenous regressors. We focus on category sales in two large stores in Chicago (referred to as Stores 1 and 2). We convert retail price, in-store promotion and sales from the UPC level to the aggregate category level. They are computed as weekly market share-weighted averages of UPC-level variables. The correlation between log retail

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

Table 2.10: Summary Statistics

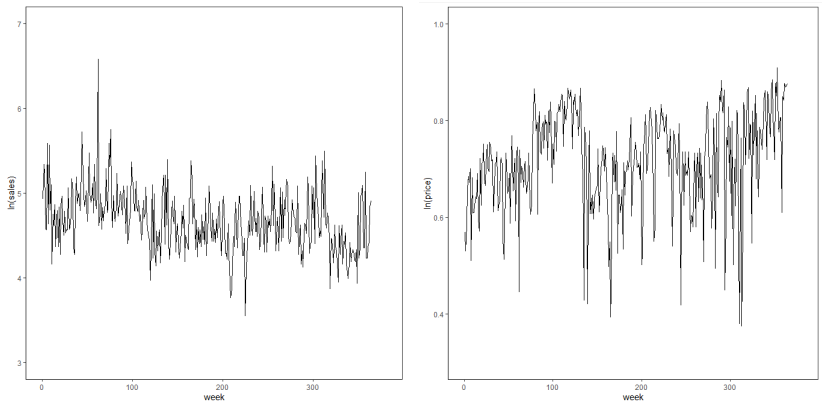
price and bonus promotion in Store 1 (Store 2) is -0.30 (-0.15), and the correlation between log retail price and price reduction promotion in Store 1 (Store 2) is -0.23 (-0.35). Both the correlations are significantly different from zero. The appreciable correlations between price and promotion variables provide a good setting for testing our method and examining the impact that our proposed method can make in the setting of correlated endogenous and exogenous regressors. Summary statistics of key variables are summarized in Table 3.2.

We estimate the following linear regression model:

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

where $t = 1, 2, ..., T$ indexes week. The vector W_t includes all exogenous regressors, which are two promotion variables, bonus promotion and price reduction, in this application.

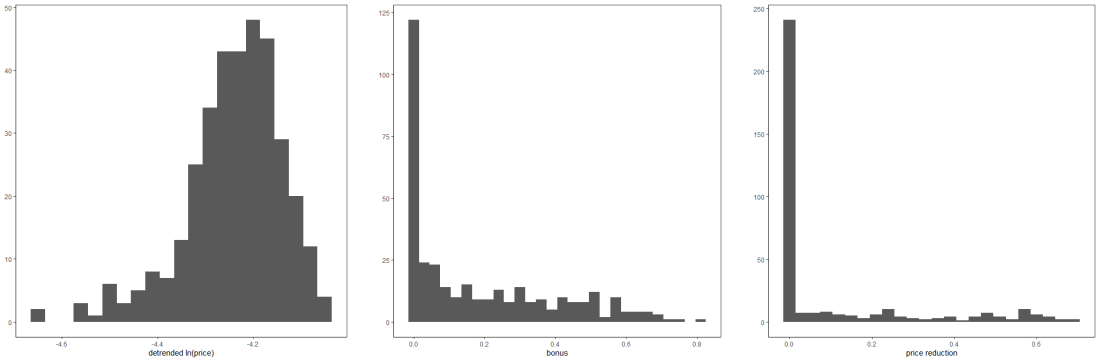
Figure 2.2 shows log sales and log retail prices of toothpaste at store 1 over time (store 2 is very similar). To control for the possible trend of retail price over time, we use the



(a) Store 1 log sales **(b)** Store 1 log retail price

Figure 2.2: Log Sales and Log Retail Price of Toothpaste in Store 1.

de-trended log retail prices (as instrumental variables as well) for estimation below.



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

Figure 2.3: Histogram of Log Retail Price, Bonus and Price Reduction in Store 1

Figure 2.3 shows the histograms of detrended log retail prices and the two promotion variables. All three variables are continuous. Moreover, except for log retail price, which is a bit close to the normal distribution, the other two regressors, bonus and price reduction, are both nonnormally distributed. Therefore, we expect that the proposed 2sCOPE method can exploit these additional features of exogenous regressors correlated with the endogenous regressor for model identification and estimation even if the endogenous regressor is close to normal distribution. We estimate the model using the OLS, two-stage least squares (TSLS), $\text{Copula}_{\text{Origin}}$ and our two proposed 2sCOPE method.

We use the IV-based TSLS estimator as a benchmark to test the validity of our proposed method. Following Park and Gupta (2012), we use retail price at the other store as an instrument for the endogenous price in the focal store. This variable can be a valid instrument as it satisfies the two key requirements. First, retail prices across stores in the same market can be highly correlated because wholesale prices are usually offered the same (or very close). The Pearson correlation between the detrended log retail prices at Stores 1 and 2 is 0.79, providing strong explanatory power on the endogenous price. The correlation is comparable to that in Park and Gupta (2012). Second, some unmeasured product characteristics such as shelf-space allocation, shelf location and category location are determined by retailers and are usually not systematically related to wholesale prices (exclusion restriction). For the three copula-based methods, we make use of information from the existing endogenous and exogenous regressors and no extra IVs are needed. In $\text{Copula}_{\text{Origin}}$, we add the copula transformation of the detrended log price, $\text{logP}^* = \Phi^{-1}(\widehat{H}(\text{logP}))$, as a "generated regressor" to the outcome regression. For the COPE method, we add another two "generated regressors", copula transformation of bonus and price reduction ($\text{Bonus}^* = \Phi^{-1}(\widehat{L}_1(\text{Bonus}))$, $\text{PriceRedu}^* = \Phi^{-1}(\widehat{L}_2(\text{PriceRedu}))$). For the 2sCOPE method, we first regress logP^* on Bonus^* and PriceRedu^*, and then add the residual as the only "generated regressor" to the outcome regression. $\widehat{H}(\cdot), \widehat{L}_1(\cdot), \widehat{L}_2(\cdot)$ are all estimated CDFs using the univariate empirical distribution for each regressor. Standard errors of parameter estimates are obtained using bootstrap.

Table 3.3 reports the estimation results. Beginning with the results from Store 1, OLS estimates are significantly different from TSLS estimates, indicating that the price endogeneity issue occurs. Instrumenting for retail price changes the price coefficient estimate from -0.767 to -1.797, implying that there is a positive correlation between unobserved product characteristics and the price. The estimates of ρ in the three IV-free copula-based methods, representing the correlation between the endogenous regressor P_t and the error

term, are all significantly positive, further confirming our previous conclusion. This direction of correlation is consistent with previous empirical findings (e.g., Villas-Boas and Winer 1999, Chintagunta et al. 2005).

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

Table 2.11: Estimation Results: Toothpaste Sales

The price elasticity estimates from the $\text{Copula}_{\text{Origin}}$, the extension COPE and the proposed method 2sCOPE are -3.082, -3.111 and -2.014, respectively. Among the three estimates, the estimate of -2.014 from the proposed 2sCOPE is close to the estimate of -1.797 from the TSLS method, whereas the existing copula and the COPE yield substantially smaller price elasticity estimates. We confirm in the literature that the TSLS and 2sCOPE estimates are reasonable because the price elasticity of the toothpaste category in the literature is around -2.0 (Hoch et al. 1995, Mackiewicz and Falkowski 2015). Comparing the estimates of ρ from the three IV-free copula-based methods, our proposed 2sCOPE provides a much smaller estimate of ρ (0.297 for 2sCOPE vs 0.521 for $\text{Copula}_{\text{Origin}}$ and 0.662 for COPE in Table 3.3), consistent with the over-correction in both $\text{Copula}_{\text{Origin}}$ and COPE.

Reasons for the substantial difference in the estimates from the $\text{Copula}_{\text{Origin}}$ include (1) it's ignoring correlated endogenous and exogenous regressors which can lead to inconsistent estimates, and (2) the unimodal close-to-normality distribution for the logarithm of the

price variable leading to potentially poor finite sample performance. In fact, the correlations between logP* and the exogenous regressors are -0.44 for Bonus and -0.26 for PriceRedu, both of which are substantially larger than the corresponding correlations (-0.30 and -0.15, respectively) between logP and the exogenous regressors. The p-value for the null hypothesis of these correlations being zeros are significantly less than 0.05 (< 0.001), indicating a violation of Assumption 4 required for $\text{Copula}_{\text{Origin}}$ to yield consistent estimates.

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

Unlike Store 1, the results from Store 2 indicate that the retail price is not endogenous. First, The estimates of ρ, which is the correlation between price and the error term, are not significantly different from 0 for both $\text{Copula}_{\text{Origin}}$ and 2sCOPE (t-value ≤ 1.96 under columns "$\text{Copula}_{\text{Origin}}$" and "2sCOPE" for Store 2 in Table 3.3), and only slightly significantly different from 0 for COPE (a t-value of 2.18, slightly larger than 1.96 under Column "COPE" in Table 3.3). The estimate of ρ for COPE, however, is questionable because of the limitations of COPE mentioned in the paragraph above. Second, the estimated price coefficient of OLS is -1.982, which is very close to the estimates of TSLS and 2sCOPE in store 1 and further confirms no endogeneity of price in store 2. Overall, the price elasticity estimates from TSLS and the three IV-free copulas-based methods are close to each other for Store 2, and the observed differences between them and the OLS estimate can be at-

tributed to estimation variability incurred from using more complicated models instead of the presence of endogeneity.

In sum, the convergence of results between TSLS and the proposed method 2sCOPE in both stores supports the validity of the proposed method in addressing the endogeneity issue. Moreover, the difference between the estimates in COPE and 2sCOPE in store 1 shows the advantages of 2sCOPE in terms of relaxing the non-normality assumption of the endogenous regressor and estimation efficiency gain by exploiting additional information from correlated exogenous regressors.

2.6 Economic Intuition and Practical Guidance of 2sCOPE

In this section, we intuitively discuss how our proposed 2sCOPE works for correcting endogeneity, and the practical guidance of 2sCOPE.

Intuitively, the proposed 2sCOPE method divides the error term into two parts, one being an endogenous part, which is the first-stage residual that is correlated with the endogenous variable P, and the other being an exogenous part. The 2sCOPE method corrects endogeneity by directly controlling for the endogenous part in the error term. Suppose we are using 2sCOPE to estimate the effect of price on quantity demanded, a classical marketing question. P is the endogenous price, W contains all observed variables that are correlated with price and would affect demand (e.g., observed product attributes), and Y is the observed demand. The remaining unobserved variables that also influence demand are all contained in the error term, and some that are correlated with the price can cause the endogeneity problem. Those unobservables can be product attributes that are observed by customers but not researchers (i.e., the smell of perfume products). Then in the first-stage regression by regressing P^* on W^*, W^* helps to explain part of the price that influences demand, and

the residual can be interpreted as containing all the unobserved variables (e.g., unobserved product attributes) that cause the endogeneity plus the errors. Adding the residual as a generated regressor in the demand model can be interpreted as controlling for all the unobserved variables that cause the correlation between price and error term. Here, we take the effect of price on demand as an example. If supply-side data are observed, 2sCOPE can also be used to estimate the effect of price on supply.

Our proposed 2sCOPE method empirically provides sound causal inference once the observational data are given. Data is not limited to cross-sectional data, 2sCOPE can also correct endogeneity for panel data. For panel data analysis, researchers use fixed effects to control for endogeneity. But adding fixed effects might not be enough, and our 2sCOPE method can use copula to further correct endogeneity besides panel fixed effects. Moreover, 2sCOPE is straightforward to be applied to a wider range of commonly used marketing models, including linear regression models, linear panel models with mixed effects, random coefficient logit models and slope endogeneity. Any model that can use the control function approach can in principle apply our method for correcting endogeneity.

We have listed in the literature review different approaches for correcting endogeneity. Each method has its own assumptions. 2sCOPE method requires the normality assumption on the error term (Villas-Boas and Winer 1999, Yang et al. 2003, Ebbes et al. 2005), and the Gaussian copula relationship among regressors and the error term. Compare 2sCOPE with IV approach, IV corrects endogeneity by using extra exogenous shock from the other side while 2sCOPE identifies the model by controlling the endogenous part from existing regressors themselves. The exclusion restriction is difficult to achieve in IV, while the Gaussian copula structure required in 2sCOPE, which results in the normality assumption of the first stage residual, is very general and robust (Danaher and Smith 2011). Eckert and Hohberger (2022) further use simulations to explore the practical usefulness of Gaussian Copula approach and compare it with OLS, IV regression, and Higher Moments (HM) esti-

mator. They find that the performance of Gaussian Copula approach is as good as a strong IV case when assumptions are met. Generally speaking, our 2sCOPE method can be used in practice when it's hard to find a strong instrumental variable or when experiments are not available to conduct. But future research is needed for comparing more methods and providing clearer guidance on when to use our 2sCOPE method.

2.7 Conclusion

Causal inference lies at the center of social science research, and observational studies often beg rigorous post-study designs and methodologies to overcome endogeneity concerns. In this paper, we focus on the instrument-free copula method to handle the problem of endogenous regressors. We propose a generalized two-stage copula endogeneity correction (2sCOPE) method that overcomes two key limitations of the existing copula-based method in Park and Gupta (2012) ($\text{Copula}_{\text{Origin}}$), and extends $\text{Copula}_{\text{Origin}}$ to more general settings. Specifically, 2sCOPE allows exogenous regressors to be correlated with endogenous regressors and relaxes the nonnormality assumption on the endogenous regressors. To demonstrate the benefits of 2sCOPE, we compare it with the direct extension of $\text{Copula}_{\text{Origin}}$ method, called COPE. Similar to the $\text{Copula}_{\text{Origin}}$, 2sCOPE method corrects endogeneity by adding "generated regressors" derived from the existing regressors and is straightforward to use. COPE is a direct extension to $\text{Copula}_{\text{Origin}}$ by adding latent copula transformation of existing regressors, while 2sCOPE has two stages and adds the residuals from regressing latent copula data for the endogenous regressor on the latent copula data for the exogenous regressors as a "generated regressor" in the structural regression model. We theoretically prove that 2sCOPE can yield consistent causal-effect estimates when exogenous regressors are correlated with the endogenous regressors, which can cause biased estimates in the method of Park and Gupta (2012). Moreover, the 2sCOPE method can further relax the nonnormality assumption on the endogenous regressors and improve estimation efficiency.

We conduct simulation studies and use an empirical marketing application to empirically verify the performance of our proposed method. The simulation results show that 2sCOPE yields consistent estimates under relaxed assumptions. Moreover, 2sCOPE method outperforms COPE in terms of dealing with normal endogenous regressors and improving estimation efficiency. Endogenous regressors are allowed to be normally distributed as long as one of the exogenous regressors is nonnormally distributed, which is a very weak assumption. The efficiency gain is substantial and can be up to $\sim$50%, meaning that the sample size can be reduced by $\sim$50% to achieve the same estimation efficiency as compared with COPE method which does not exploit the correlations between endogenous and exogenous regressors. Last but not least, our robustness checks show that the proposed method 2sCOPE is reasonably robust to the structural error distributional assumption and non-Gaussian copula correlation structure (see details in Appendix). We further apply our 2sCOPE method to a commonly used public dataset in marketing. When dealing with endogenous price, we find that the estimated price coefficient using our proposed 2sCOPE is very close to the TSLS estimate, while OLS and $\text{Copula}_{\text{Origin}}$ show large biases. Moreover, the difference between the results of 2sCOPE and COPE demonstrates the advantage of 2sCOPE in dealing with (nearly) normal endogenous regressors and improving estimation efficiency.

These findings have rich implications for guiding the practical use of copula-based instrument-free methods to handle endogeneity. A known critical assumption for $\text{Copula}_{\text{Origin}}$ is the non-normality of endogenous regressors. The users of the method in the literature have all been practicing the check and verification of this assumption. However, our work shows that this is insufficient: one also needs to check Assumption 4 for the one-endogenous-regressor case, and Assumption 4(b) for the multiple-endogenous-regressors case. Note that neither assumption is the same as checking the pairwise correlations between the endogenous and exogenous regressors. Assumption 4 evaluates pairwise correlations involving copula transformation of the endogenous regressor, which, as shown in our empirical

application, can be substantially different from the pairwise correlations using the regressor itself (Danaher and Smith 2011). Assumption 4(b) evaluates the correlations between exogenous regressors and the linear combination of generated regressors, which are even more different from checking pairwise correlations on the regressors themselves. When the above assumptions are satisfied, $\text{Copula}_{\text{Origin}}$ is preferred to our proposed 2sCOPE method, since the simpler and valid model outperforms more general but more complex models.

If any endogenous regressor fails to have sufficient departure from being normally distributed, or any exogenous regressor violates Assumption 4 or 4(b), our proposed 2sCOPE method should be used instead of $\text{Copula}_{\text{Origin}}$. The 2sCOPE is straightforward to extend to many other settings, and we have derived 2sCOPE for a range of commonly used marketing models, including linear regression models, linear panel models with mixed-effects, random coefficient logit models and slope endogeneity. The 2sCOPE method proposed here can be applied to these cases and many other cases not studied here, while accounting for correlations between exogenous and endogenous regressors and exploiting the correlations for model identification in the presence of insufficient non-normality of endogenous regressors.

Although the proposed 2sCOPE contributes to the literature by relaxing key assumptions of the existing copula method $\text{Copula}_{\text{Origin}}$ and extending it to more general settings, it is not without limitations. For the 2sCOPE to work best, the distributions of the endogenous regressors need to contain adequate information. The condition is violated when the endogenous regressors follow Bernoulli distributions or discrete distributions with small support, as noted in Park and Gupta (2012). The proposed 2sCOPE method does not address this limitation. Developing instrument-free methods to handle such inadequately distributed endogenous regressors is an important topic for future research. The simplicity of 2sCOPE hinges on the normal structural error and Gaussian copula dependence structure. Although 2sCOPE demonstrates reasonable robustness to departures from these assumptions as shown in the Appendix, future research is needed for more flexible methods testing and relaxing

these assumptions. Despite these limitations, we expect that the proposed 2sCOPE will provide a useful alternative to a broad range of empirical problems when instruments are not available.

Chapter 3

Lasso-based Instrument-free Causal Inference

3.1 Introduction

Copula-based instrument-free method to correct endogeneity and draw causal inference has aroused wide interest because of its simplicity to implement as no instruments are needed. Yang et al. (2022) proposed a general two-stage copula method called 2sCOPE to account for endogeneity by adding a generated regressor, which is the first-stage residual, to the structural regression model, analogous to the control function approach. The statistical tool copula is a cumulative distribution function used to model multiple variables jointly. Yang et al. (2022) used copula to model the joint distribution among the endogenous regressors, exogenous regressors and the error term, and obtain the first-stage residual by regressing latent copula data for each endogenous regressor on the latent copula data for the exogenous regressors. Intuitively, the residual in the first stage contains all the unobserved variables that make the endogenous regressor and the error term correlated. By controlling the residual,

endogeneity can be corrected because all the unobservables that cause the endogeneity are controlled for. Thus, the residual plays a central role in correcting for endogeneity using 2sCOPE, and the wrong estimation of the residual can cause bias in the estimation of the endogenous regressor.

A problem empirical researchers face when using the conditional-on-residual identification strategy (i.e., control function approach) is that the residual is unknown and is relying on knowing what variables to include in the first-stage regression. The problem is inevitable, especially for high-dimensional data, which are becoming increasingly common in this big data era. Belloni et al. (2014a) pointed out that high-dimensional data can arise through the inherent high-dimensional features. Examples are conventional data such as census and survey data, scanner data and genomics data (Peng et al. 2010) , and unconventional data that is too high-dimensional for standard estimation methods, including image (Zhang et al. 2017) and language/text data (Amado et al. 2018) that we conventionally had not even thought of as data we can work with (Mullainathan and Spiess 2017). We cannot avoid high-dimensional data in empirical analysis. Sometimes economic intuition would help in suggesting a set of variables that might be important, but it cannot identify exactly what variables are important. With high-dimensional data and too many irrelevant variables, the traditional endogeneity-correction methods may have poor performance with finite-sample bias. Moreover, the dimension of variables can even be larger than the sample size, making the traditional estimation methods infeasible.

In this paper, we combine the causal inference method with machine learning techniques to deal with the high-dimensional problem in drawing causal inference. Machine learning (ML) is a powerful tool for data analysis, especially for large or high-dimensional data sets. Researchers are recently trying to apply the powerful machine learning tool to diverse areas, such as Economics (Athey and Imbens 2019), Marketing (Cui et al. 2006, Ascarza 2018, Ngai and Wu (2022)) and operation research (Feldman et al. 2022). However, unlike

the single-equation prediction setting in most papers that apply the ML tool, our causal inference setting, which takes a two-stage structure, makes the application of ML different and more complicated. Several papers recently have tried to apply ML to some economic models for causal inference. For example, Belloni et al. (2012) applied ML instrumental variable models, and Belloni et al. (2014a) and Belloni et al. (2014b) used ML to study treatment effects.

Our paper studies the two-stage instrument-free method 2sCOPE to correct endogeneity with machine learning techniques. Specifically, we use lasso-based feature selection methods for the first-stage estimation of 2sCOPE because the copula relationship among the endogenous and exogenous regressors in the first stage are linearly correlated. Lasso method has been widely used for feature selection in the literature (Tibshirani 1996, Belloni et al. 2012, Belloni et al. 2014a, Belloni et al. 2014b, Belloni and Chernozhukov 2013, Bai and Ng 2008, Bai and Ng 2009, Javanmard and Montanari 2014a). But since Lasso selects important variables by adding a penalty term of non-zero coefficients to the least squares optimization function, the Lasso estimator has estimation bias. After noticing this drawback, researchers are devoted to methodology development in alleviating or even eliminating the bias. For example, Belloni et al. (2012) proposed a post-Lasso method to alleviate the bias of the Lasso estimator, and Javanmard and Montanari (2014a) proposed a de-biased lasso method. In this chapter, we propose a method combining the 2sCOPE method with different Lasso-based methods (Lasso, post-Lasso, de-biased Lasso) in the estimation of first-stage residual, and examine how Lasso-based methods work for correcting endogeneity and drawing causal inference for high-dimensional data. We demonstrate the performance of the proposed lasso-based 2sCOPE, compared with the regular 2sCOPE method without feature selection methods in the first stage, via simulation studies and real-data application. The simulation result shows that using Lasso-based methods in the first stage can improve the estimation accuracy and efficiency by around 50%, as compared with the regular 2sCOPE

method in high-dimensional data. We further apply our lasso-based 2sCOPE method to real data, examining how governments' policy stringency in the COVID-19 period affects citizens' happiness. We use country-level cross-sectional data for analysis. The dimension of country-level characteristics is relatively large, around half of the sample size. The estimation result shows that using Lasso-based methods in the first stage makes the effect of policy stringency on happiness more significantly negative compared with the regular 2sCOPE method without feature selection. Moreover, the estimates using different Lasso-based methods are quite robust.

The remainder of this paper unfolds as follows. Section 3.2 reviews the related literature on causal inference with feature selection. In Section 3.3, we propose a method that combines 2sCOPE with lasso-based methods. In Section 3.4, we evaluate the performance of our proposed lasso-based 2sCOPE method using simulation studies and compare the performance with the 2sCOPE method without feature selection. In Section 3.5, we apply the proposed lasso-based 2sCOPE method to examine the effect of policy strictness of COVID-19 on citizens' happiness using high-dimensional country-level cross-sectional databases. We conclude the paper and discuss future work in Section 3.6.

3.2 Literature Review

There is a large literature in marketing, economics and statistics, focusing on approaches to addressing endogeneity when inferring causal effects. Among them, the most commonly used approach is the instrumental variable approach (Angrist and Krueger 2001, Rossi 2014). Though the theory of instrumental variables is well-developed, good instruments are extremely difficult to find in practice. Rossi (2014) summarized the most frequently used instrumental variables in the literature, and pointed out that many of them are weak IVs with poor performance or even fail the validity (exclusion restriction) condition. This

forces researchers to seek for new methodologies to correct endogeneity and draw causal inference.

One stream of methodologies called the instrument-free method has recently aroused wide interest. Literally speaking, instrument-free method means no instrumental variables are needed. Park and Gupta (2012) proposed a method using the statistic tool copula to correct endogeneity by directly modeling the association between the endogenous variable and the error term via copula. By doing so, it can address the lack of suitable instruments issue using information in the existing regressors themselves without any extra information. After that, the copula method has been rapidly adopted by researchers to deal with the endogeneity problem because of its feasibility (Burmester et al. 2015, Datta et al. 2015, Gruner et al. 2019, Keller et al. 2019, Bombaij and Dekimpe 2020, Guitart et al. 2018, Lamey et al. 2018, Wetzel et al. 2018, Heitmann et al. 2020, Atefi et al. 2018, Elshiewy and Boztug 2018). Regarding methodology development, Yang et al. (2022) and Haschka (2021) further extend the method to a more general setting. Both the papers allow exogenous regressors that are correlated with the endogenous regressor to be included in the model, while the original paper doesn't allow that. But compared with Haschka (2021), the method proposed in Yang et al. (2022) has weaker assumptions on the regressors, is simpler to implement and can be applied to a wider range of models because it uses the control function approach by adding a generated regressor to the outcome regression. However, none of the papers studies causal inference when high-dimensional data are present, which is quite common in this big data era. The performance of the method relies on the first-stage residual (in the control function approach), conditional on which the endogeneity can be corrected. But the residual is unknown and needs estimation. With too many irrelevant controls in the first stage, the performance of first-stage regression can be poor with finite-sample bias, thus causing amplifying bias in the estimation of the endogenous variable.Moreover, sometimes the dimension of controls can even be larger than the sample size, making the traditional

estimation methods infeasible. To solve the high-dimensional problem, we propose a combined 2sCOPE method in Yang et al. (2022) with lasso-based feature selection methods that help to select important control variables in the first stage Our paper contributes to the literature on instrument-free approach to addressing the endogeneity problem by combining with machine learning methods for estimating the first-stage regression on high-dimensional exogenous control variables.

Another related literature is feature selection in high-dimensional data for causal inference. There is a large and growing literature on Lasso-based methods for feature selection (Tibshirani 1996, Belloni et al. 2012, Belloni et al. 2014a, Belloni et al. 2014b, Belloni and Chernozhukov 2013, Bai and Ng 2008, Bai and Ng 2009, Javanmard and Montanari 2014a). Recently, researchers start to consider applying Lasso-based methods to economic models for causal inference. For example, Belloni et al. (2012) proposed a Lasso-based method for instrument selection in linear instrumental variables models with many instruments. Belloni et al. (2014b) and Belloni et al. (2014a) used Lasso methods to study the treatment effects with high-dimensional controls in a partially linear model. But the challenge is that when the goal is causal inference, model selection can be problematic. The model selection procedures are originally designed for forecasting (prediction). When doing model selection for causal inference, it's a two-stage model instead of just a one-equation prediction, and the model selection method is acting directly on the endogenous variable in the first stage. One cannot guarantee that exactly all the variables with nonzero coefficients are perfectly selected, and the omission of some variables with small effects can contaminate causal inference results based on the selected set of variables (Leeb and Pötscher 2008a, Leeb and Pötscher 2008b). As mentioned previously, several papers have tried applying machine learning methods to draw the causal inference. Belloni et al. (2012) studied instrument selection in instrumental variables models with many instruments. Belloni et al. (2014b) and Belloni et al. (2014a) studied the treatment effects with high-dimensional controls.

In this paper, we address the endogeneity problem using an instrument-free approach 2sCOPE with high-dimensional exogenous variables. We apply Lasso-based methods in the first stage to select important and meaningful controls in explaining the endogenous variable and use the estimated residual to draw causal inference. It's natural to use the Lasso-based methods because of the linear relationship among the copula-transformed regressors in the first stage. But one drawback of the lasso method is that it has shrinkage bias. Belloni et al. (2012) proposed a post-Lasso method to alleviate the bias, and Javanmard and Montanari (2014a) developed a de-biased Lasso method. But no paper has examined how those lasso-based methods work in the instrument-free model to correct endogeneity with high-dimensional data. Our paper contributes to the feature selection for causal inference literature by being one of the first to study feature selection in the copula-based instrument-free method. We propose a combined lasso-based method with 2sCOPE, and compare the performance using different Lasso-based methods to draw causal inferences in the instrument-free approach setting.

3.3 Methods

In this section, we develop a copula-based instrument-free method that is combined with machine learning methods to handle endogenous regressors when the dimension of the exogenous variables is huge, relative to the sample size. Specifically, we build on the two-stage Copula (2sCOPE) method that deals with endogeneity (Yang et al. 2022) and then use the Lasso-based methods to select important features among the exogenous variables in the first stage to better explain the endogenous variable and control the endogeneity.

3.3.1 Copula Endogeneity Correction Method (2sCOPE) in Standard Case

We first introduce the concept of copula and review the 2sCOPE method (Yang et al. 2022) in the standard low-dimension case. Consider the following linear structural regression model with an endogenous regressor and a vector of exogenous regressors:

$$Y_t = \mu + P_t\alpha + W_t'\beta + \xi_t \tag{3.1}$$

where $t = 1,2,...,T$ indexes either time or cross-sectional units, Y_t is a (1×1) dependent variable, P_t is a (1×1) endogenous regressor, W_t is a $(\mathrm{k}\times1)$ vector of exogenous regressors, ξ_t is the structural error term, and (μ,α,β) are model parameters. P_t is correlated with ξ_t, which generates the endogeneity problem.

The key idea of the 2sCOPE method is to divide the error term into an endogenous and an exogenous part by jointly modeling the relationship among the endogenous regressor P_t, exogenous regressors W_t and the error term ξ_t using copula, and then control for the endogenous part, the first-stage residual, to correct endogeneity. The statistical tool copula is a cumulative distribution function used to model multiple variables jointly. It outperforms the traditional multivariate distribution in that the marginal distributions are not needed to be from the same distribution family. (ξ_t,P_t,W_t) is assumed to follow Gaussian Copula for its many desirable properties (Danaher 2007; Danaher and Smith 2011).

In statistics, a copula is a multivariate cumulative distribution function that models the dependence among variables without imposing assumptions on marginal distributions. Let $F(P,W,\xi)$ be the joint cumulative distribution function (CDF) of the endogenous regressor P_t and the structural error ξ_t with marginal CDFs $H(P)$, $L(W)$ and $G(\xi)$, respectively. For notational simplicity, we may omit the index t in P_t and ξ_t below when appropriate. According to Sklar's theorem (Sklar 1959), there exists a copula function $C(\cdot,\cdot)$ such that for

all P, W and ξ,

$$F(P,W,\xi) = C(H(P),L(W),G(\xi)) = C(U_p,U_w,U_\xi), \tag{3.2}$$

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

To capture the association between the endogenous regressor P, W_t and the error ξ, we use the Gaussian copula (Park and Gupta 2012, Haschka 2021, Yang et al. 2022) for its many desirable properties (Danaher and Smith 2011):

$$\begin{aligned} F(P,W,\xi) &= C(U_p,U_w,U_\xi) = \Psi_\Sigma(\Phi^{-1}(U_p),\Phi^{-1}(U_w),\Phi^{-1}(U_\xi)) \\ &= \frac{1}{\sqrt{(2\pi)^3|\Sigma|}} \int_{-\infty}^{\Phi^{-1}(U_p)} \int_{-\infty}^{\Phi^{-1}(U_w)} \int_{-\infty}^{\Phi^{-1}(U_\xi)} \exp\left[\frac{-[s,t,q]'\Sigma^{-1}[s,t,q]}{2}\right] dsdtdq, \end{aligned} \tag{3.3}$$

where $\Phi(\cdot)$ denotes the univariate standard normal distribution function and $\Psi_\Sigma(\cdot,\cdot)$ denotes the multivariate standard normal distribution with the covariance matrix Σ. In the Gaussian copula model, Σ captures the correlation among variables.

The Gaussian copula assumption on (P_t, W_t, ξ_t) is equivalent to that $[P_t^*, W_t^*, \xi_t^*]$ follows the multivariate normal distribution:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{p\xi} \\ \rho_{pw} & 1 & 0 \\ \rho_{p\xi} & 0 & 1 \end{bmatrix} \right), \tag{3.4}$$

where $P_t^* = \Phi^{-1}(H(P_t))$, $W_t^* = \Phi^{-1}(L(W_t))$, and $\xi_t^* = \Phi^{-1}(G(\xi_t))$. $H(\cdot)$, $L(\cdot)$ and $G(\cdot)$

are marginal CDFs of P_t, W_t and ξ_t respectively. We call P_t^*, W_t^*, ξ_t^* the Gaussian copula transformations of P_t W_t, ξ_t, respectively.

The above multivariate distribution can be rewritten as follows:

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

$$\begin{pmatrix} \omega_{1,t} \\ \omega_{2,t} \\ \omega_{3,t} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right). \tag{3.5}$$

Then, we have a linear relationship between P^* and W^*, and here comes the first-stage regression,

$$P_t^* = \rho_{pw} W_t^* + \varepsilon_t \tag{3.6}$$

and the structural error in Equation (3.1) can be re-expressed as

$$\begin{aligned} \xi_t = \sigma_\xi \cdot \xi_t^* &= \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* + \frac{-\sigma_\xi \rho_{pw}\rho_{p\xi}}{1-\rho_{pw}^2} W_t^* + \sigma_\xi \sqrt{1-\rho_{p\xi}^2-\frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}}\,\omega_{3,t}, \\ &= \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2}(P_t^* - \rho_{pw} W_t^*) + \sigma_\xi \sqrt{1-\rho_{p\xi}^2-\frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}, \\ &= \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2}\varepsilon_t + \sigma_\xi \sqrt{1-\rho_{p\xi}^2-\frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t} \end{aligned} \tag{3.7}$$

In this way, the structural error term ξ_t is split into two parts: one part is the first-stage residual, which is a function of P_t^* and W_t^* that captures the endogeneity of P_t and the association of W_t with $\xi_t|P_t$ [1], and the other part is an independent new error term. Then, we

[1] Although the exogenous regressor W_t and ξ_t are uncorrelated, W_t and $\xi_t|P_t$ (the error component in ξ_t remaining after removing the effect of the endogenous regressor P_t) can be correlated.

substitute Equation (A.2) into the main model in Equation (3.1), and obtain the following regression equation:

$$Y_t = \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2}\varepsilon_t + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}. \tag{3.8}$$

Equation (3.8) suggests adding the estimate of the residual ε_t from the first stage regression (Equation 3.6) as a generated regressor to the outcome regression. Adding the first stage helps model identification with extra information, the correlation between P_t^* and W_t^*, making the new error term $\omega_{3,t}$ to be uncorrelated with all the regressors in Equation (3.8) and thus consistent estimates can be obtained. This two-step 2sCOPE method adds the first-stage residual term $\widehat{\varepsilon}_t$ to control for endogeneity and in this aspect is similar to the control function approach of Petrin and Train (2010), while requires no use of instrumental variables.

Intuitively, the usage of the copula function allows us to directly model the relationship among the endogenous regressors, the observed exogenous regressors and the structural error. The correlated exogenous regressors added in the first stage help to capture and explain part of the endogenous regressor. By controlling the residual ε_t obtained from the first stage, we actually control all the unobserved variables that cause the endogeneity problem, and thus can correct endogeneity and get consistent estimates.

However, it becomes problematic when we have a large dimension of W_t, and we don't know which ones in W_ts are important in explaining the endogenous regressor. This problem is similar to the many-instruments problem (Belloni et al. 2012). A naive approach is to include all the possible exogenous variables in the first stage (regular 2sCOPE). However, doing so would make the performance of finite-sample prediction worse and inefficient. In the section below, we propose a lasso-based method to select important features in the first-stage of 2sCOPE in helping explain the endogenous regressor.

3.3.2 2sCOPE with Lasso and post-Lasso

In this subsection, we propose a method to deal with the case when the dimension of exogenous regressors is huge, but only a few are correlated with the endogenous regressor. This is actually very common in practice. Specifically, we combine the 2sCOPE method in dealing with endogeneity with a feature selection machine learning method in the first stage to select important W_ts when the dimension of W_t is large relative to the sample size. If too many irrelevant variables are included in the first stage to explain the endogenous regressor, it would affect both the finite-sample estimation accuracy and efficiency.

To eliminate the too-many-control issue, we use a Lasso method in the first stage of the 2sCOPE method in Equation (3.6). Under the Gaussian copula assumption in Equation (3.4), we have a linear relationship between P_t^* and all the W_t^*s,

$$P_t^* = W_t^* \gamma + \varepsilon_t$$

where P_t^* and W_t^* are the Gaussian copula transformations of P_t and W_t, respectively. Consider the usual least squares optimization function:

$$\hat{Q}(\gamma) := \frac{1}{2n} \|P^* - W^* \gamma\|_2^2, \tag{3.9}$$

Since the dimension of W is large, which can even be larger than the sample size, we need to select important features to obtain better estimation in the first stage, and thus improve the estimation accuracy and efficiency of the endogenous regressor to draw the causal inference. Because of the linear relationship between P^* and W^* and the sparsity of important Ws, it's very natural to use Lasso regression to penalize too many parameters. Lasso solves for regression coefficients by minimizing the sum of the usual least squares objective function and a penalty for model size. The Lasso estimator (Tibshirani 1996) is defined as a solution

of the optimization program below:

$$\hat{\gamma}_l(P^*, W^*; \lambda) = \arg\min_{\gamma \in R^k} \hat{Q}(\gamma) + \lambda \|\gamma\|_1, \tag{3.10}$$

where λ is the penalty level. In the estimation procedure, we use cross validation to tune the penalty level. Specifically, we set $\lambda = \lambda_{\min}$, for which we can get the minimum mean cross-validated error. Using Lasso method to form first-stage prediction provides an effective approach to obtain efficiency gains or even accuracy for finite sample estimation. However, since the Lasso estimator reduces the dimension of variables and estimates the first-stage regression coefficients by adding a penalty of non-zero coefficients based on the usual least squares objective function, the estimates of Lasso would be biased. Thus, we further use a post-lasso method to alleviate the bias. The post-Lasso estimator discards the Lasso estimates, but will take advantage of variables selected by Lasso and then refit the first-stage regression using the selected variables via OLS estimation.

Denote I_l the variables selected among W^* in the first stage using the Lasso method. The post-Lasso estimator is defined as the ordinary least squared estimator among variables $I_{pl} \supseteq I_l$. I_{pl} contains the variables in I_l and some meaningful variables according to real data application (I_m). The post-Lasso estimator $\hat{\gamma}_{pl}$ is

$$\hat{\gamma}_{pl} \in \arg \min_{\gamma \in R^k : \gamma_{(\hat{I}_l \cup I_m)^c} = 0} \hat{Q}_l(\gamma) \tag{3.11}$$

The Lasso and post-Lasso methods are motivated by the desire to predict the first-stage residual well without overfitting, and thus gain more estimation accuracy and efficiency for correcting the endogeneity. It can also deal with the case when the dimension of variables is even larger than the sample size, which would cause a problem for OLS estimation. Once we get estimates using Lasso and post-Lasso in the first stage, we calculate the residual and then add it as a generated regressor to the outcome regression model. In this way, we

combine the 2sCOPE method with the lasso-based machine learning methods to deal with the high-dimension case.

3.3.3 2sCOPE with De-biased Lasso

As mentioned above, though the Lasso method is powerful in reducing the dimension of variables, the Lasso estimator is biased. The post-Lasso method introduced above can alleviate the bias to some extent by running OLS estimation after variables are selected by Lasso. However, the post-Lasso estimator can still have bias. In this section, we introduce the de-biased Lasso method proposed by Javanmard and Montanari (2014a) that can eliminate the bias using Lasso-based methods, and combine it with the 2sCOPE method to address the endogeneity problem with high-dimensional exogenous controls.

The de-biased Lasso estimator $\hat{\gamma}_{dl}$ of the first-stage regression is,

$$\hat{\gamma}_{dl} = \hat{\gamma}_l(\lambda) + \frac{1}{n} M (W^*)^T (P^* - W^* \hat{\gamma}_l(\lambda)). \tag{3.12}$$

where $\hat{\gamma}_l(\lambda)$ is the Lasso estimator in Equation (3.10), and M is a chosen matrix depending on the P^*. We can find that the second part $\frac{1}{n} M (W^*)^T (P^* - W^* \hat{\gamma}_l(\lambda))$ is proportional to a subgradient of the l_1 norm at the Lasso estimator, $(W^*)^T (P^* - W^* \hat{\gamma}_l(\lambda))/(n\lambda)$. So intuitively, the de-biased lasso estimation $\hat{\gamma}_{dl}$ can compensate the bias introduced by the l_1 penalty in the lasso by adding a term proportional to the subgradient.

The quality of the debiasing procedure depends on the choice of M. Javanmard and Montanari (2014b) suggests the choice $M = c\Sigma^{-1}$ with $\Sigma = E\{P^*(P^*)^T\}$ being the population covariance matrix and c a positive constant. Javanmard and Montanari (2014a)

suggested setting $M = (m_1, m_2, ..., m_k)^T$ by solving a convex program below.

$$\begin{aligned} &\text{minimize} \quad m^T \hat{\Sigma} m \\ &\text{subject to} \quad \|\hat{\Sigma} m - e_i\|_\infty \leq \mu, \end{aligned} \tag{3.13}$$

where $e_i \in R^k$ is the vector with one at the i^{th} position and zero everywhere else, and $\hat{\Sigma} = ((P^*)^T P^*)/n$. The benefit of this choice M is that the method requires weaker assumptions on P^*. It can be applied to general covariance structures Σ, while Javanmard and Montanari (2014b) can only be applied to sparse Σ. In this paper, we use the de-biased lasso estimator in Javanmard and Montanari (2014a) (Equation 3.12) with M solved by Equation 3.13 for the first-stage estimation in 2sCOPE in Equation (3.6).

3.4 Simulation Study

In this section, we conduct Monte Carlo simulation studies for the following goals: (a) to assess the performance of the proposed method for high-dimensional correlated regressors, and (b) to compare the performance of the proposed methods with existing methods.

We examine the case when the dimension of Ws is large relative to the sample size, and only a few of them play important roles in explaining endogeneity. Specifically, we generate the sample size N = 1000 and the dimension of W, Nw = 600. The specific data-generating process (DGP) is summarized below:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{p\xi} \\ \rho_{pw} & 1 & 0 \\ \rho_{p\xi} & 0 & 1 \end{bmatrix} \right) \tag{3.14}$$

$$\xi_t = G^{-1}(U_{\xi,t}) = G^{-1}(\Phi(\xi_t^*)) = \Phi^{-1}(\Phi(\xi^*)) = 1 \cdot \xi_t^*, \tag{3.15}$$

$$P_t = H^{-1}(U_{P,t}) = H^{-1}(\Phi(P_t^*)), \quad W_t = L^{-1}(U_{W,t}) = L^{-1}(\Phi(W_t^*)), \tag{3.16}$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + \xi_t = 1 + 1 \cdot P_t + (-1) \cdot W_t + \xi_t. \tag{3.17}$$

where ξ_t^* and P_t^* are correlated with the correlation coefficient $\rho_{p\xi} = 0.2$, and thus ξ_t and P_t are correlated, generating the endogeneity problem. W_t^*s are exogenous and are not correlated with ξ_t^*. W_t^* and P_t^* are set to be correlated with different correlation levels, which determines how important each W^* is in explaining the endogenous P. Specifically, we set two different levels of the correlation, $\rho_{pw} \in \{0.1, 0.3\}$. Among all the possible W^*s, only 13 are important ones that are correlated with P^* with non-zero ρ_{pw}, among which three of them are the most important ones with correlation $\rho_{pw} = 0.3$. For simplicity, we assume that there is no correlation among W^*s. We consider six different estimation methods: (i) OLS, (ii) 2sCOPE with all W_t^*s included in the first stage (2sCOPE$_{All}$), (iii) 2sCOPE with selected W_t^*s using Lasso with penalty parameter $\lambda = \lambda_{min}$ (Lasso), (iv) 2sCOPE with selected W_t^*s using post-Lasso (post-Lasso), (v) 2sCOPE with selected W_t^*s using De-biased Lasso (De-biased Lasso), and (vi) 2sCOPE with optimal W_t^*s (Golden).

We generate 1000 data sets as replicates using the DGP above. In the simulation, we use the gamma distribution $Gamma(1,1)$ with shape and rate equal to 1 for P_t and the exponential distribution $Exp(1)$ with rate 1 for W_t. Models are estimated on all generated data sets, providing the empirical distributions of the parameter estimates.

Table 3.1 reports estimation results. In the table, we only list the first three $(\beta_1, \beta_2, \beta_3)$ of all the 600 W_t coefficients to save space. As expected, OLS estimates of all μ, α and β are biased as a result of the regressor endogeneity. However, unexpectedly, the estimation result of 2sCOPE$_{All}$ has a significant bias, especially for the estimated coefficient of P_t (0.118 away from the true value). This tells us that, for finite sample estimation, including too many irrelevant or unimportant variables in the first stage of the model would make the estimation

		OLS		2sCOPE$_{All}$		Lasso		post-Lasso		De-biased Lasso		Golden	
Parameters	True	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
μ	1	1.166	1.099	1.088	1.105	1.187	1.085	1.061	1.072	1.045	1.078	1.013	1.078
α	-1	-0.749	0.060	-0.882	0.082	-1.019	0.105	-1.006	0.104	-0.981	0.100	-1.009	0.102
β_1	1	0.931	0.049	0.966	0.051	0.988	0.054	1.002	0.052	0.992	0.054	0.999	0.055
β_2	1	0.934	0.050	0.968	0.050	0.990	0.052	0.996	0.049	0.994	0.053	1.001	0.053
β_3	1	0.933	0.052	0.967	0.054	0.989	0.055	0.995	0.050	0.993	0.056	1.000	0.056
MSE(α)		0.0668		0.0207		0.0114		0.0108		0.0104		0.0105	
MSE(ALL)		0.00420		0.00399		0.00385		0.00384		0.00383		0.00380	

Table 3.1: Simulation Results of Lasso-based 2sCOPE

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

biased. This actually demonstrates the importance of using feature selection in the first stage to get unbiased causal estimates in the high-dimensional case. We next examine some lasso-based machine learning methods. All three types of lasso methods (Lasso, post-Lasso and De-biased Lasso) can significantly reduce estimation biases. Moreover, post-Lasso can reduce more bias and get better performance in estimation accuracy than Lasso. The estimated coefficient α using post-Lasso is -1.006, very close to the true value. The mean squared error (MSE) is 0.0108, which is 47.8% smaller than the MSE using 2sCOPE$_{All}$. It means that using post-Lasso in the first stage can improve estimation efficiency by around 47.8%, compared with the naive approach of including all variables (2sCOPE$_{All}$). De-biased Lasso shows comparable performance with post-Lasso. The estimated coefficient α using De-biased Lasso is -0.981, and the mean squared error (MSE) is 0.0104, which is 49.8% smaller than the MSE using 2sCOPE$_{All}$.

3.5 Empirical Application

In this section, we apply our method to a real data application. We are interested in how COVID-19 changes individual life. Specifically, we want to examine how governments' policies to contain the spread of COVID-19 affect citizens' happiness. We use the Stringency Index [2] constructed by Oxford Coronavirus Government Response Tracker (Ox-CGRT) to measure the strictness of policy. The COVID-19-related data is from the *Our World in Data* team and the *Covid-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)* [3]. The data contains daily confirmed COVID-19 cases, daily stringency level, etc. We further collect happiness data from the World Happiness Report [4], and country-level characteristics from the World Bank [5] and Fraser Institute [6]. Country-level characteristics include economy (GDP, inflation), population, gender, labor market-related, governance-related, legal system characteristics and so on.

Government policies in response to COVID-19 can be endogenous. The endogeneity can come from unmeasured country characteristics such as ideology, democracy, and freedom that can influence both people's happiness and governments' decisions. Since these variables are unobserved by researchers, they are absorbed into the structural error, leading to the endogeneity problem. Besides the endogenous stringency index, many country-level exogenous variables are included. In total, there are 41 country characteristics and 102 country observations. Table 3.2 summarizes the statistics of several key variables.

[2]The nine metrics used to calculate the Stringency Index are: school closures; workplace closures; cancellation of public events; restrictions on public gatherings; closures of public transport; stay-at-home requirements; public information campaigns; restrictions on internal movements; and international travel controls. https://ourworldindata.org/covid-stringency-index

[3]COVID-19 data: https://github.com/owid/covid-19-data/tree/master/public/data.

[4]https://worldhappiness.report;

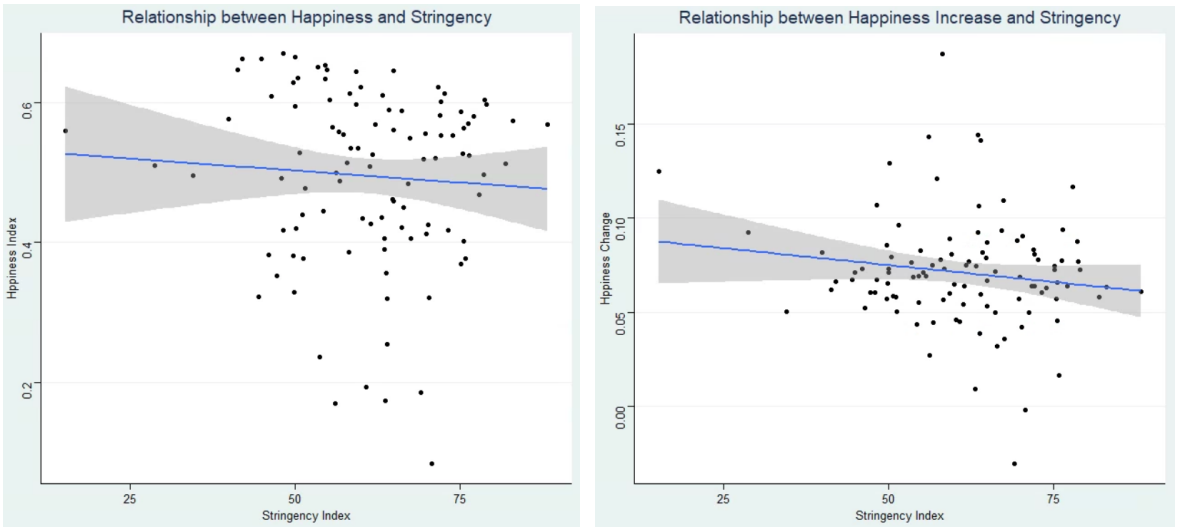
[5]https://databank.worldbank.org/source/world-development-indicators

[6]https://www.fraserinstitute.org/economic-freedom/map?geozone=worldpage=mapyear=2019.

Variables	Mean	SD	Max	Min
Happiness	0.49	0.13	0.67	0.08
HappinessGrowth	0.07	0.03	0.19	-0.03
Stringency/10	6.12	1.23	8.84	1.51
LogCases/million	8.87	1.83	11.01	2.11
LogPopuden	4.34	1.42	8.98	0.68
LogGDP	9.66	1.09	11.50	7.17
LogGDP Growth	-0.04	0.04	0.06	-0.23

Table 3.2: Summary Statistics

Figure 3.1 shows data evidence of the relationship between happiness and stringency level. It shows that both the happiness and the happiness growth in the year 2020 are decreasing with the stringency level. The stricter the policy in response to COVID-19, the less happy people will be.



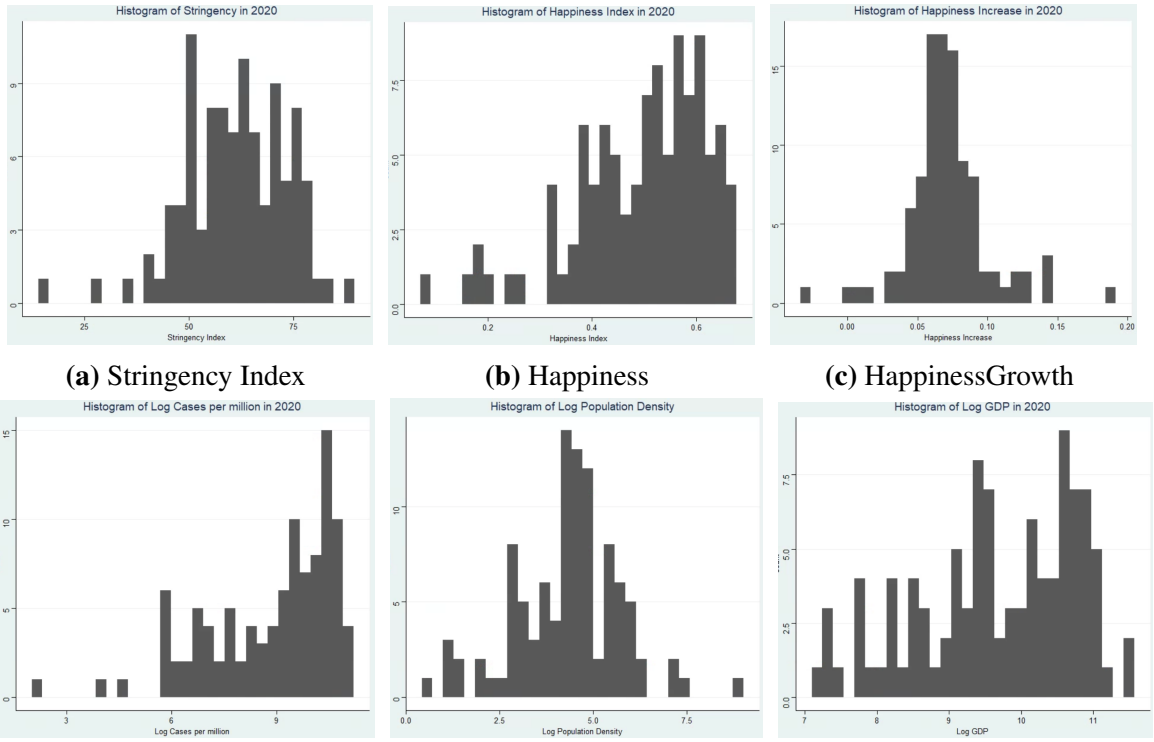
(a) Happiness **(b)** Happiness Growth

Figure 3.1: Relationship Between Happiness and Stringency Index.

We then consider the following linear regression model:

$$\text{HappyGrowth}_i = \beta_0 + \text{StringencyChange}_i \cdot \beta_1 + W_i'\beta_2 + \xi_i, \tag{3.18}$$

where $i = 1, 2, ..., I$ indexes country. W_i includes all country characteristics, such as change of confirmed COVID-19 cases, GDP growth, population, freedom and legal-system-related characteristics in this application. We are interested in examining how the stringency change during the COVID-19 period and country characteristics affect people's happiness growth.



(a) Stringency Index **(b)** Happiness **(c)** HappinessGrowth

(d) LogCases permillion **(e)** Log Populationdensity **(f)** Log GDP

Figure 3.2: Histogram of Stringency, Happiness, and Country Characteristics in 2020

Figure 3.2 shows the histograms of the Stringency index and some country-level characteristics. All the variables are continuous. Different country characteristics can have different correlations with the endogenous stringency index, which means they might have different importance levels in explaining the endogenous variable. For example, the correlation between stringency and log number of confirmed COVID-19 cases/million is 0.24

(p-value = 0.017). The correlation between stringency and logGDP growth is -0.27 (p-value = 0.006). Both the correlations are significantly different from zero. These are two characteristics that have high correlations with the endogenous variable, and many characteristics don't have high pairwise correlations with the endogenous stringency. Importantly, whether a variable is important in explaining the endogenous variable is hard to determine. It is complicatedly affected by all variables together, and the simple pairwise correlation may not fully measure the importance. That's why we use the lasso-based methods to fully measure and select important features in explaining the endogenous stringency in the first stage.

The many exogenous variables and the different correlations between the endogenous variable and exogenous variables provide a good setting for examining the performance of our proposed lasso-based 2sCOPE method. We estimate the linear regression model in Equation (3.18) using the OLS, the 2sCOPE method with all exogenous variables included in the first stage (2sCOPE$_{All}$), the 2sCOPE method using Lasso in the first stage (Lasso), and the 2sCOPE method using De-biased Lasso in the first stage (De-biased Lasso).

	OLS			2sCOPE$_{All}$			Lasso			post-Lasso			De-biased Lasso		
Parameters	Est	SE	t-value	Est	SE	t-value	Est	SE	t-value	Est	SE	t-value	Est	SE	t-value
Constant	0.226	0.153	1.479	0.219	0.158	1.385	0.266	0.167	1.597	0.249	0.157	1.590	0.242	0.153	1.581
Stringency	-0.025	0.010	-2.394	-0.023	0.012	-1.866	-0.031	0.014	-2.152	-0.031	0.013	-2.331	-0.034	0.014	-2.494
LogCases	-0.009	0.007	-1.314	-0.009	0.007	-1.304	-0.007	0.008	-0.926	-0.008	0.007	-1.018	-0.007	0.007	-0.987
LogGDP	0.207	0.067	3.109	0.208	0.067	3.090	0.200	0.068	2.943	0.198	0.068	2.905	0.200	0.067	2.995
ρ				-0.0207			0.378			0.258			0.465		

Table 3.3: Estimation Results: How COVID-19 affects Happiness

Table 3.3 reports the estimation results. The estimated OLS coefficient of the Stringency is -0.025, which means that increasing Stringency by 10 units would decrease happiness growth by 0.025. The estimated coefficient is significantly different from zero. Look at the estimates of 2sCOPE$_{All}$, the 2sCOPE method with all the exogenous variables included in

the first stage, the stringency coefficient is -0.023, which is very close to the OLS estimate and becomes even less significant. This can result from adding too many irrelevant variables in the first stage. When using Lasso-based methods to select important variables in the first stage, the estimated coefficients of stringency are -0.031 for Lasso, -0.031 for post-Lasso, and -0.034 for De-biased Lasso, all significantly different from zero. Those estimates are quite close, indicating the robustness of different Lasso-based methods in obtaining causal inference using 2sCOPE under this specific setting. Overall speaking, using Lasso-based methods to select important features in the first stage changes the stringency coefficient estimate from -0.023 to up to -0.034, implying that there is a positive correlation between unobserved country characteristics and stringency. For example, unobserved ideological characteristics such as conservative level can be positively correlated with both stringency and happiness. The positive estimates of ρ in the three Lasso-based 2sCOPE methods also confirm the positive correlation.

In sum, including too many irrelevant variables in the first stage in the high-dimensional case would weaken the power of 2sCOPE in correcting endogeneity and achieving accurate causal inference. After the feature selection procedure using Lasso-based methods in the first stage of 2sCOPE, we obtain a more negative and significant coefficient of stringency than using 2sCOPE with all observed exogenous variables included in the first stage. The estimated coefficient is -0.034, which means that increasing Stringency by 10 units would decrease happiness growth by 0.034. Moreover, the estimated coefficients are quite close using different Lasso-based methods, showing the robustness of the estimates.

3.6 Conclusion

Causal inference lies at the center of social science research, but the more and more common high-dimensional data in this big data era might make the traditional causal inference

methods fail to work. On the one hand, too many (irrelevant) variables added might bring finite-sample bias, reduce estimation efficiency, or even make the traditional regression infeasible when the dimension of variables is larger than the sample size. That's why feature selection machine learning methods are needed for causal inference methods. On the other hand, the more complicated structure of causal inference methods (e.g., two-stage estimation) other than a single-equation regression provides a distinct setting and a different perspective for feature selection methods. Adaption of traditional causal inference methods for high-dimensional data is much needed.

In this paper, we are interested in the copula instrument-free causal inference method for correcting endogeneity, and adapting the method to a high-dimensional setting. The copula method, first proposed by Park and Gupta (2012), has aroused wide interest because of its feasibility that no instruments are needed. After that, many empirical researchers have applied the copula method for real-data analysis (Burmester et al. 2015, Datta et al. 2015, Gruner et al. 2019, Keller et al. 2019, Bombaij and Dekimpe 2020, Guitart et al. 2018, Lamey et al. 2018, Wetzel et al. 2018, Heitmann et al. 2020, Atefi et al. 2018, Elshiewy and Boztug 2018). We consider a two-stage copula method, called 2sCOPE in Yang et al. (2022), that extends the original one to a much more general setting, and adapt it to the high-dimensional setting. Specifically, 2sCOPE corrects endogeneity by adding a generated regressor, which is the first-stage residual, to the outcome regression, just like what the control function approach does. By controlling the residual estimated in the first stage, endogeneity can be corrected. Thus, the estimation of the first-stage residual plays a central role. We propose a method combining the generalized two-stage copula method (2sCOPE) with some lasso-based methods (lasso, post-lasso, de-biased lasso) in selecting important variables in the first stage. We call it the lasso-based 2sCOPE method.

We conduct simulation studies and use an empirical real-data application to empirically verify the performance of our proposed method. The simulation results show that 2sCOPE

without feature selection suffers the finite-sample bias when the dimension is relatively large. Moreover, the lasso-based 2sCOPE method can substantially improve estimation accuracy and efficiency, by around 50%, as compared with the 2sCOPE without feature selection. We further apply our method to a current data set about COVID-19, and examine the effect of governments' policy strictness in response to COVID-19 on citizens' happiness. We use country-level cross-sectional data. The dimension of country characteristics is relatively large, around half of the sample size. The result shows that the estimated coefficient of the endogenous policy stringency level from the 2sCOPE with all controls included in the first stage is very close to the OLS estimate and becomes even insignificant. In contrast, our proposed lasso-based 2sCOPE method makes the effect of policy strictness on happiness more significantly negative. Moreover, the estimates using different lasso-based methods are quite robust in this application. This application result confirms the findings in the simulation study. Under the lasso-based 2sCOPE estimation, the application tells us that policy strictness has a significant negative effect on people's happiness during the COVID-19 period. Increasing policy strictness by 10 units would decrease happiness growth by 0.034, which is more than one standard deviation of happiness growth.

In the above analysis, we only consider using feature selection methods in the first stage of 2sCOPE, which means that we assume what variables are going to have effects on the main outcome are known. It is worth exploring and incorporating cases where one has less certainty on important controls in the outcome regression. Our next step is to further explore how to apply feature selection methods in the main regression model as well, which we label as double lasso-based 2sCOPE method for correcting endogeneity. Moreover, for the current simulation analysis on 2sCOPE combined with the lasso-based methods, we only conduct simulation studies for one single setting. In reality, lasso can be sensitive to variations in real data. For future research, we will conduct more simulation studies in terms of parameter variation, and theoretically prove what are the requirements for our method to

work.

Chapter 4

Vertical Differentiation in Two-sided Markets: Evidence from A Ride-hailing Platform

4.1 Introduction

Two-sided (or, more generally, multi-sided) markets are roughly defined as markets in which one or several platforms enable interactions between end-users and try to get the two (or multiple) sides "on board" by appropriately charging each side (Rochet and Tirole (2006)). Leading examples of such platforms include Uber and Lyft for rides; Airbnb for accommodation; and eBay and Taobao for E-commerce. Statistics show that consumers spend $65 billion on Uber [1], and $3.5 trillion on goods on E-commerce sites [2] in 2019. Given the growing economic significance of such platforms, it is important to understand the two-sided markets—what makes them different from traditional markets, and whether traditional

[1] https://www.businessofapps.com/data/uber-statistics/

[2] https://www.statista.com/topics/871/online-shopping/

business strategies are still applicable to them.

One distinct feature of two-sided markets is network externalities, which can exist both within and across sides. The demand (supply) of a product is dependent on the demand (supply) from other users considering that product. In addition, demand will attract more service providers (supply) to join the platform, and meanwhile, more supply will further make the platform more attractive to consumers. Since users' utilities from the product depend on other users' decisions, network externalities create and provide users with another product quality dimension, which I call the network value. That is, in markets where products are subject to network externalities, the number of users also determines, at least partially, the perceived quality of the product. For example, users of a ride-sharing platform value short waiting time, which comes from a large scale of drivers available on the platform, whereas drivers value short cruising time, which results from a large demand base from the rider side.

Vertical differentiation is a business strategy commonly used by firms in both conventional markets and two-sided markets. By designing products with different qualities, a firm can attract and satisfy consumers with heterogeneous preferences and thus expand the market. For example, Uber designs Uber Black to target high-end customers and UberX to target low-end customers. Several studies have examined whether price discrimination is profitable for a monopoly. For example, Gabszewicz et al. (1986), Maskin and Riley (1984), Mussa and Rosen (1978), Salant (1989) and Stokey (1979) discussed vertical differentiation in a one-sided market scenario. In particular, Salant (1989) and Stokey (1979) identify conditions under which the monopolist finds it optimal to price discriminate. These papers are developed in the context of conventional markets without network externalities. Jing (2007) further investigates how network externalities affect a firm's decision to price discriminate in a one-sided market. They find that network externality is a factor that makes firms favor extending a product line. Moreover, they find that the monopolist employs two

qualities of products for rather different purposes: the low-end product is used mainly to expand its network size, and the high-end product is its primary source of profits. However, the paper only considers vertical differentiation with network externalities in one-side markets. In a two-sided market, both consumers (demand side) and service providers (supply side) are affected by network externalities. Moreover, the network externalities can exist both within and across sides. All these make vertical differentiation more complicated in two-sided markets.

In the two-sided markets, it is unclear whether network externalities still make firms favor extending a product line. On the one hand, according to the literature on one-sided markets, firms can be better off by designing vertically differentiated products to expand the market, especially when network externalities exist. On the other hand, offering vertically differentiated products in a two-sided market might be less optimal than offering a homogeneous product if the further segmented demand and supply limit positive network effects. Firms might want to unite markets from different products into one scaled-size market to achieve large positive network effects. Thus, there is a tradeoff when using vertical differentiation in two-sided markets. The main goal of this paper is to examine whether vertical differentiation is better than a homogeneous product design in two-sided markets. This paper contributes to the literature by being one of the first to empirically examine vertical differentiation in two-sided markets.

I use a distinct dataset from a leading ride-hailing company located in New York City to empirically analyze and answer the research question above. On this platform, there are two types of differentiated products with different qualities of car make and service quality, one is called premium and the other is called standard. The data is from January 2016 to May 2016, and include all individual order requests from riders and drivers' corresponding responses about whether to pick up the riders. I build up a structural model to simultaneously estimate demand and supply, and use the Bayesian Markov chain Monte Carlo

(MCMC) method for model estimation. The results show that both intrinsic product quality and network value are significant components for riders' and drivers' decisions, and thus for determining the degree of vertical differentiation. I also conduct counterfactual analyses based on the estimation results. In counterfactual analysis, I evaluate alternative strategies by using the estimated parameters to solve the fixed points and obtain the equilibrium in the new setting. Specifically, I compare the current vertical differentiation case with a homogeneous product case. The result shows that using vertical differentiation under the current pricing strategy is better than a standard-only homogeneous product in terms of both market size expansion and profit maximization, which means that the positive effect of market expansion from vertical differentiation can offset the loss from network segmentation. By comparing different pricing strategies under the same vertical differentiation setting, I find that the roles of the two vertically differentiated products are different. Premium product is more profitable, but also costs the firm more to maintain the network because of the smaller network size (value). Low-end product is not profitable, but is valuable to the firm in enlarging user size. In a word, network externalities play important roles in two-sided markets, and two-sided platforms should understand users' preferences over both the network value and the product value when determining the optimal product strategy and pricing strategies.

The remainder of the paper is organized as follows. In section 4.2, I reviewed the marketing and economics literature relevant to vertical differentiation and ride-hailing platforms. Section 4.3 describes the data used in the empirical implementation and presents model-free evidence of network externalities and how network externalities influence riders' and drivers' behavior. Section 4.4 describes the model and estimation method used for estimating the simultaneous demand and supply model and quantifying network externalities. Section 4.5 presents the detailed estimation results. Section 4.6 further conducts counterfactual analysis and provides managerial implications. Section 4.7 concludes the paper with key results and suggestions for future work.

4.2 Literature Review

This paper is broadly related to three streams of literature. I first discuss the general vertical differentiation literature, then link vertical differentiation to two-sided markets, and discuss the relevant work on the ride-hailing industry in the end.

Vertical differentiation research has attracted considerable attention in both economics and marketing (Lancaster (1990), Mussa and Rosen (1978), Shaked and Sutton (1982), Sridhar Moorthy (1984)). In a vertically differentiated product space, more attributes will attract consumers with different preferences, who may have different willingness to pay. Vertical differentiation has already been widely discussed in one-sided markets. For example, Mussa and Rosen (1978), Sridhar Moorthy (1984), Salant (1989) and Stokey (1979) studied vertical differentiation for monopolists. In particular, Salant (1989) and Stokey (1979) identified conditions under which the monopolist finds it optimal to price discriminate. Shaked and Sutton (1982) built up a theoretical model for vertical differentiation in a competitive setting. These papers focused on one-dimensional product differentiation, many papers also discussed multidimensional product differentiation. For example, Vandenbosch and Weinberg (1995) studied product and price competition in a two-dimensional vertical differentiation model, whereas product qualities are differentiated into two product attributes.

All those papers are developed in the context of conventional markets without network externalities. Jing (2007) further investigated how network externalities affect a firm's decision to price discriminate in a one-sided market. They find network externalities are a factor that favors extending a product line. Specifically, they find that monopolists employ the two qualities for rather different purposes: The low-end product is used mainly to expand its network size, which is why the monopolist would lower the price of the low-end, and the high-end product is its primary source of profits with a rather higher price than in

conventional markets. However, the paper only considers the effect of network externalities in one-side markets. That is, they assume network externalities only affect the demand side and assume supply is sufficient.

Recently, two-sided markets have raised wide concern in academia. In a two-sided market setting, not only consumers are affected by network externalities, but service providers (supply) are also affected. The within- and cross-network externalities would make vertical differentiation more complicated. Only a few papers in the literature have studied vertical differentiation in two-sided markets. For example, Liu and Serfes (2013) theoretically compared the effect of vertical differentiation on competition in two-sided markets with that in a one-sided market. They find that price discrimination in a two-sided market may actually soften competition. Gabszewicz and Wauthy (2014) developed a theoretical model and demonstrated that a unique vertical differentiated pricing equilibrium exists for two firms competing in a two-sided market. They also assume that products are vertically differentiated only by network size. While in this paper, I focus on vertical differentiation within a platform and products are vertically differentiated not only by network value but also by product intrinsic quality. Lin (2020) studied vertical differentiation in a monopoly case for media platforms. These papers theoretically study vertical differentiation in two-sided markets, while this paper studies vertical differentiation using empirical analysis within a two-sided platform. There is also some empirical work in two-sided markets. For example, Zervas et al. (2017) and Li and Srinivasan (2019) studied the impact of the two-sided platform Airbnb's entry on hotels. However, they didn't study vertical differentiation. This paper is one of the first papers to empirically analyze vertical differentiation in the two-sided market context.

With the increasing scale of ride-hailing platforms, it has attracted wide attention from researchers. Most papers in ride-hailing literature focus on driver-side behavior. They study the effect of the new ride-hailing technology on drivers. For example, Chen et al. (2019)

studied the value of ride-hailing platforms in creating flexible work for drivers. Wang et al. (2019) studied the impact of mobile hailing technology on taxi driving behaviors. Frechette et al. (2019) studied matching frictions and technologies in the taxi industry. Guda and Subramanian (2019) studied the effect of surge pricing on worker incentives with flexible work. Few papers have studied vertical differentiation. Lin (2020) studied vertical differentiation in a monopoly case for media platforms. They theoretically develop a two-sided media model and examine how a monopoly ad-financed media can price discriminate through versioning using membership-based pricing. While our paper empirically studies usage-based pricing strategy for a non-monopoly firm. Another related paper is Bryan and Gans (2019). It theoretically studied vertical differentiation in a ride-hailing setting. But they assume that the number of drivers is endogenously determined by the platform. In this way, they only model how the network size of drivers would influence riders through waiting time, but didn't consider the effect of rider size on drivers' decisions. While this paper models demand and supply simultaneously. Moreover, they assume products are vertically differentiated only in network size, while I consider the case when products are vertically differentiated in two quality dimensions, product intrinsic quality and network value.

In a word, the contribution of this paper to the literature on vertical differentiation in two-sided markets is mainly twofold. First, I study two-dimension vertical differentiation in two-sided markets. That is, products are assumed to be vertically differentiated in two dimensions, product intrinsic quality and network value, and both are important in determining the level of vertical differentiation. Second, this paper is one of the first papers to empirically study vertical differentiation in two-sided markets and quantify network externalities by modeling demand and supply simultaneously.

4.3 Data and Model-Free Evidence

In this section, I will introduce the data, and provide data patterns and evidence about why network externalities are important in riders' and drivers' choices towards vertically differentiated products.

4.3.1 Data Background

The global ride-hailing market is valued at USD 113 billion in 2020. Overall, Didi, Uber, Lyft, and Grab are major players who have prominent market share globally. UBER and Lyft have a prominent share in the North American region, whereas, in China, Didi Chuxing Technology Co [3]. contributes to the high market share.

The focal platform under study is a leading on-demand ride-hailing company located in New York City, where the major competitors are yellow/green taxi, Uber and Lyft. Riders on this platform order a car by using the company's location-based smartphone app. They can place order requests at any time and anywhere they want. Once a rider placed an order request, drivers nearby who are active on the platform can receive the offer notifications from the platform, and then decide whether to accept the offer.

This platform provides two main ride types that riders can choose from, which are called standard and premium. Standard type serves affordable everyday rides and premium type serves a bit more expensive high-end rides, which are similar to UberX and UberBlack. Riders are charged fees for each realized trip, and drivers earn income from trip fares. The fares for the two types of products are different for both riders and drivers. Note that the fare structure for drivers varies in peak and off-peak time, and also changes once during the time window, while remaining the same for riders over time. Tables 4.1 and 4.2 show the

[3]https://www.mordorintelligence.com/industry-reports/ride-hailing-market

detailed fare structure for rider and driver side separately. On the rider side, the standard type charges a fixed fee of $8.98 for each trip, and the premium type charges a two-part tariff consisting of a one-time fixed fee and a duration-based fee. The rate is $8 for the first 15 minutes, $0.5 per minute between 15-29 minutes, $0.75 per minute above 30 minutes, and $1 per minute above 40 minutes.

Table 4.1: Price Structure for Rider

Time	Price Structure	Standard	Premium
Jan – May 2016	Base fee	$ 8.98	$8
	Per additional minute (15-29 mins)	0	$0.5
	Per additional minute (30-39 mins)	0	$0.75
	Per additional minute ($\geq$ 40 mins)	0	$1

On the driver side, earnings all follow a two-part tariff, and a significant fare change was implemented in February 2016. Before February 14, 2016, drivers earn the same fixed fee, $8, for both types, while the number of minutes to earn the base fee is shorter ($13 < 15$) and the per-minute payment for an additional minute is higher for the premium type ($\$0.62 > \$0.55$). Throughout the whole period, there was a surcharge of $0.75 for premium and $0.9 for standard in peak times during 7:00 a.m-9:00 a.m, 6:00 p.m.- 8:00 p.m., and night hours between 8:00 p.m. and 3:00 a.m on weekends. After February, the platform changes the fixed fee to $7 and also lowers per-minute earnings for both types of products in both peak and off-peak hours. The above shows the fare structures for rider and driver, and the platform further takes a fixed rate of fares as commission from drivers. The commission rate of Uber is 25% while that is 20% on this platform. All order requests from riders and all the drivers' corresponding responses in the Manhattan area from January 9 to May 15, 2016, a total of 18 weeks, are observed. The raw data comprises 98,342 riders who request at least one ride and 3,509 drivers available during the period. On average, each rider requested 8 times, for a total of 760,298 requests from all riders.

Table 4.2: Fare Income Structure for Driver

		Standard		Premium	
		off-peak	peak	off-peak	peak
Before Feb. 14, 2016	Base fee ($)	8			
	Minutes to earn minimum fare (min)	15	11	13	9
	Per additional minute ($)	0.55	0.75	0.62	0.9
After Feb. 14, 2016	Base fee ($)	7			
	Minutes to earn minimum fare (min)	13	13	14	11
	Per additional minute ($)	0.54	0.6	0.52	0.72

4.3.2 Variables Related to Network Externalities

I first clean the raw data, and construct key variables that are related to network externalities in both rider (demand) and driver (supply) sides for further analysis.

Rider's Expected Waiting Time

On the rider side, the rider's waiting time for each type of product, a direct indicator of network effect, is a key factor to affect their choices. In this ride-hailing platform, though the time riders have to wait is unobserved, it's reasonable to assume that they can form expectations on the waiting time because they can observe the real traffic environment. Below shows the procedure of how I construct the rider's expected waiting time for each type.

- Define and calculate the rider's waiting time of each trip as the total time from order creation to estimated driver's arrival, both of which are observed by researchers;
- Divide the total area into smaller markets for constructing expected waiting time at a specific time and location. Specifically, I first divide the Manhattan area into 29 smaller locations, according to a common demarcation of Manhattan[4], which is visualized in Figure 4.1. Then define a market at a week-hour-location level. That is, order

[4]https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas-NTA-/cpf4-rkhq

requests at a specific location in a specific hour and week are in the same market. In the end, there are 5,672 markets in total.

- After the market is defined, I construct the rider's ex-ante expected waiting time for each type by averaging the waiting time of all same-type trips in the same market.

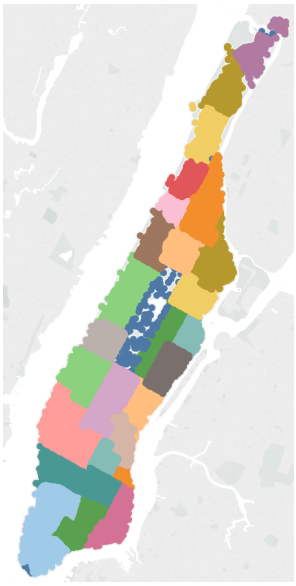


Figure 4.1: Manhattan Neighborhoods.

Driver's Cruising Time

On the driver side, cruising time for a trip is essential for their decisions on whether to take the trip. Cruising time is defined as the estimated time to pick up the rider once an offer is received. The driver's cruising time is not observed directly by the driver, but the rider's location and the distance from the rider at the time the driver receives an offer are observed. And researchers can observe the cruising time for drivers to pick up riders for realized trips. We assume drivers can approximate the cruise time, and fill in the missing cruising time for unrealized trips by using the observed distance from riders of all trips and the observed cruising time for realized trips (see more details in Appendix). It's reasonable to approximate the cruising time in this way because drivers must be familiar with the geographic

information once the location of the rider is revealed. The reason I want to construct cruising time instead of using distance alone is two folds. First, drivers care about both distance and time spent. Second, distance itself cannot fully reveal the dynamic traffic environment, while cruising time can because traffic environment can be different in different times and areas because of the real-time network effect. For example, two trips with the same distance from riders, one in rush hour while the other not, can have exact different cruising times.

Table 4.3: Summary statistics

	Variables	Unit	Mean	SD	Max	Min
Premium Type	Trip fare	$	10.76	4.15	24.44	5.38
	Rider waiting time	min.	7.54	4.47	24.37	0.20
	Driver wage	$	13.6	7.68	62.8	7.0
	Driver wage (no cruise)	$/min	0.73	0.25	3	0.28
	Driver wage (incl cruise)	$/min	0.54	0.19	4.57	0.10
	Trip duration	min.	18.8	9.0	69.6	4.2
	Driver cruise time	min.	5.81	2.60	20.35	0.40
	Demand size	Units	4.98	4.78	50.5	1.0
	Supply size	Units	3.30	2.98	29.75	0.25
Standard Type	Trip fare	$	8.98	0.23	24.00	8.00
	Rider waiting time	min.	7.38	4.35	24.37	0.20
	Driver wage	$	14.4	7.03	48.5	7.0
	Driver wage (no cruise)	$/min	0.68	0.23	3.7	0.39
	Driver wage (incl cruise)	$/min	0.53	0.18	5.15	0.10
	Trip duration	min.	21.4	9.74	69.8	4.2
	Driver cruise time	min.	5.44	2.09	19.50	0.32
	Demand size	Units	8.28	7.04	55.75	1.0
	Supply size	Units	5.72	5.07	31.0	0.25

After the construction of riders' expected waiting time and drivers' cruise time for each trip, I further calculate the network size of both demand and supply in each market to analyze network externalities. Here, I define demand as the number of total trip requests from riders, and measure supply using the number of trips that drivers accept. Until now, I have finished the construction of rider's expected waiting time, driver's expected cruising time, demand size and supply size for each type. I further clean the data by removing erroneous observations and outliers, and summarize the descriptive statistics of key variables in Ta-

ble 4.3. It shows that on average, the premium type has a higher trip fare with $10.76 than the standard type with $8.98 for riders, and provides a higher per-minute payment with $0.73 than the standard type with $0.68 for drivers. However, both riders' waiting time and drivers' cruising time are longer for the premium type. After accounting for cruising time, drivers' per-minute payments from the two types become quite close. The statistics provide some key factors that might influence drivers' and riders' choices towards the two types, which I will show detailed evidence in the next section. In the end, 304,044 trips with 63,260 available riders and 3,509 drivers on this platform are surviving for further analysis.

4.3.3 Data Evidence in Riders' Choice

In this section, I will show data evidence about how riders choose between the two types, and what factors would affect their choice decisions.

Figure 4.2 shows riders' aggregate choices. Around two-thirds of riders choose the standard type and one third choose the premium. What are the key factors or the differences between the two types that differentiate riders' preferences? What do high-end riders pay higher prices for?

First, the products' intrinsic qualities are different, which is the core to help differentiate in traditional one-sided markets. The quality difference between the two types mainly comes from car quality and service quality. Data shows that cars are more luxury and rating is higher for premium (see details about the quality difference between the two types in Appendix B.1). This quality difference provides different product intrinsic value to riders that help to differentiate riders with heterogeneous preferences [5].

Second, besides product intrinsic quality, waiting time is another important factor that

[5]Note that some proportion of standard drivers in this platform is allowed to serve both the premium and standard riders. No need to worry, as the market is still differentiated. See more details in Appendix B.1

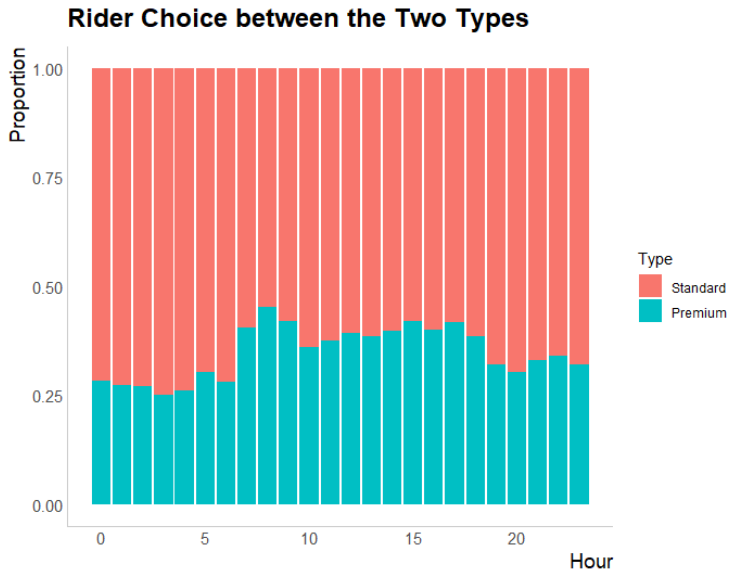


Figure 4.2: Choice Occasion between The Two Types.

plays an important role in riders' choice decisions. Figure 4.3 shows the relationship between market share and waiting time. The line with the negative slope indicates that the market share of the premium type decreases with the waiting time difference between premium and standard type. That is, the longer the waiting time for premium type compared with that for standard type, the smaller the market share of premium type would be. This model-free evidence demonstrates that riders prefer shorter waiting time, and waiting time is an important factor that affects riders' preferences over the two types of products. In a word, data evidence shows that product intrinsic quality and waiting time are two important attributes for product differentiation in two-sided markets.

4.3.4 Data Evidence in Drivers' Behavior

In this subsection, I further explore what would affect driver behavior.

First, drivers care about earnings from the received offer. Since time is money for

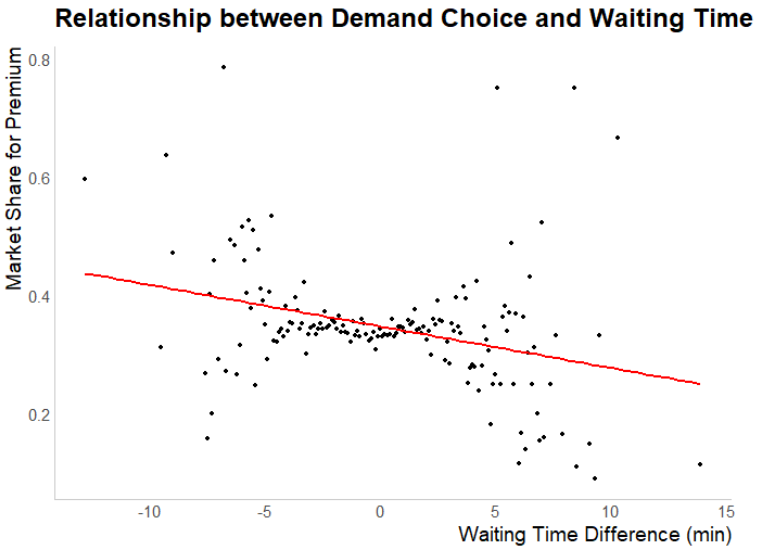
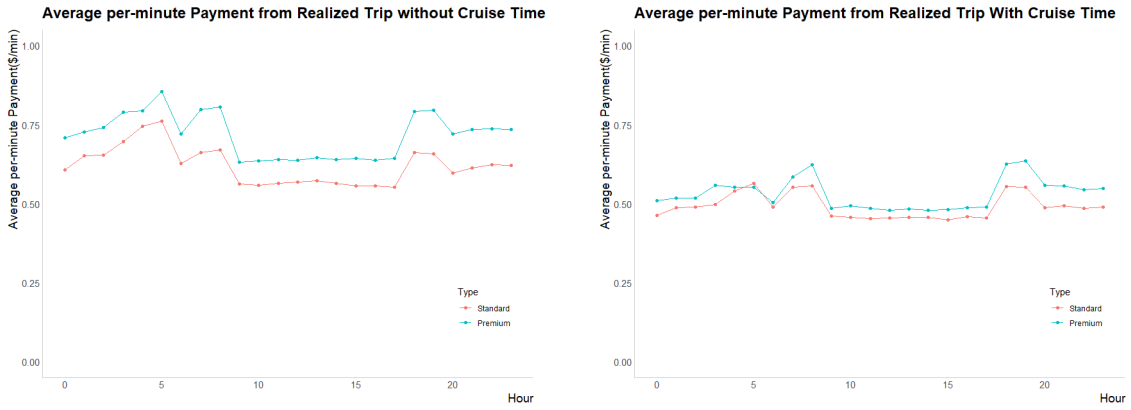


Figure 4.3: Relationship Between Market Share and Waiting Time.



(a) Payment without Cruise Time **(b)** Payment with Cruise Time

Figure 4.4: Drivers' Average per-minute Payment from Realized Trips.

drivers, per-minute payment should be a better measure for drivers' earnings instead of a total fare from a trip. Figure 4.4(a) shows the average per-minute payment drivers can get from the two types of products across hours. This per-minute payment is calculated by dividing drivers' earnings by the driving time with passengers. It shows that generally speaking, drivers receive higher payment from the premium than from the standard-type trips. However, not only does the driving time which creates value matter, but the cruising

time for drivers to pick up the rider also matters. Since drivers can observe the location of the rider, cruising time can be another key factor that would affect drivers' decision about whether to accept the current offer. Figure 4.4(b) shows drivers' per-minute payment after accounting for the cruising time to pick up riders. It shows that after accounting for cruising time, drivers' average per-minute payment decreases from around \$0.7 to \$0.5, a more than 20% decrease, which indicates that cruising time accounts for non-negligible importance. Moreover, drivers' per-minute payments from the two types become surprisingly close. This tells that though premium has higher payment, cruising time for drivers to fulfill premium is also longer. Thus, cruising time can be an important factor that would affect drivers' decisions. Figure 4.5 further shows the relationship between drivers' average acceptance rate within each hour and the cruising time in that hour. The negative slope tells us that drivers favor trips with shorter cruising times.

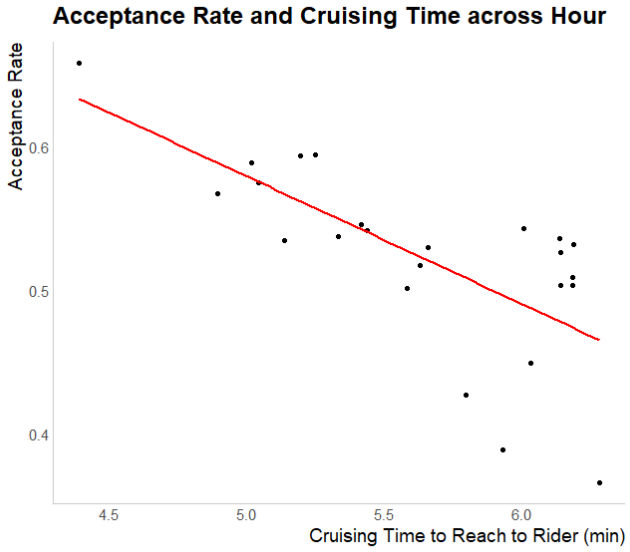


Figure 4.5: Relationship Between Drivers' Acceptance Rate and Cruise Time.

Second, besides the earning from the current offer, whether there exist other possible opportunities can also be an important factor that would affect drivers' decisions on whether to accept the current offer. Drivers can be picky about offers if there are many options for them. Figure 4.6 shows that drivers' acceptance rate of offers is negatively correlated with

the total demand size in that market. It tells us that once drivers have more options (demand), they will be more selective about offers and less likely to accept the current offer. This is actually one way how network externalities play roles in affecting drivers' decisions. A larger demand size provides more opportunities to drivers, which will further affect drivers' decisions.

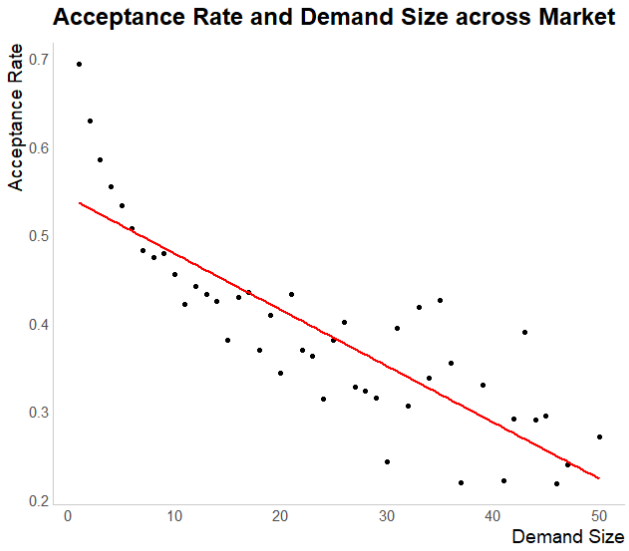


Figure 4.6: Relationship Between Drivers' Acceptance Rate and Demand Size.

4.4 Model

In this section, I develop a structural model to capture both riders' and drivers' choices simultaneously, together with the IV and control function approach to correct endogeneity, and use the Bayesian MCMC method for estimation.

4.4.1 Rider's Decision

An individual rider needs to decide which car type to choose before placing an order request. Premium type and the standard type are different in prices and some quality-level attributes. The demand function is specified as a multinomial logit model. Suppose requests from a group of riders ($i = 1,2,\ldots,I$) for k types of choices ($k = 1,2$) across m markets ($m = 1,2,\ldots,M$) are observed. The utility rider i receives from type k product at market m is specified as:

$$U_{ikm} = \beta_{0i}^d + \beta_{1i}^d P_{ikm} + \beta_{2i}^d Ewait_{km} + X_m^d \beta^d + \xi_{km}^d + \varepsilon_{ikm}, \tag{4.1}$$

In the case of choosing the outside option, I denote it as $k = 0$ and the associated utility function is:

$$U_{i0tm} = \varepsilon_{i0m}. \tag{4.2}$$

where P_{ikm} is the price charged for that ride, $Ewait_{km}$ is riders' expected waiting time for type k at market m, which is the key variable that indicates network externalities in the current market, and X_{km}^d is a vector of market-level exogenous controls for each type k. $Ewait_{km}$ depends on the current total demand and the total supply of drivers for that type in the same market.

I further model riders' expected waiting time as a function of current demand and supply.

$$\text{Ewait}_{km} = \gamma_{k0}^d + \gamma_{k1}^d \text{D}_{km} + \gamma_{k2}^d \text{S}_{km} + \varepsilon_{km}^d. \tag{4.3}$$

β_{0i}, β_{1i} and β_{2i} are individual-level response coefficients, and γ_{k0}, γ_{k1},γ_{k2} are market-level response coefficients. D_{km} and S_{km} are the logarithms of demand and supply size for type k of market m. ε_{ikm} and ε_{km} are independent individual-level and market-level unobserved error terms separately. I make assumptions on the error terms and response coefficients as

the following:

$$\varepsilon_{ikm} \sim \text{extreme value } (0,1), \tag{4.4}$$

$$\varepsilon_{km} \sim N(0,\sigma_1^2), \tag{4.5}$$

$$\theta_i = (\beta_{0i},\beta_{1i},\beta_i) \sim MVN(\bar{\theta}_i,\Sigma_{\theta i}). \tag{4.6}$$

According to Equations (4.1) and (4.3), I can easily find that riders' expected waiting time will affect their choice of which type of product to request, and meanwhile, the number of total requests for that type will in turn influence riders' waiting time for that type. Thus, this is a system of simultaneous equations. However, this system of simultaneous equations is different from traditional ones, as the demand is modeled non-linearly using a choice model. The type I extreme value specification of ε_{ikm} leads to a standard logit choice probability for rider i choosing type k at market m,

$$Pr(y_{ikm}=1) = \frac{\exp(\mathrm{u_{ikm}})}{1+\exp(\mathrm{u_{i1m}})+\exp(\mathrm{u_{i2m}})}, (k=1,2). \tag{4.7}$$

where

$$u_{ikm} = \beta_{0i}^d + \beta_{1i}^d P_{ikm} + \beta_{2i}^d Ewait_{km} + X_{tm}^d \beta^d + \xi_{km}^d. \tag{4.8}$$

and the predicted demand using this choice model will further influence $Ewait_{km}$. I later use 3SLS estimation and the control function approach to estimate this nonlinear system of simultaneous equations. After riders make their choice decisions, the platform will send those order requests to nearby drivers. Then drivers will make decisions on whether to accept the offer.

4.4.2 Driver's Decision

Once a driver receives an offer from a rider, some offer-specific information such as the type of the offer, price, and time they have to cruise around to pick up the rider (distance away from the rider) is revealed to the driver. He/she might also obtain some information about the current market, such as the expected demand and supply. Based on all those observed and inferred information, they make decisions on whether to accept the offer. I model drivers' behavior in the following way. After receiving an offer, a driver will decide whether to accept the offer by comparing the earning from the current trip with his reservation wage in the current market. He will accept the offer if and only if the per-minute payoff from the trip is larger than or equal to his reservation wage. That is, I assume the driver j will accept the offer f in market m if and only if his expected per-minute pay exceeds his reservation wage at market m.

$$Y_{jfm} = \begin{cases} 1 & \text{if Payment}_{jfm} \geq R_{jm}, \\ 0 & \text{otherwise.} \end{cases}$$

Then the utility of driver j by accepting the offer f in market m is specified as

$$V_{jfm} = \beta^s_{1j} \frac{\text{Wage}_{jfm}}{\text{RideDur}_{jfm} + \text{Cruise}_{jfm}} - R_{jm} + \varepsilon^s_{jfm}, \tag{4.9}$$

where Wage_{jfm} is the total payment from offer f, RideDur_{jfm} and Cruise_{jfm} are ride duration and cruising time of offer f. R_{jm} is individual and market-level reservation wage, which depends on the driver's specific preference and market environment, and indicates the possible opportunities for the driver j in the market m. The reservation wage of driver j at market m is modeled as

$$R_{jm} = \beta^s_{0j} + \beta^s_{2j} D_{1m} + \beta^s_{3j} S_{1m} + \beta^s_{4j} D_{2m} + \beta^s_{5j} S_{2m} + \beta^s_{6j} X^s_m + \xi^s_m + \varepsilon^s_{jm}. \tag{4.10}$$

where X_m includes all market-specific variables such as week, hour, and location dummies. D_{km}, S_{km} (k=1,2) are network sizes in rider and driver sides for each product separately. Here, I model that network externalities would affect drivers' decisions by also influencing their reservation wages (opportunities) in the focal market. This model setting is reasonable because, besides the payment from the current offer, the driver also cares about other possible opportunities in the current market.

I make similar assumptions on the error terms and response coefficients for drivers as the following:

$$\varepsilon^s_{jfm} \sim \text{extreme value } (0,1) \tag{4.11}$$

$$\varepsilon^s_m \sim N(0,\sigma_2^2) \tag{4.12}$$

$$\theta_j = (\beta^s_{0j}, \beta^s_{1j}, ..., \beta^s_{6j}) \sim MVN(\bar{\theta}_j, \Sigma_{\theta j}) \tag{4.13}$$

The type I extreme value specification of ε_{jfm} leads to a standard logit choice probability for driver j choosing to accept offer f at market m,

$$Pr(Y_{jfm} = 1) = \frac{\exp(v_{jfm})}{1+\exp(v_{jfm})}. \tag{4.14}$$

where

$$V_{jfm} = v_{jfm} + \varepsilon^s_{jfm}. \tag{4.15}$$

4.4.3 Hierarchical model

After modeling drivers' and riders' decisions, I further add a hierarchical layer to estimate individual parameters. The individual parameters, $(\theta_i \equiv (\beta^d_{0i}, \beta^d_{1i}, \beta^d_i), \theta_j \equiv (\beta^s_{0i}, \beta^s_{1i}, \beta^s_i))$,

are determined by a vector of individual-related W_i, W_j separately as follows:

$$\begin{aligned} \theta_i &= G^d W_i + \eta_i, \eta_i \sim N(0, \Sigma_{\theta i}), \\ \theta_j &= G^s W_j + \eta_j, \eta_j \sim N(0, \Sigma_{\theta j}). \end{aligned} \tag{4.16}$$

where $G^d(G^s)$ is a matrix of parameters that indicate how each demographic variable influences the individual parameter θ. $\Sigma_{\theta i}(\Sigma_{\theta j})$ is the variance-covariance matrix capturing the interdependence among $\theta_i(\theta_j)$. The $W_i(W_j)$ includes riders' (drivers') demographic variables that will influence their decision-making process. In the current estimation, I include intercepts only in $W_i(W_j)$.

After model setup, I next introduce the endogeneity issue and some exogenous variables in this setting that could be used as IVs to deal with the endogeneity of the simultaneous-equation system.

4.4.4 3SLS/2SLS and Control Function Approach

In this simultaneous-equation system, riders' expected waiting time and demand (supply) are endogenous variables. On the one hand, riders' expected waiting time for a specific type of product will affect their decisions of whether to choose that type. On the other hand, in this dynamic request and pick-up process, each rider's choice will influence supply and demand and in turn affect riders' waiting time. For example, suppose there's a sudden demand shock (e.g., a concert is just over), many people are requesting for rides and thus the waiting time in that place must be quite long because of no enough supply. This would cause the waiting time and demand to be positively correlated, giving us the wrong and confusing intuition that riders prefer long waiting time. Similar to the driver side. network size would simultaneously influence drivers' decisions.

I use the instrumental variable approach to correct the endogeneity of expected waiting time (expected cruising time) and demand (supply). Specifically, I use the expected waiting time in the last week at the same hour and location as an instrumental variable for endogenous waiting time. Last-period waiting time will not influence the current demand, but has a strong correlation with the waiting time in the current period. For the endogenous demand and supply in Equations (4.3) and (4.10), I use last-hour demand and supply in the same week at the same location as instrument variables. Another important IV for supply is the exogenous subsidy level, which is defined as the difference between driver's wage and rider's price. On this platform, riders' price structure remains the same over time, while drivers' wage structure changes a couple of times. This exogenous variation can help to correct endogeneity. To sum up, I use 3-stage (2-stage) lease squares to correct endogeneity on the rider (driver) side.

Stage 1: In the first stage, I use lag demand and lag supply as IV to correct endogeneity in waiting time and reservation wage.

$$\text{Demand}_{km_{(h)}} = \delta^d_{k0} + \delta^d_{k1}\text{Demand}_{km_{(h-1)}} + \varepsilon^d_{km}, \tag{4.17}$$

$$\text{Supply}_{km_{(h)}} = \delta^s_{0k} + \delta^s_{k1}\text{Supply}_{km_{(h-1)}} + \delta^s_{k2}\text{Subsidy}_{km_{(h)}} + \varepsilon^s_{km}. \tag{4.18}$$

The IVs of demand and supply are valid that on the one hand, the last-hour demand (supply) is correlated with the current demand (supply). On the other hand, the IV satisfies the exclusion restriction that last-hour demand (supply) should not affect the current waiting time directly, but instead will affect waiting time only by affecting the current demand (supply).

Stage 2: Then I substitute the predicted demand and supply into the waiting time stage on the rider side (reservation wage on the driver side), and add the waiting time from last week

at the same hour and location as IV to correct endogeneity in the choice stage.

$$\text{Ewait}_{km} = \gamma^d_{k0} + \gamma^d_{k1}\widehat{Demand}_{km} + \gamma^d_{k2}\widehat{Supply}_{km} + \gamma^d_{k3}\text{Ewait}_{k(w-1)m} + \mu^d_{ktm}. \tag{4.19}$$

Stage 3: Since the choice stage is not linear, I cannot simply substitute the predicted Ewait into the third stage. Here, I use the control function approach. Following Petrin and Train (2010), I use the simplest version of the control function, which is a linear function of second-stage residual.

$$CF(\mu;\lambda) = \lambda\mu, \tag{4.20}$$

where μ is residual from the waiting time stage, and λ is the parameter to be estimated. Then utility with the control function is

$$U_{ikm} = u_{ikm} + \lambda\mu_{km} + \sigma_m\eta_{km} + \varepsilon_{ikm}. \tag{4.21}$$

where u_{ikm} is the mean utility in Equation (4.8), and η_{km} is i.i.d standard normal. In the estimation below, I use the random coefficient model so that σ_m here acts as the standard deviation of the intercept. I use a similar way on the supply side. The only difference is that there are only 2 stages (i.e., stages 1 & 3) on the supply side. No cruising time stage is needed as cruising time is observed by drivers. I replace network size in reservation wage with predicted network size to deal with the simultaneous endogeneity issue. After I set up models and address the endogeneity problem, I use Bayesian MCMC for model estimation (see more details about the estimation procedure in Appendix).

4.5 Estimation Result

In the sections above, I focus on the model-free evidence of network externalities and the structural model development. In this section, I will show both the reduced-form and

MCMC estimation results to quantify the network externalities and product intrinsic value in riders' and drivers' choice decisions.

4.5.1 Estimation for Network Externalities

I first show estimation results in terms of how network sizes affect riders' waiting time and drivers' cruising time. Note that, the cruising time stage is not used for estimation because drivers can observe cruising time directly. However, the estimates of the cruising time stage will be used in the following counterfactual analysis.

I use IVs in section 3.2 to correct endogeneity in each stage. Table 4.4 shows the 2SLS estimation results in terms of how network size on each side influences riders' waiting time and drivers' cruising time. The dependent variables of the four columns are riders' waiting time, drivers' cruising time, the demand size, and the supply size separately. The first column shows that riders' waiting time is influenced by both the direct and indirect network externalities. The waiting time is increasing with demand (direct network externalities), and is decreasing with supply (indirect network externalities). For example, a 10% increase in supply would decrease waiting time by 0.38 (=$\log(1.1) * 3.97$) minutes, and a 10% increase in demand would increase waiting time by 0.357 (=$\log(1.1) * 3.75$) minutes. Similarly, drivers' cruising time is influenced by both the direct and indirect network externalities as well. It is decreasing with demand (indirect network externalities) and is increasing with supply (direct network externalities). In a word, Table 4.4 shows that there exist both direct and indirect network externalities on both rider and driver sides. Next, I will show both the reduced-form and MCMC estimation results in riders' and drivers' choice decisions.

Table 4.4: How Network Externalities Influence Waiting/Cruising Time

	Wait_{2SLS}	Cruise_{2SLS}	IV_{Demand}	IV_{Supply}
Intercept	8.01 (0.29)***	4.77 (0.17)***	-0.17 (0.07)*	-0.30 (0.06)***
Log demand	3.75 (0.11)***	-1.09 (0.12)***		
Log supply	-3.97 (0.14)***	1.11 (0.13)***		
LagWeek wait	0.10 (0.008)***			
LagWeek cruise		0.15 (0.009)***		
Laghour log demand			0.59 (0.01)***	0.28 (0.01)***
Laghour log supply			0.20 (0.02)***	0.29 (0.01)***
Subsidy level				0.41 (0.01)***
Fixed Effects				
Week	YES	YES	YES	YES
Hour	YES	YES	YES	YES
Location	YES	YES	YES	YES
R^2	0.49	0.55	0.83	0.81
Num. obs.	5,672	5,672	5,672	5,672

***p<0.001, **p<0.01, *p<0.05
Note: Dependent variables are market-level waiting time, cruising time, log of demand and log of supply respectively. Standard errors are in parentheses.

4.5.2 Reduced-form Estimation

Previous model-free evidence in Figure 4.3 shows that riders favor short waiting time, and the section above already shows how direct and indirect network externalities affect riders' waiting time. In this section, I further show reduced-form evidence about how the waiting time influences riders' choices toward the two types of products on this platform. Moreover, besides waiting time, whether there exist other factors that also help to determine the degree of vertical differentiation.

Table 4.5 shows the reduced-form estimation results. On the demand side, after adding IVs to control for endogeneity (Model 2), waiting time adds a significant negative value to

Table 4.5: Reduced-Form Estimation Results for Rider and Driver

	Rider Side Model		Driver Side Model	
Models	Model 1	Model 2	Model 3	Model 4
PremiumDummy	0.22 (0.006)***	0.19 (0.007)***		
LogPrice	-3.02 (0.01)***	-3.02 (0.02)***		
Wait	-0.015 (0.002)***	-0.10 (0.01)***		
Intercept			-2.81 (0.05)***	-2.75 (0.06)***
Wage			6.33 (0.03)***	6.35 (0.03)***
LogDemand_p			-0.43 (0.02)***	-0.64 (0.05)***
LogSupply_p			0.09 (0.02)***	0.26 (0.06)***
LogDemand_s			-0.59 (0.02)***	-0.53 (0.05)***
LogSupply_s			0.65 (0.02)***	0.61 (0.05)***
Num. Obs.	391,642	391,642	388,296	388,296
Log Likelihood	-183,242	-183,195	-233,874	-232,307
IV	NO	YES	NO	YES

***p<0.001, **p<0.01, *p<0.05
Note: DV of Model 1 and 2 is a binary choice of whether the rider chooses the premium type; DV of Model 3 and 4 is a binary choice of whether the driver accepts the offer; Wage is driver's per-minute payment after considering cruising time to pick up rider; Standard errors are in parentheses.

riders, and premium product adds significant positive value to riders. This result is consistent with model-free evidence in Figure 4.3. In a word, both product intrinsic quality and network externalities are crucial in affecting riders' choices. On the driver side, according to the result of Model 4 in Table 4.5, network externalities would affect drivers' decisions in two ways. First, drivers prefer offers with higher wages, which is the per-minute wage that takes cruising time into consideration, and thus drivers favor offers with shorter cruising time. Second, network size on both sides would affect drivers' decisions by influencing drivers' opportunities to get other offers. More demand indicates more opportunities, and would decrease drivers' probability to accept the current offer; While more supply indicates more competitors and fewer opportunities, and would increase driver's probability to accept the current offer.

The reduced-form estimation results show that network externalities are key components to affect riders' and drivers' choices. Both the product intrinsic quality and the network externalities play important roles in determining the degree of vertical differentiation in two-sided markets. Next, I will add outside demand from Green/Yellow taxi in NYC, and use Bayesian MCMC to estimate the model with heterogeneous parameters.

4.5.3 MCMC Estimation Results for Heterogeneous Choice Model

The above reduced-form estimation results show that riders' waiting time and drivers' cruising time are affected by network size, both within and across sides. Next, I will show the MCMC estimation result, and quantify the value of each component for riders' and drivers' decisions. Before estimation, I first add outside options for riders. When a rider decides to order a ride, he might also consider outside options such as taxi and Uber. I include public taxi data in New York City as the outside option to control the total market size in the following analysis. The taxi data in New York City is public online, including all realized trips with trip start time, trip fare, distance and geographic information. Please find details about how I match outside demand with the data and how I construct alternative options for those who choose the outside option in the Appendix. The utility for the outside option is normalized to zero. Then I follow the estimation procedure in the model section and use the MCMC method to estimate the random coefficient model. I run a total of 300,000 MCMC iterations and report the posterior distribution of the parameters based on the last 100,000 draws. Table 4.6 shows the estimation results of the mean and standard deviation of parameters among the population.

Table 4.6 shows the MCMC estimation results on both the rider and driver sides. Let's first look at the estimation result on the demand (rider) side. First, product intrinsic value matters for riders when vertical differentiation is used. Choosing a premium trip would

Table 4.6: MCMC Estimation Results for Demand and Supply

	Variables	Mean Parameter	Standard Deviation σ
Demand Side	Intercept	9.10 (8.83, 9.36)	7.08 (6.90, 7.26)
	Premium Dummy	0.08 (0.06, 0.11)	1.77 (1.74, 1.80)
	Log Price	-2.93 (-2.99, -2.88)	2.33 (2.27, 2.40)
	Avg wait	-0.13 (-0.15, -0.10)	0.45 (0.43, 0.47)
Supply Side	Intercept	-3.76 (-3.98, -3.59)	2.68 (2.46, 2.83)
	Payment ($/min)	8.00 (7.78, 8.23)	4.66 (4.41,4.87)
	Log Demand$_p$	-0.53 (-0.69, -0.33)	0.81 (0.74, 0.93)
	Log Supply$_p$	0.29 (0.09, 0.50)	0.82 (0.71, 0.92)
	Log Demand$_s$	-0.23 (-0.37, -0.09)	0.71 (0.68, 0.75)
	Log Supply$_s$	0.22 (0.09, 0.34)	0.73 (0.65, 0.83)

Note: Numbers in parentheses indicate the 2.5 and 97.5 percentiles; Avg wait is the expected waiting time in that market; Subscript p represents the premium network and s represents the standard network.

make riders gain 0.08 more utility than choosing a standard trip. Second, I quantify network value to riders. The result shows that riders' waiting time, the indicator of network externalities, provides negative values to riders. One minute increase in waiting time would decrease riders' utility by 0.13, which is much larger than the utility gain if choosing the premium type. I further link the network externalities with waiting time. According to the waiting time stage result in Table 4.4, a 10% increase in supply productivity, keeping all others equal, can achieve a 0.38-minute (=$\log(1.1) * 3.97$) decrease in waiting time, and finally increase the utility of choosing the option by 0.05. Thus, besides product intrinsic value, network externality is a crucial factor in affecting riders' choices by influencing riders' waiting time.

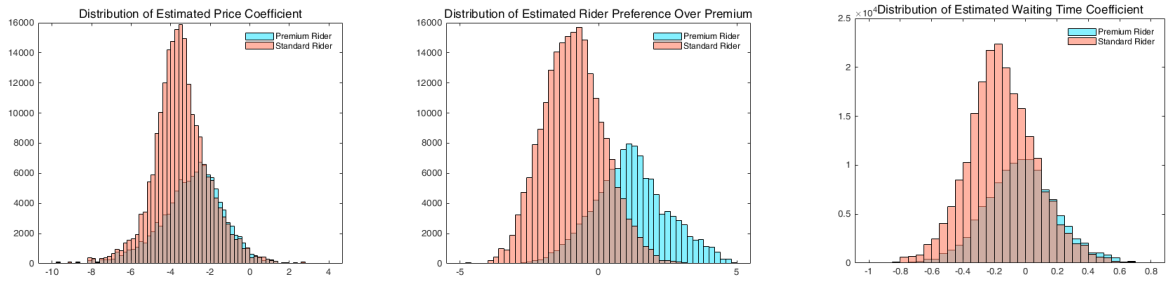
On the driver side, network externalities would affect drivers' decisions in two ways. First, network externalities can influence drivers' earnings by affecting the time drivers need to cruise around to pick up riders. Drivers prefer higher per-minute wages, which will increase when cruising time decreases. A $0.01 per-minute wage increase would increase the odds of accepting the offer by 8.33% = ($e^{8*0.01}$-1), which means the probability of

accepting the offer increases from 50% to 52% if fixing the original acceptance rate to 50%. Second, besides the effect on cruising time, network size would also affect drivers' probability of accepting the current offer by directly affecting their potential opportunities to get other offers. A 10% increase in premium demand would cause the odds of accepting the current offer to decrease by 5.2%, and a 10% increase in standard demand would cause the odds to decrease by 2.3%.

4.5.4 Visualization of MCMC Estimates

Visualization of Rider Heterogeneity

In the analysis above, I have used a heterogeneous model to estimate and quantify network value in riders' decision-making process. In this section, I plot the distribution of those estimates among riders, and get some visual insights about rider heterogeneity on the platform. Figure 4.7 shows the distribution of the estimated coefficients of some key variables among riders. First, I find that riders, on average, have negative price coefficients (Figure 4.7(a)), and the price coefficients of standard riders are more negative than those of premium riders. This tells us that the riders who choose standard type are more sensitive to price than riders who choose premium type. Second, the coefficient of the premium dummy is much larger for premium riders (Figure 4.7(b)), which means that riders who choose the premium type indeed value the high intrinsic quality of the premium type. Third, the coefficient of waiting time is more negative for standard riders (Figure 4.7(c)), which means standard riders are more sensitive to waiting time than premium riders. In a word, the estimated coefficients mainly tell us that the premium riders on the platform are less sensitive to price than standard riders. They are willing to pay higher prices for the premium type because they value mostly on the high intrinsic quality of the premium type, and can bear a bit longer waiting time.



(a) Price Coefficient **(b)** Premium Intrinsic Value **(c)** Waiting Time Coefficient

Figure 4.7: Distribution of Estimated Riders' Heterogeneous Coefficients

Visualization of Driver's Reservation Wage

In the analysis above, I have used model estimation to quantify network value in both riders' and drivers' decision-making processes. I further use estimates to calculate and visualize drivers' reservation utility to test the rationality of the model.

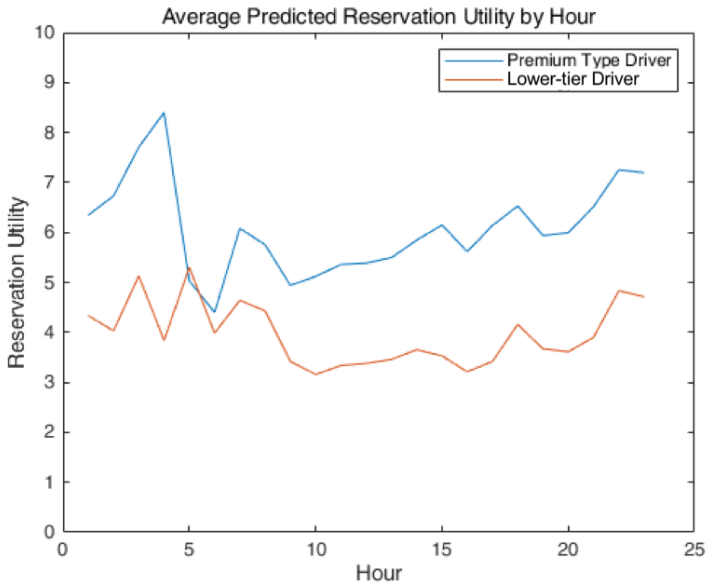


Figure 4.8: Average Predicted Reservation Utility by Hour.

Figure 4.8 shows drivers' average reservation utility across hours. First, on average, premium drivers have higher reservation utility than lower-tier drivers do. Second, for both types of drivers, they have higher reservation utility during rush hours and midnight periods.

The prediction results are quite reasonable. During rush hours, drivers have more opportunities because of more demand requests and thus have higher reservation utility. During the midnight period, drivers have high reservation utility because most of them would not go out for work and are used to staying at home for rest.

In this section, I estimate the model to quantify the network value and product intrinsic value to riders and drivers, and I also visualize drivers' reservation wages using the estimates. Next, I will do some counterfactual analysis to get managerial implications.

4.6 Platform Pricing Policy and Counterfactual Analysis

Previously, I build up a structural model to estimate and evaluate riders' and drivers' behavior under the platform's current pricing strategy. Drivers' (riders') choices toward the two types are determined by the earnings (prices), product value and the network value from the two types of product. The network size in equilibrium and the firm's profit will be fully determined by the firm's product strategy (vertical differentiation vs. homogenous project) and the price strategies. Once the two-sided platforms know the preferences of their target users (via data analysis like what I did above or via other approaches), they can design optimal strategies.

From the analysis above, I have quantified users' preferences over product value and network value, and users' price (earning) sensitivities on this platform. In this section, I will introduce more details about the platform's current strategy, and further do some counterfactual analysis using the model estimates above to examine whether there are better strategies for the platform and get some managerial implications.

4.6.1 Platform's Current Pricing and Subsidy Policy

According to Tables 4.3, 4.1 and 4.2, the platform's current strategy is that it charges different prices for different types. Moreover, the prices charged from riders and the earnings the platform provides to drivers are not the same.

Table 4.3 shows that there's a significant gap between payment from riders and payment given to drivers, especially for standard-type rides. That is, under the current strategy, the platform is not simply giving drivers what riders pay, but is actually subsidizing drivers, especially for drivers who serve standard-type trips. The data shows that 53.4% of premium trips offered subsidies to drivers, and 62.9% of the standard-type trips offered subsidies. Ride-sharing platforms usually charge a fixed commission rate from drivers based on drivers' earnings from trip fares. For example, Uber charges a 25% commission from drivers. The commission rate that the platform charges is 20% during the time window. On average, a premium trip provides a $0.12 subsidy to drivers, while a standard trip provides a $2.54 subsidy.

The possible reason why the platform wants to subsidize drivers is that they want to attract more users to gain enough network value. Once the platform gets a large scale, it can start to earn profit from the two types. Comparing the subsidy to premium and standard, the firm is subsidizing more to the standard type, which can be interpreted as that the platform is 'earning' more from premium trips than from standard trips. This strategy is actually consistent with the theoretical work in one-sided markets (Jing (2007)) that when there exist network externalities, firms can provide a low-quality product to enlarge the market size and expand the market, and earn profit from the high-end product. But is the current strategy optimal? Specifically, is vertical differentiation better than a homogeneous product? If yes, is there any better pricing strategy? Next, I will do counterfactual analyses to address these questions.

4.6.2 Counterfactual Analysis

According to the summary of the platform's current strategy above, the platform is using vertical differentiation and 'earns' more profit from the premium type. The network size of the standard type seems to be stably around two times the size of the premium product under the current pricing strategy. In this section, I use the above MCMC estimates and do counterfactual analyses regarding what strategy would be better for the platform. Specifically, I discuss several counterfactual scenarios below.

Base Case. The platform uses vertical differentiation under the current pricing strategy;

Case 1. The platform provides a homogeneous product, and only focuses on the low-end market, which is the standard product;

Case 2. The platform provides no subsidy to drivers. That is, earnings for drivers are set to be equal to what riders pay for each trip.

To predict demand and supply in the new counterfactual settings, I have to solve fixed points in the simultaneous system by substituting predicted waiting time and cruising time as a function of demand and supply into the choice stage for both rider and driver sides. Table 4.7 shows the detailed algorithm for solving fixed points.

Table 4.7: Algorithm to Solve Fixed Points

Steps:
1. Start with initial demand, d^0, and supply, s^0 in each market;
2. Calculate estimated waiting time w^1 and cruising time c^1 using d^0 and s^0;
3. Calculate new demand, d^1, and supply, s^1 using updated w^1 and c^1 in the choice model;
4. Repeat steps 2-3 until demand and supply converge. That is $\max(\lvert \mathrm{d}^n - d^{n-1} \rvert, \lvert \mathrm{s}^n - s^{n-1} \rvert) \leq \varepsilon$.

Table 4.8 shows the counterfactual results, listing the pricing strategies for both sides

of each product and the predicted number of realized trips and profit in each counterfactual scenario. I discuss the three counterfactual cases one by one.

Table 4.8: Counterfactual Results

	Base	Case 1	Base_2	Case 1_2	Case 2
Price (premium)	13.15	-	13.15	-	13.15
	(4.16)		(4.16)		(4.16)
Wage (premium)	16.6	-	16.6	-	13.15
	(7.89)		(7.89)		(4.16)
Price (standard)	9.0	9.0	9.0	9.0	9.0
	(0.23)	(0.23)	(4.16)	(4.16)	(4.16)
Wage (standard)	14.2	14.2	14.2	14.2	9.0
	(6.44)	(6.44)	(6.44)	(6.44)	(4.16)
# realized trips	189,872	173,150	175,324	173,872	110,941
Total profit ($)	-504,353	-444,413	-298,607	-390,796	129,622
Profit/trip ($)	-2.65	-2.57	-1.70	-2.25	1.17

I first compare the current vertical differentiation strategy with a homogeneous product case. Under the platform's current pricing strategy (Base case), the average premium price is \$13.15 and the standard price is a constant, \$9.0. And the average corresponding earning for drivers is \$16.6 from a premium trip and \$14.2 from a standard trip. For the homogeneous standard-only case (Case 1), I use the same pricing for the standard product. Compare Case 1 with the Base case, I find that the market size is expanded using vertical differentiation (189,872 $>$ 173,150). However, it seems more costly for the platform to maintain two networks as the per-trip cost is higher for vertical differentiation than the homogeneous product case (\$2.65$>$\$2.57). After further exploration, I find it's actually not the case. There is a tricky part of the platform's current strategy. According to the current pricing strategy, the price for the standard type is fixed, no matter how long the trip is, while the price for the premium type is positively correlated with trip duration. In this way, we find riders actually strategically choose the premium when their rides are short and choose standard when their rides are long. This can explain why the cost to maintain the two products is higher than the homogeneous case, as the platform extracts less surplus from riders because of their 'smart' behavior. This result also tells us that for ride-hailing companies, it's not wise to set

constant prices, especially when there are multiple choices for riders. Using constant price would undermine the benefit of vertical differentiation for usage-based platforms.

To answer whether it's better to use vertical differentiation, I next use a clean setting in Base_2 and Case 1_2, in which the price for the standard ride is set to be equal to the price for the premium type minus \$4, instead of using a constant price. In this way, we can somehow avoid riders' 'smart' behavior because the price of the premium is strictly higher than that of the standard type, regardless of what duration the trip has. Under this clean setting, I find the market size is still expanded when using vertical differentiation (175,324 $>$ 173,872). This tells us that the size gain from market expansion by using vertical differentiation can offset the loss from network size segmentation. Moreover, profit from vertical differentiation is also higher than the homogeneous product case (-\$1.7 $>$ -\$2.25). This tells us that by using vertical differentiation, the platform can extract more surplus from high-end customers.

Next, I further explore how the network size of the two products would change with the platform's pricing strategy. Specifically, I examine what would happen if no subsidy is provided to drivers when using vertical differentiation. In case 2, I set drivers' earnings from trips equal to the prices charged from riders. Compare Case 2 with the base case in Base_2, by providing subsidy to drivers, the platform in total spend \$(298,607 + 129,622) = \$428,229 and attract (175,324 $-$ 110,941) = 64,383 more realized trips. Thus, the cost of attracting an extra realized trip is $\frac{428{,}229}{64{,}383} = \$6.7$. I further calculate the cost of attracting an extra premium or standard realized trip separately. The result shows that the cost of attracting an extra premium realized trip is (6,857 + 72,200) / (34,792 $-$ 26,536) = \$9.6, while the cost of attracting an extra standard realized trip is (291,750 + 57,422) / (140,532 $-$ 84,405) = \$6.2. The cost of attracting an extra premium trip (marginal cost) is higher than that of a standard trip, \$9.6$>$\$6.2. Thus, it's more costly for the platform to expand and maintain the premium network. The network size of the standard is around three times the size of the premium, and the corresponding cost of attracting an additional standard realized trip

is only 64.6% (=6.2/9.6) of the cost of attracting an additional premium realized trip. This tells us the value of network size (scale effect). A large network size would generate a large positive network value, while the expansion of the product with a relatively small network size might be restrained.

4.6.3 Managerial Implications

I further get managerial implications from the above counterfactual analyses. First, for ride-hailing platforms (usage-based), prices charged from riders should be positively correlated with trip duration instead of being constant, especially when multiple options are offered to riders. Otherwise, it would undermine the platforms' benefit from using vertical differentiation. Second, extending a product line by offering another high-end product is better than just providing a low-end product for ride-hailing platforms under the competitive environment in New York City. That is, network externalities still make firms favor product differentiation in a two-sided market setting. On the one hand, market size can be expanded, as the effect of market expansion from vertical differentiation can offset the loss of a single network size advantage. On the other hand, firms can make more profit by extracting more surplus from high-end customers. Finally, even though high-end customers are more profitable, the high-end network is more costly to maintain and harder to expand because of its relatively smaller network value (smaller user group). In a word, two-side platforms should pay high attention to the network externalities when deciding whether to use vertical differentiation, and if yes, what price surcharges between products on each side to use.

4.7 Conclusion

The distinct feature in two-sided markets, network externalities, makes the conventional vertical differentiation strategy more complicated. On the one hand, platforms might be

better off designing vertically differentiated products in terms of attracting customers with heterogeneous preferences and expanding the market. On the other hand, offering multiple differentiated products might be less optimal than offering a homogeneous product if the segmented demand and supply limit the positive network effects in two-sided markets. Platforms might want to unite the markets from different products into one larger market to achieve greater positive network effects. Thus, understanding and quantifying the economic impact of network externalities on both demand and supply in two-sided markets is of great importance to both industry practitioners and the academic audience. This paper is among the first to empirically examine how vertical differentiation works in two-sided markets. I use a distinct data set from a two-sided ride-hailing platform that provides two vertically differentiated types of car service. I develop a structural model to simultaneously model the demand and supply, and use the Bayesian MCMC method to estimate and quantify users' preferences over intrinsic product value and network value. I present the main results concerning riders' and drivers' behavior, the counterfactual analyses, and the managerial implications for the platform below.

On the rider side, both the intrinsic product value and the network value are crucial in affecting riders' choices toward vertically differentiated products. According to the evidence from the data, riders who choose the premium product can get a better quality car and possibly better service quality, but has to on average wait longer because of the smaller network size and pay more in trip fare. Overall, one-third of riders are willing to order a premium trip. According to the estimation results in Tables 4.6 and 4.4, both intrinsic product value and network externalities are important for riders' decisions. First, the premium product adds value for riders. The utility obtained from the premium product is 0.08 larger than that from the standard product. Second, a large network size also provides value to riders. Specifically, a 10% increase in supply productivity would decrease riders' waiting time by 0.38 minutes on average and result in a 0.05 utility gain for riders.

On the driver side, network externalities also play important roles in affecting drivers' decisions about whether to accept the offer. Specifically, network externalities affect drivers' decisions in two ways. First, network externalities can influence drivers' earnings by affecting the time drivers need to cruise around to pick up riders. Data statistics show that, although a driver's per-minute earning from a premium trip is \$0.05 more than that from a standard trip, the cruising time for the premium trip is 0.34 minutes longer, making the earnings from the two types, taking cruising time into account, in equilibrium surprisingly close, around \$0.53 per minute. According to the estimation results in Tables 4.6 and 4.4, a 10% increase in demand would decrease cruising time by 0.10 minutes, which will further driver's per-minute earning. The exact increase in earnings is not determined, as it depends on other information about the offer, such as the price and trip duration. Suppose the demand variation can bring a \$0.01 per-minute wage increase, it would increase drivers' odds of accepting the offer by 8.33%. Second, network externalities would also affect drivers' decisions of whether to accept the current offer by influencing the driver's potential opportunity to receive other offers. Estimation results show that a 10% increase in premium demand would cause the odds of accepting the current offer to decrease by 5.2%, and a 10% increase in standard demand would cause the odds to decrease by 2.3%. More opportunities to receive other offers would decrease drivers' probability of accepting the current received offer. Thus, network externalities play important roles in both riders' and drivers' behavior, and the intrinsic product value and network value together determine the degree of vertical differentiation in equilibrium.

For two-sided platforms, it is crucial to understand users' preferences for both intrinsic product quality and network value when deciding what strategy to use. Once users' preferences are revealed using model analysis, the platform can further decide whether to use vertical differentiation, and if yes, what pricing strategies on each side for different products to use. Taking the focal ride-hailing platform as an example, I conduct several counterfactual

studies based on the estimation results to answer those questions. Specifically, I compare the platform's current strategy with two counterfactual settings, one being a standard-only homogenous product case and the other being vertical differentiation with a different pricing strategy, to obtain managerial implications about what can be better strategies for the firm. These counterfactual analyses yield several managerial implications. First, for ride-hailing (usage-based) platforms, prices charged from riders should be monotonically increasing with trip duration. The platforms should never use a constant price, especially when multiple options are offered. Second, extending the product line by offering a second high-end product is better for the ride-hailing platform than just providing a low-end product in the competitive environment of New York City. That is, network externalities still lead firms to favor product differentiation in two-sided markets. The benefit from the market expansion by using vertical differentiation can offset the loss from market segmentation. Moreover, firms can make more profit by extracting more surplus from high-end customers. Finally, even though high-end customers are more profitable, the high-end network is more costly to maintain and harder to expand because of its smaller network value (smaller user group). Two-sided platforms should take both the benefit and the cost of vertical differentiation into consideration, and if using vertical differentiation, carefully decide the optimal price surcharges between products on both demand and supply sides.

The findings should be interpreted based on the limitations inherent to this context. Though my research context offers a clean and detailed setting to observe riders' individual decisions towards two vertically differentiated products, it is also a setting lacking other ride-hailing competitors such as Uber. I try to make up by treating the green and yellow taxis in New York City as the main competitors and assume the three choices—the premium and standard types on the focal platform and the outside option—satisfy the IIA assumption in my model. I leave it to future research when more information about other ride-hailing competitors is available.

Chapter 5

Conclusion

This dissertation addresses the endogeneity problem for causal inference, and applies causal inference methods based on observational data to some interesting marketing and economic topics. For methodology development, I propose a new instrument-free method for correcting endogeneity. The method requires no instrument variables or even extra information. It corrects endogeneity using a control function approach by adding a generated regressor constructed on existing variables. I also extend the method for high-dimensional data. Besides the methodology development, I empirically apply causal inference methods to learn consumer behavior and get managerial insights.

The first essay develops a generalized two-stage copula endogeneity correction (2sCOPE) instrument-free method to correct endogeneity. I theoretically prove that 2sCOPE can yield consistent and efficient causal-effect estimates under a much weaker assumption than the proposed copula methods in Park and Gupta (2012) and Haschka (2021). The simulation results show that 2sCOPE yields consistent estimates under relaxed assumptions and improves estimation efficiency by up to 50%. Moreover, simulation studies also show that the 2sCOPE method can deal with normal endogenous variables, while other copula methods

in Park and Gupta (2012) and Haschka (2021) cannot even handle close-to-normal endogenous variables such as t distribution with the degree of freedom equals to 10. The estimation results in data application also confirm the performance of 2sCOPE because the estimated price coefficient using 2sCOPE is very close to the TSLS estimate, while OLS and previous copula method in Park and Gupta (2012) show large biases. Finally, the 2sCOPE is straightforward to implement, and can be widely applied to many other models, including linear regression models, linear panel models with mixed effects, random coefficient logit models, slope endogeneity, etc.

The second essay further studies the 2sCOPE method and extends it to the high dimensional setting. As described in detail above, 2sCOPE corrects endogeneity by adding the first-stage residual as a generated regressor to the outcome regression. Thus, the estimation of the first-stage residual plays a central role. With high-dimensional data (dimension can even be larger than sample size), estimation of the residual using the traditional method (i.e., OLS estimation) would be problematic. To address the high-dimension problem, I propose a method combining the 2sCOPE with some lasso-based methods (lasso, post-lasso, debiased lasso) to select important variables in the first stage. I call it the lasso-based 2sCOPE method. Simulation results show that 2sCOPE without feature selection in the first stage suffers a large finite-sample bias when the dimension is relatively large, while the lasso-based 2sCOPE method can improve substantial estimation accuracy and efficiency, by around 50%. I further apply the method to examine the effect of the government's policy stringency in response to COVID-19 on citizens' happiness. I use country-level cross-sectional data with a large dimension of country characteristics (i.e., around half of the sample size). The result shows that the estimated coefficient of the proposed lasso-based 2sCOPE method makes the effect of policy strictness on happiness more significantly negative. Increasing policy strictness by 10 units would decrease happiness growth by 0.034, which is larger than one standard deviation of happiness growth.

The above two essays focus on methodology development in correcting endogeneity for causal inference. The third essay empirically applies causal inference methods to solve the endogeneity problem in an interesting marketing scenario, two-sided markets. Specifically, I apply structural modeling with IV approach to correct the endogeneity of network size in examining vertical differentiation in two-sided markets. Data is from a leading ride-hailing platform that provides two vertically differentiated types of car service. I build up a simultaneous structural model together with the IV approach on both demand and supply sides, use the Bayesian MCMC method to quantify users' preferences over intrinsic product value and network value, and also conduct counterfactual analyses to get managerial insights. The result shows that both intrinsic product value and network externalities are important in determining the degree of vertical differentiation. Users favor high intrinsic product quality and hate long waiting (cruising) time affected by network size. For two-sided platforms, it is crucial to understand users' preferences over both intrinsic product quality and network value when deciding what product strategy (i.e., vertical differentiation vs. homogeneous product) and price strategies to use.

In the end, I wish to point out a few general limitations and research directions regarding the dissertation. First, in terms of methodology development to correct endogeneity, the dissertation proposes an instrument-free method and also adapts it to the high-dimensional setting combined with machine learning techniques. Though the method relaxes some key assumptions in the literature, it still relies on some assumptions. For example, for the 2sCOPE in the first essay to work best, the distributions of the endogenous regressors need to contain adequate information, which is violated for Bernoulli distributions or discrete distributions with small support. Moreover, 2sCOPE hinges on the normal structural error and Gaussian copula dependence structure. Future research is needed for more flexible methods that test and relax these assumptions. Second, it is interesting to apply our proposed method (2sCOPE or lasso-based 2sCOPE) for more classical empirical applications. For example,

we could further apply the proposed instrument-free method to quantify the endogenous network value in the third essay in the future.

Bibliography

Aghion, P., Bloom, N., Blundell, R., Griffith, R., and Howitt, P. (2005). Competition and innovation: An inverted u relationship. *Quarterly Journal of Economics*, 120:701–728. → page 9

Amado, A., Cortez, P., Rita, P., and Moro, S. (2018). Research trends on big data in marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics*, 24(1):1–7. → page 61

Anderson, E. T. and Simester, D. I. (2004). Long-run effects of promotion depth on new versus established customers: three field studies. *Marketing Science*, 23(1):4–20. → page 15

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014. → page 6

Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4):69–85. → pages 15, 63

Arora, N. and Huber, J. (2001). Improving parameter estimates and model prediction by aggregate customization in choice experiments. *Journal of Consumer Research*, 28(2):273–283. → page 36

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1):80–98. → page 61

Ataman, M. B., Van Heerde, H. J., and Mela, C. F. (2010). The long-term effect of marketing strategy on brand sales. *Journal of Marketing Research*, 47(5):866–882. → page 15

Atefi, Y., Ahearne, M., Maxham III, J. G., Donavan, D. T., and Carlson, B. D. (2018). Does selective sales force training work? *Journal of Marketing Research*, 55(5):722–737. → pages 17, 64, 82

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497. → page 15

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725. → page 61

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317. → pages 62, 65

Bai, J. and Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics*, 24(4):607–629. → pages 62, 65

Becker, J.-M., Proksch, D., and Ringle, C. M. (2021). Revisiting gaussian copulas to handle endogenous regressors. *Journal of the Academy of Marketing Science*, pages 1–21. → pages 9, 10, 11, 12, 13, 18, 23, 39, 47

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429. → pages 62, 65, 66, 70

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547. → pages 62, 65

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50. → pages 61, 62, 65

Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650. → pages 62, 65

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262. → page 16

Blundell, R. and Matzkin, R. L. (2014). Control functions in nonseparable simultaneous equations models. *Quantitative Economics*, 5(2):271–295. → page 13

Blundell, R. and Powell, J. L. (2003). Endogeneity in nonparametric and semiparametric regression models. *Econometric society monographs*, 36:312–357. → page 13

Blundell, R. W. and Powell, J. L. (2004). Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679. → page 13

Bombaij, N. J. and Dekimpe, M. G. (2020). When do loyalty programs work? the moderating role of design, retailer-strategy, and country characteristics. *International Journal of Research in Marketing*, 37(1):175–195. → pages 17, 64, 82

Bryan, K. A. and Gans, J. S. (2019). A theory of multihoming in rideshare competition. *J. Econ. Manag. Strateg.*, 28(1):89–96. → page 91

Burmester, A. B., Becker, J. U., van Heerde, H. J., and Clement, M. (2015). The impact of pre-and post-launch publicity and advertising on new product sales. *International Journal of Research in Marketing*, 32(4):408–417. → pages 17, 64, 82

Chen, M. K., Rossi, P. E., Chevalier, J. A., and Oehlsen, E. (2019). The value of flexible work: Evidence from uber drivers. *Journal of Political Economy*, 127(6):2735–2794. → page 90

Chintagunta, P., Dubé, J.-P., and Goh, K. Y. (2005). Beyond the endogeneity bias: The effect of unmeasured brand characteristics on household-level brand choice models. *Management Science*, 51(5):832–849. → page 52

Chintagunta, P., Erdem, T., Rossi, P. E., and Wedel, M. (2006). Structural modeling in marketing: review and assessment. *Marketing Science*, 25(6):604–616. → page 16

Cui, G., Wong, M. L., and Lui, H.-K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, 52(4):597–612. → page 61

Danaher, P. (2007). Modeling page views across multiple websites with an application to internet reach and frequency prediction. *Marketing Science*, 26:422–437. → pages 19, 67

Danaher, P. and Smith, M. (2011). Modeling multivariate distributions using copulas: Applications in marketing (with discussion and rejoinder). *Marketing Science*, 30:4–21. → pages 7, 20, 55, 58, 67, 68

Datta, H., Foubert, B., and Van Heerde, H. J. (2015). The challenge of retaining customers acquired with free trials. *Journal of Marketing Research*, 52(2):217–234. → pages 17, 64, 82

Dotson, J. P. and Allenby, G. M. (2010). Investigating the strategic influence of customer and employee satisfaction on firm financial performance. *Marketing Science*, 29(5):895–908. → page 16

Dubé, J.-P., Chintagunta, P., Petrin, A., Bronnenberg, B., Goettler, R., Seetharaman, P. S., Sudhir, K., Thomadsen, R., and Zhao, Y. (2002). Structural applications of the discrete choice model. *Marketing letters*, pages 207–220. → page 16

Ebbes, P., Wedel, M., and Böckenholt, U. (2009). Frugal iv alternatives to identify the parameter for an endogenous regressor. *Journal of Applied Econometrics*, 24(3):446–468. → page 16

Ebbes, P., Wedel, M., Böckenholt, U., and Steerneman, T. (2005). Solving and testing for regressor-error (in) dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics*, 3(4):365–392. → pages 7, 16, 55

Eckert, C. and Hohberger, J. (2022). Addressing endogeneity without instrumental variables: An evaluation of the gaussian copula approach for management research. *Journal of Management*, DOI: 10.1177/01492063221085913. → pages 8, 12, 13, 55

Elshiewy, O. and Boztug, Y. (2018). When back of pack meets front of pack: How salient and simplified nutrition labels affect food sales in supermarkets. *Journal of Public Policy & Marketing*, 37(1):55–67. → pages 17, 64, 82

Erickson, T. and Whited, T. M. (2002). Two-step gmm estimation of the errors-in-variables model using high-order moments. *Econometric Theory*, 18(3):776–799. → pages 7, 16

Feldman, J., Zhang, D. J., Liu, X., and Zhang, N. (2022). Customer choice models vs. machine learning: Finding optimal product displays on alibaba. *Operations Research*, 70(1):309–328. → page 61

Frechette, G. R., Lizzeri, A., and Salz, T. (2019). Frictions in a competitive, regulated market: Evidence from taxis. *American Economic Review*, 109(8):2954–92. → page 91

Gabszewicz, J. J., Shaked, A., Sutton, J., and Thisse, J.-F. (1986). Segmenting the market: The monopolist's optimal product mix. *Journal of economic theory*, 39(2):273–289. → page 86

Gabszewicz, J. J. and Wauthy, X. Y. (2014). Vertical product differentiation and two-sided markets. *Economics Letters*, 123(1):58–61. → page 90

Godes, D. and Mayzlin, D. (2009). Firm-created word-of-mouth communication: Evidence from a field test. *Marketing science*, 28(4):721–739. → page 15

Greene, W. H. (2012). *Econometric Analysis. 7th Edition*. Prentice Hall, Upper Saddle River, NJ. → page 32

Gruner, R. L., Vomberg, A., Homburg, C., and Lukas, B. A. (2019). Supporting new product launches with social media communication and online advertising: sales volume and profit implications. *Journal of Product Innovation Management*, 36(2):172–195. → pages 17, 64, 82

Guda, H. and Subramanian, U. (2019). Your uber is arriving: Managing on-demand workers through surge pricing, forecast communication, and worker incentives. *Management Science*, 65(5):1995–2014. → page 91

Guitart, I. A., Gonzalez, J., and Stremersch, S. (2018). Advertising non-premium products as if they were premium: The impact of advertising up on advertising elasticity and brand equity. *International journal of research in marketing*, 35(3):471–489. → pages 17, 64, 82

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209. → page 15

Hartmann, W., Nair, H. S., and Narayanan, S. (2011). Identifying causal marketing mix effects using a regression discontinuity design. *Marketing Science*, 30(6):1079–1097. → pages 15, 16

Haschka, R. E. (2021). Express: Handling endogenous regressors using copulas: A generalization to linear panel models with fixed effects and correlated regressors. *Journal of Marketing Research, First appeared online on Dec 18, 2021*. → pages 3, 10, 11, 12, 14, 17, 31, 32, 39, 64, 68, 125, 126

Heckman, J. J. and Robb Jr, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of econometrics*, 30(1-2):239–267. → page 13

Heitmann, M., Landwehr, J. R., Schreiner, T. F., and van Heerde, H. J. (2020). Leveraging brand equity for effective visual product design. *Journal of Marketing Research*, 57(2):257–277. → pages 17, 64, 82

Hoch, S. J., Kim, B.-D., Montgomery, A. L., and Rossi, P. E. (1995). Determinants of store-level price elasticity. *Journal of marketing Research*, 32(1):17–29. → page 52

Hogan, V. and Rigobon, R. (2003). Using unobserved supply shocks to estimate the returns to education. *Unpublished manuscript*. → pages 7, 16

Javanmard, A. and Montanari, A. (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909. → pages 62, 65, 66, 73, 74

Javanmard, A. and Montanari, A. (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554. → pages 73, 74

Jing, B. (2007). Network externalities and market segmentation in a monopoly. *Economics Letters*, 95(1):7–13. → pages 86, 89, 117

Johnson, G. A., Lewis, R. A., and Nubbemeyer, E. I. (2017). Ghost ads: Improving the economics of measuring online ad effectiveness. *Journal of Marketing Research*, 54(6):867–884. → page 15

Keller, W. I., Deleersnyder, B., and Gedenk, K. (2019). Price promotions and popular events. *Journal of Marketing*, 83(1):73–88. → pages 17, 64, 82

Kleibergen, F. and Zivot, E. (2003). Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics*, 114(1):29–72. → page 15

Lamey, L., Deleersnyder, B., Steenkamp, J.-B. E., and Dekimpe, M. G. (2018). New product success in the consumer packaged goods industry: A shopper marketing approach. *International Journal of Research in Marketing*, 35(3):432–452. → pages 17, 64, 82

Lancaster, K. (1990). The Economics of Product Variety: A Survey. *Mark. Sci.*, 9(3):189–206. → page 89

Leeb, H. and Pötscher, B. M. (2008a). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376. → page 65

Leeb, H. and Pötscher, B. M. (2008b). Guest editors'editorial: Recent developments in model selection and related areas. *Econometric Theory*, 24(2):319–322. → page 65

Lewbel, A. (1997). Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and r&d. *Econometrica: journal of the econometric society*, pages 1201–1213. → pages 7, 16

Li, H. and Srinivasan, K. (2019). Competitive dynamics in the sharing economy: an analysis in the context of airbnb and hotels. *Marketing Science*, 38(3):365–391. → page 90

Lin, S. (2020). Two-sided price discrimination by media platforms. *Marketing Science*, 39(2):317–338. → pages 90, 91

Liu, Q. and Serfes, K. (2013). Price Discrimination in Two-Sided Markets. *J. Econ. Manag. Strateg.*, page 19. → page 90

Mackiewicz, R. and Falkowski, A. (2015). The use of weber fraction as a tool to measure price sensitivity: a gain and loss perspective. *Advances in Consumer Research*, 43. → page 52

Maskin, E. and Riley, J. (1984). Monopoly with incomplete information. *The RAND Journal of Economics*, 15(2):171–196. → page 86

Mendelson, H. (2000). Organizational architecture and success in the information technology. *Management Science*, 46:513–529. → page 8

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106. → page 61

Mussa, M. and Rosen, S. (1978). Monopoly and product quality. *Journal of Economic theory*, 18(2):301–317. → pages 86, 89

Narayanan, S. and Kalyanam, K. (2015). Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Science*, 34(3):388–407. → page 15

Ngai, E. W. and Wu, Y. (2022). Machine learning in marketing: A literature review, conceptual framework, and research agenda. *Journal of Business Research*, 145:35–48. → page 61

Novak, S. and Stern, S. (2009). Complementarity among vertical integration decisions: Evidence from automobile product development. *Management Science*, 55(2):311–332. → page 15

Otter, T., Gilbride, T. J., and Allenby, G. M. (2011). Testing models of strategic behavior characterized by conditional likelihoods. *Marketing Science*, 30(4):686–701. → page 16

Papies, D., Ebbes, P., and Van Heerde, H. J. (2017). Addressing endogeneity in marketing models. In *Advanced methods for modeling markets*, pages 581–627. Springer. → pages 9, 15

Park, S. and Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, 31(4):567–586. → pages 2, 3, 7, 8, 9, 11, 12, 15, 16, 17, 18, 19, 20, 21, 23, 27, 34, 37, 44, 48, 51, 56, 58, 64, 68, 82, 125, 126, 139, 15

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The annals of applied statistics*, 4(1):53. → page 61

Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of marketing research*, 47(1):3–13. → pages 13, 26, 70, 108

Qian, Y. (2008). Impacts of entry by counterfeiters. *Quarterly Journal of Economics*, 123:1577–1609. → page 15

Qian, Y. and Xie, H. (2021). Simplifying bias correction for selective sampling: A unified distribution-free approach to handling endogenously selected samples. *Marketing Science, Forthcoming, available at https://www.nber.org/papers/w28801*. → page 36

Rigobon, R. (2003). Identification through heteroskedasticity. *Review of Economics and Statistics*, 85(4):777–792. → pages 7, 16

Rochet, J.-c. and Tirole, J. (2006). Two-sided markets : a progress report. *RAND J. Econ.*, 37(3):645–667. → page 85

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. → pages 7, 15, 63

Rutz, O. J. and Watson, G. F. (2019). Endogeneity and marketing strategy research: An overview. *Journal of the Academy of Marketing Science*, 47(3):479–498. → page 15

Salant, S. W. (1989). When is inducing self-selection suboptimal for a monopolist? *The Quarterly Journal of Economics*, 104(2):391–397. → pages 86, 89

Shaked, A. V. and Sutton, J. (1982). Relaxing Price Competition Through Product Differentiation. *Rev. Econ. Stud.*, page 11. → page 89

Shi, H., Sridhar, S., Grewal, R., and Lilien, G. (2017). Sales representative departures and customer reassignment strategies in business-to-business markets. *Journal of Marketing*, 81(2):25–44. → page 15

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231. → pages 19, 67

Sorescu, A., Warren, N. L., and Ertekin, L. (2017). Event study methodology in the marketing literature: An overview. *Journal of the Academy of Marketing Science*, 45:186–207. → page 8

Sridhar Moorthy, K. (1984). Market Segmentation, Self-Selection, and Product Line Design. Technical Report 4. → page 89

Stokey, N. L. (1979). Intertemporal price discrimination. *The Quarterly Journal of Economics*, pages 355–371. → pages 86, 89

Sudhir, K. (2001). Competitive pricing behavior in the auto market: A structural analysis. *Marketing Science*, 20(1):42–60. → page 16

Sun, B. (2005). Promotion effect on endogenous consumption. *Marketing science*, 24(3):430–443. → page 16

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288. → pages 62, 65, 71

Van Heerde, H. J., Gijsenberg, M. J., Dekimpe, M. G., and Steenkamp, J.-B. E. (2013). Price and advertising effectiveness over the business cycle. *Journal of Marketing Research*, 50(2):177–193. → page 15

Vandenbosch, M. and Weinberg, C. (1995). Product and Price Competition in a Two-Dimensional Vertical Differentiation Model. *Mark. Sci.*, page 26. → page 89

Villas-Boas, J. M. and Winer, R. S. (1999). Endogeneity in brand choice models. *Management science*, 45(10):1324–1338. → pages 52, 55

Wang, Y. and Blei, D. (2019). The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596. → page 16

Wang, Y., Wu, C., and Zhu, T. (2019). Mobile hailing technology and taxi driving behaviors. *Marketing Science*, 38(5):734–755. → page 91

Wetzel, H. A., Hattula, S., Hammerschmidt, M., and van Heerde, H. J. (2018). Building and leveraging sports brands: evidence from 50 years of german professional soccer. *Journal of the Academy of Marketing Science*, 46(4):591–611. → pages 17, 64, 82

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2):420–445. → page 13

Yang, F., Qian, Y., and Xie, H. (2022). Addressing endogeneity using a two-stage copula generated regressor approach. Technical report, National Bureau of Economic Research. → pages 60, 64, 65, 66, 67, 68, 82

Yang, S., Chen, Y., and Allenby, G. M. (2003). Bayesian analysis of simultaneous demand and supply. *Quantitative marketing and economics*, 1(3):251–275. → pages 16, 55

Zervas, G., Proserpio, D., and Byers, J. W. (2017). The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Journal of marketing research*, 54(5):687–705. → page 90

Zhang, S., Lee, D., Singh, P. V., and Srinivasan, K. (2017). How much is an image worth? airbnb property demand estimation leveraging large scale image analytics. *Airbnb Property Demand Estimation Leveraging Large Scale Image Analytics (May 25, 2017)*. → page 61

Appendix A

Chapter 2 Appendices

Appendix: Proofs

A.1 Proof of Theorem 1

Under the Gaussian copula assumption for structural error term ξ_t and the endogenous regressor P_t, and the normality assumption of ξ_t, the outcome regression becomes (Equation 2.6)

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

Because of the exogeneity assumption of W_t in linear model (Equation 2.1), $Cov(W_t, \xi_t) = 0$,

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

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

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

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

A.2 Assumption 4.b in CopulaP&G

According to Park and Gupta (2012), under a Gaussian copula model for $(P_{1,t}, P_{2,t}, \xi_t)$, the structural model in Equation (2.13) with two endogenous regressors can be re-expressed as

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

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

A.3 COPE Method Development

Under the Gaussian copula model for the endogenous regressor, P_t, the correlated exogenous regressor, W_t, and the structural error term, ξ_t in Equation (3.5), the structural error in

Equation (2.1) can be re-expressed as

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

In this way, the structural error term ξ_t is split into two parts: one part as a function of P_t^* and W_t^* that captures the endogeneity of P_t and the association of W_t with $\xi_t|P_t$ [1], and the other part as an independent new error term. Then, we substitute Equation (A.2) into the main model in Equation (2.1), and obtain the following regression equation:

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

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

Theorem A1. Estimation Consistency. Assuming (1) the error term is normal, (2) the endogenous regressor P_t and exogenous regressors W_t are non-normally distributed, and (3) a Gaussian Copula for the error term, P_t and W_t, $Cov(\omega_{3,t}, W_t) = Cov(\omega_{3,t}, P_t) = Cov(\omega_{3,t}, W_t^*) = Cov(\omega_{3,t}, P_t^*) = 0$ and thus the OLS estimation of Equation (A.3) yields consistent estimates of model parameters.

Proof: See Online Appendix, Proof of Theorem A1

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

As shown in Theorem A1, the proposed COPE method does not require the uncorrelatedness between P_t^* and W_t for consistent model estimation, an assumption needed for CopulaP&G. In fact, CopulaP&G can be obtained as a special case of the COPE: when W_t is uncorrelated with P_t (i.e. $\rho_{pw} = 0$) and also uncorrelated with P_t^* under the joint copula model, $\frac{-\sigma_\xi \rho_{pw} \rho_{p\xi}}{1-\rho_{pw}^2} W_t^*$ in Equation (A.3) vanishes and COPE based on Equation (A.3) reduces to CopulaP&G base on Equation (2.6). This broader applicability of COPE is a merit of COPE. However, similar to CopulaP&G, COPE requires the non-normality of the endogenous regressor P_t to fulfill the full-rank identification assumption. Moreover, a correlation between endogenous regressor P and the exogenous regressors W will cause CopulaP&G to transfer the endogeneity from P to W; the correction for the induced endogenous regressor W should have the same non-normality assumption for model identification as with P. In the next subsection, we will develop a novel two-stage COPE method that relaxes the regressor non-normality assumption. We further extend the model to incorporate multiple endogenous regressors in the section 4.4.

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

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

we have:

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

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

where m is a function of all the ρs. Under the Gaussian Copula assumption above, we can derive ξ^*_t as a function of P_t and W_t. After simplification, the structural error in Equation (2.13) can be decomposed as

$$\xi_t = \sigma_\xi \xi^*_t = \eta_1 P^*_{1,t} + \eta_2 P^*_{2,t} - (\eta_1\rho_{wp1} + \eta_2\rho_{wp2})W^*_t + \sigma_\xi \cdot m \cdot \omega_{4,t}. \tag{A.5}$$

where

$$\begin{aligned} \eta_1 &= \frac{\sigma_\xi\rho_{\xi p1}(1-\rho_{wp2}^2) - \sigma_\xi\rho_{\xi p2}(\rho_p - \rho_{wp1}\rho_{wp2})}{1-\rho_p^2-\rho_{wp1}^2-\rho_{wp2}^2+2\rho_p\rho_{wp1}\rho_{wp2}}, \\ \eta_2 &= \frac{\sigma_\xi(\rho_{wp1}\rho_{wp2}\rho_{\xi p1} + \rho_{\xi p2} - \rho_p\rho_{\xi p1} - \rho_{wp1}^2\rho_{\xi p2})}{1-\rho_p^2-\rho_{wp1}^2-\rho_{wp2}^2+2\rho_p\rho_{wp1}\rho_{wp2}}. \end{aligned} \tag{A.6}$$

The COPE method with one endogenous regressor in Equation (A.3) is then extended to

$$Y_t = \mu + P_{1,t}\alpha_1 + P_{2,t}\alpha_2 + W_t\beta + \eta_1 P^*_{1,t} + \eta_2 P^*_{2,t} - (\eta_1\rho_{wp1} + \eta_2\rho_{wp2})W^*_t + \sigma_\xi \cdot m \cdot \omega_{4,t}. \tag{A.7}$$

In Equation (A.7), the new error term $\omega_{4,t}$ is uncorrelated with all the regressors on the right-hand side of Equation (A.7). Thus, the OLS estimation of Equation (A.7) provides consistent estimates of structural regression model parameters $(\mu, \alpha_1, \alpha_2, \beta)$.

A.4 Proof of Theorem A1

Under the Gaussian copula model for (P_t, ξ_t) and the normality assumption of the error term ξ_t, we can divide ξ_t into an endogenous and exogenous part and our proposed COPE method is based on the OLS estimation of the regression below (Equation A.3) by adding P_t^* and W_t^* as generated regressors.

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

We want to prove that the new error term $\omega_{3,t}$ is uncorrelated with all terms of the right-hand side. Since $\omega_{1,t}$, $\omega_{2,t}$ and $\omega_{3,t}$ follow a standard multivariate Gaussian distribution (Equation 3.5), they are independent. According to the same equation, W_t^* and P_t^* are linear functions of $\omega_{1,t}$ and $\omega_{2,t}$. Thus, P_t^* and W_t^* are normally distributed and are independent of $\omega_{3,t}$. Since functions of independent variables are still independent, P_t (W_t), as a function of P_t^* (W_t^*), would be uncorrelated with $\omega_{3,t}$ and thus $\omega_{3,t}$ is not correlated with P_t, P_t^*, W_t and W_t^* on the right-hand side of Equation (A.3). Since P_t and W_t are nonnormal distributed, the full rank assumption is satisfied and thus COPE yields consistent estimates. **Theorem proved**.

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

the linear regression is:

$$Y_t = \beta_0 + \beta_1 P_t + \beta_2 W_{1,t} + \beta_3 W_{2,t} + \xi_t \tag{A.8}$$

Under the Gaussian Copula assumption,

$$\begin{pmatrix} P_t^* \\ W_{1,t}^* \\ W_{2,t}^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_\xi \\ \rho_1 & 1 & \rho_w & 0 \\ \rho_2 & \rho_w & 1 & 0 \\ \rho_\xi & 0 & 0 & 1 \end{bmatrix} \right) \tag{A.9}$$

The multivariate normal distribution can be written as follows:

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

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

$$\xi_t = \sigma_\xi \xi_t^* = \frac{\sigma_\xi \rho_\xi (1-\rho_w^2)}{1-\rho_1^2-\rho_2^2+2\rho_1\rho_2\rho_w+\rho_w^2} \left(P_t^* - \frac{\rho_1 - \rho_2\rho_w}{1-\rho_w^2} W_{1,t}^* - \frac{\rho_2 - \rho_1\rho_w}{1-\rho_w^2} W_{2,t}^* \right) + \sigma_\xi \gamma \cdot \omega_{4,t}. \tag{A.10}$$

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

$$Y_t = \beta_0 + \beta_1 P_t + \beta_2 W_{1,t} + \beta_3 W_{2,t} + \beta_4 P_t^* + \beta_5 W_{1,t}^* + \beta_6 W_{2,t}^* + \sigma_\xi \gamma \cdot \omega_{4,t} \tag{A.11}$$

where

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

Since $\omega_{4,t}$ is independent of P_t^*, $W_{1,t}^*$ and $W_{2,t}^*$, it would also be uncorrelated with any functional form of P_t^*, $W_{1,t}^*$ and $W_{2,t}^*$, and thus $\omega_{4,t}$ is uncorrelated with any other terms in Equation (A.11). The COPE method can easily be extended to the case with multiple endogenous regressors by adding copula transformation of each regressor as generated regressors into the outcome regression, and the proof of estimation consistency is similar.

A.5 Proof of Theorem 2: Consistency of 2sCOPE

We have shown the derivation of 2sCOPE method in Section 3. The system of equations used in 2sCOPE method (Equations 3.1, 2.9) leads to the following equations

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

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

Next we show that this result can be easily extended to the multi-dimension W_t case.

We first derive the system of equations of the 2sCOPE method. Here we take 2-dimension W_t as an example. Because of the Gaussian relationship among P_t^* and W_t^* we assumed in Equation (A.9), the first stage regression becomes

$$\begin{aligned} P_t^* &= \frac{\rho_1 - \rho_2\rho_w}{1-\rho_w^2} W_{1,t}^* + \frac{\rho_2 - \rho_1\rho_w}{1-\rho_w^2} W_{2,t}^* + \sqrt{1-\rho_1^2 - \frac{(\rho_2 - \rho_1\rho_w)^2}{1-\rho_w^2}}\,\omega_{3,t} \\ &= \frac{\rho_1 - \rho_2\rho_w}{1-\rho_w^2} W_{1,t}^* + \frac{\rho_2 - \rho_1\rho_w}{1-\rho_w^2} W_{2,t}^* + \varepsilon_{2,t} \\ &= \gamma_1 W_{1,t}^* + \gamma_2 W_{2,t}^* + \varepsilon_{2,t}. \end{aligned} \tag{A.12}$$

The structural error ξ_t in Equation (A.8) and the first-stage error term $\varepsilon_{2,t}$ are linear transformations of the Gaussian data $(\xi_t, P_t^*, W_{1,t}^*, W_{2_t}^*)$ and thus follow a bivariate normal distribution. Thus, ξ_t can be decomposed to a sum of one term containing $\varepsilon_{2,t}$ and an independent new error term, resulting in the following regression equation:

$$Y_t = \beta_0 + \beta_1 P_t + \beta_2 W_{1,t} + \beta_3 W_{2,t} + \beta_4 \varepsilon_{2,t} + \sigma_\xi \gamma \cdot \omega_{4,t}. \tag{A.13}$$

where

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

Since $\omega_{4,t}$ is independent of P_t^*, $W_{1,t}^*$ and $W_{2,t}^*$, it is uncorrelated with any functional form of P_t^*, $W_{1,t}^*$ and $W_{2,t}^*$, and thus $\omega_{4,t}$ is uncorrelated with P_t, $W_{1,t}$, $W_{2,t}$ and $\varepsilon_{2,t}$ in Equation (A.13). Thus, 2sCOPE that performs OLS regression of Equation (A.13) yields consistent model estimates. Without loss of generality, the result can be extended to cases with any dimension of W_t.

A.6 Proof of Theorem 3: Nonnormality Assumption Relaxed

In this section, we prove that our proposed 2sCOPE method can relax the nonnormality assumption on the endogenous regressors imposed in CopulaP&G, while COPE does not.

We first examine the COPE method in Equation (A.3),

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

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

We then examine the 2sCOPE method in Equation (3.8).

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

$$\varepsilon_t = P_t^* - \rho_{pw}W_t^*.$$

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

A.7 Proof of Theorem 4: Variance Reduction

According to the COPE method in Equation (A.3),

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

Note that all the regressors on the right-hand side of the equation above are not endogenous. The coefficients of P_t^* and W_t^* follows a linear relationship. Denote δ_3 and δ_4 the coefficients of P_t^* and W_t^* respectively. Then,

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

With the two-stage estimation in 2sCOPE (Equation 3.8), ρ_{pw} is estimated in the first stage and is thus treated as a known parameter in the main regression. That is, 2sCOPE can be viewed as the COPE method with a linear restriction. The linear restriction is,

$$\delta_4 + \hat{\rho}_{pw}\delta_3 = 0. \tag{A.14}$$

In this case, the two-stage copula method (2sCOPE) can be viewed as one kind of restricted least square estimation based on COPE. We next prove that restricted least square can achieve reductions in standard errors. Suppose we simplify the regression expression in Equation (A.3) as

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

where $\varepsilon \sim N(0, \sigma^2 I)$, $X \equiv (1, P_t, W_t, P_t^*, W_t^*)$, and $\theta = (\mu, \alpha, \beta, \delta_3, \delta_4)$. The restriction in Equation (A.14) becomes

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

Thus, the 2sCOPE yields the least square estimates $\hat{\theta}_2$ of Equation (A.3) subject to the above restriction, whereas COPE yields the unrestricted least square estimates, $\hat{\theta}_1$, as follows.

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

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

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

Therefore,

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

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

We have proved that there would be variance reduction when there exist restriction of

parameters. When the exogenous regressor W_t is a scalar, the linear restriction is shown in Equation (A.14). We next show that when W_t is extended to a multi-dimension vector, there are still linear restrictions and variance reduction of 2sCOPE. We take a 2-dimension W_t as an example below. According to the 2sCOPE method with 2-dimension W_t in Equations (A.12, A.13), 2sCOPE is equivalent to adding two restrictions to COPE in Equation (A.11). The two restrictions are:

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

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

where $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are estimated and obtained in the first stage in Equation (A.12). Thus, compared with COPE, we still have variance reduction using 2sCOPE in the 2-W case. Without loss of generality, this result can be extended to cases with any dimension of W_t.

A.8 2sCOPE for Slope Endogeneity with Correlated and Normally Distributed Regressors

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

$$Y_t = \mu + P_t \alpha_t + W_t' \beta_t + \eta_t, \qquad \text{where } \alpha_t = \bar{\alpha} + \xi_t, \tag{A.15}$$

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

endogeneity problem is derived as follows.

$$\begin{aligned} Y_t &= \mu + P_t(\bar{\alpha} + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* + \frac{-\sigma_\xi \rho_{pw}\rho_{p\xi}}{1-\rho_{pw}^2} W_t^* + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}}\omega_{3,t}) + W_t'\beta_t + \eta_t \\ &= \mu + P_t\bar{\alpha} + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t \times P_t^* + \frac{-\sigma_\xi \rho_{pw}\rho_{p\xi}}{1-\rho_{pw}^2} P_t \times W_t^* + W_t'\beta_t + \\ &\quad \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} P_t \times \omega_{3,t} + \eta_t. \end{aligned} \tag{A.16}$$

Given both $P_t \times P_t^*$ and $P_t \times W_t^*$ in Equation (A.16), the unobserved variable $w_{3,t}$ is independent of all regressors (P_t, W_t, P_t^*, W_t^*) and uncorrelated with functions of these regressors. Thus, Equation (A.16) can be estimated using standard methods for random-effects models with $\omega_{3,t}$ as the random effect and ($P_t \times P_t^*$, $P_t \times W_t^*$) as generated regressors. The method of Park and Gupta (2012) adds only $P_t \times P_t^*$ as a generated regressor, and may fail to yield consistent estimates when P_t and W_t are correlated, resulting in the correlation between the random effect in their method and the regressor W_t.

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

$$\begin{aligned} Y_t &= \mu + P_t(\bar{\alpha} + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2}\varepsilon_t + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}) + W_t'\beta_t + \eta_t \\ &= \mu + P_t\bar{\alpha} + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* \times \varepsilon_t + W_t'\beta_t + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} P_t \times \omega_{3,t} + \eta_t, \end{aligned} \tag{A.17}$$

where only one generated regressor, $P_t^* \times \varepsilon_t$, is needed, given which the random effect $\omega_{3,t}$ is independent of all regressors in Equation (A.17).

Both COPE and 2sCOPE can be implemented using the standard methods for random effects models by simply adding generated regressors to control for endogenous regressors. By contrast, the maximum likelihood approach requires constructing a complicated joint

likelihood of $(\xi_t, \eta_t, P_t^*, W_t^*)$, which is not what the standard random effects method uses and thus requires separate development and significantly more computation involving numerical integration.

A.9 2sCOPE for Random Coefficient Logit Model with Correlated and Normally Distributed Regressors

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

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

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

The 2sCOPE approach can be used to address the endogeneity issue using the following

two-step procedure. In the first step, we estimate the model

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

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

$$\widehat{\delta}_{jt} = \mu_j + P'_{jt}\bar{\alpha} + W'_{jt}\bar{\beta} + \xi_{jt} + \eta_{jt}, \tag{A.18}$$

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

Appendix: Additional Results

A.10 Additional Results for Smaller Sample Size

In previous simulation studies in section 4, we use sample size N=1000. We want to further check the robustness of sample size. That is, whether our proposed methods can be applied to smaller sample size. We simulate sample N=200 for T=1000 times, and use the same DGP for continuous endogenous and exogenous regressors as in case 1. Table A.1 shows that 2sCOPE has unbiased estimates for small sample size N=200. Hence, our proposed method is robust and can be applied to very small sample size.

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

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

A.11 Misspecification of ξ_t

Table 2.9 reports estimation results of misspecification of ξ_t using uniform[-0.5,0.5], beta(0.5,0.5) and t(df=3) distributions. We provide estimation results of beta and t distributions with different distribution parameters below.

		OLS			CopulaP&G			COPE			2sCOPE		
Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
μ	1	0.948	0.008	6.928	1.036	0.013	2.909	1.000	0.014	0.022	1.000	0.010	0.009
α	1	1.095	0.006	16.61	1.011	0.011	1.024	0.999	0.011	0.057	1.000	0.010	0.044
β	-1	-1.043	0.005	8.149	-1.048	0.005	9.267	-1.000	0.011	0.004	-1.000	0.007	0.030
$\rho_{p\xi}$	0.5	-	-	-	0.565	0.046	1.414	0.499	0.053	0.010	0.499	0.037	0.025
σ_ξ	0.167	0.144	0.003	7.969	0.168	0.006	0.136	0.167	0.006	0.077	0.167	0.006	0.011

Table A.2: Results of the Simulation Study: Misspecification of ξ_t (Beta(4,4))

		OLS			CopulaP&G			COPE			2sCOPE		
Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
μ	1	0.603	0.059	6.723	1.270	0.115	2.343	0.992	0.126	0.066	0.997	0.080	0.039
α	1	1.727	0.053	13.65	1.095	0.114	0.836	1.006	0.118	0.053	1.006	0.113	0.057
β	-1	-1.328	0.043	7.642	-1.366	0.043	8.554	-0.997	0.090	0.030	-1.002	0.067	0.037
$\rho_{p\xi}$	0.5	-	-	-	0.549	0.059	0.829	0.483	0.065	0.254	0.486	0.047	0.292
σ_ξ	1.291	1.118	0.049	3.506	1.293	0.074	0.030	1.292	0.071	0.008	1.289	0.070	0.032

Table A.3: Results of the Simulation Study: Misspecification of ξ_t (t(5))

A.12 Misspecification of Copula

In the proposed methods, we use the Gaussian copula to capture the dependence structure among the regressors and error term (U_p, U_w and U_ξ). In practice, the dependence might

come from an economic mechanism (such as marketing strategic decisions) and thus might be different from what the Gaussian copula generates. In this section, we examine the robustness of the Gaussian copula in simulated data. Specifically, we generate the dependence among U_p, U_w and U_ξ using copula models other than the Gaussian copula. Specifically, we consider the following T copula models which provide flexible random general generation from arbitrary and heterogeneous correlation structures among more than two variables:

$$C(U_p,U_w,U_\xi)=\int_{-\infty}^{t_\nu^{-1}(U_p)}\int_{-\infty}^{t_\nu^{-1}(U_w)}\int_{-\infty}^{t_\nu^{-1}(U_\xi)}\frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\pi\nu)^d|\Sigma|}}\left(1+\frac{x'\Sigma^{-1}x}{\nu}\right)dx, \quad \text{(A.19)}$$

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

$$\begin{pmatrix}P_t^*\\W_t^*\\\xi_t^*\end{pmatrix}\sim t_\nu^d\left(\begin{bmatrix}0\\0\\0\end{bmatrix},\begin{bmatrix}1&\rho_{pw}&\rho_{p\xi}\\\rho_{pw}&1&0\\\rho_{p\xi}&0&1\end{bmatrix}\right)=t_\nu^d\left(\begin{bmatrix}0\\0\\0\end{bmatrix},\begin{bmatrix}1&0.5&0.5\\0.5&1&0\\0.5&0&1\end{bmatrix}\right) \quad \text{(A.20)}$$

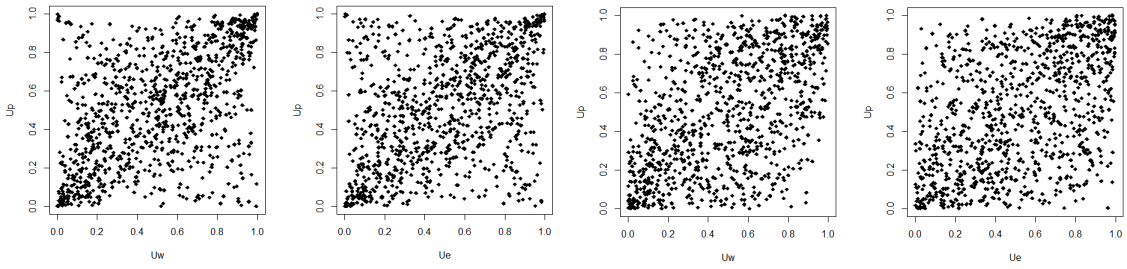
Figure A.1 shows the scatter plots of randomly generated (U_p,U_w,U_ξ) pairs from the above copulas, as well as the Gaussian copula with the same correlation of 0.5. The figure shows disparate dependence structures between U_p and ξ_t (U_p and U_w) for these two copulas.

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

$$\xi_t=G^{-1}(U_\xi)=\Phi^{-1}(U_\xi), \quad \text{(A.21)}$$

$$P_t=H^{-1}(U_p), W_t=L^{-1}(U_w), \quad \text{(A.22)}$$

$$Y_t=1+1\cdot P_t+(-1)\cdot W_t+\xi_t. \quad \text{(A.23)}$$



(a) T Copula ($\rho = 0.5$) **(b)** Gaussian Copula ($\rho = 0.5$)

Figure A.1: Scatter plots of Randomly Generated Pairs U_p, U_w (U_p,U_ξ) for Considered Copulas.

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

Table A.4 summarizes the estimation results. OLS estimates are still biased for all parameters. By contrast, estimates from the proposed COPE and 2sCOPE methods are centered closely around the true values. Therefore, the proposed methods based on the Gaussian copula are reasonably robust to the misspecification of the copula dependence structure among the regressors and the structural error.

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

Table A.4: Results of the Simulation Study Case D2: Misspecification of Copula

Appendix B

Chapter 4 Appendices

B.1 Quality Difference Between Premium and Standard

One thing to note is that on this platform, each driver is not just belongs to a single type. Among the 3,509 drivers, 1,074 drivers with luxury cars are assigned to serve premium riders only, which I call the high-tier drivers. However, the remaining 2,435 drivers with relatively lower quality cars are allowed to serve both premium and standard riders, and I call them low-tier drivers. One possible reason why the platform allows the mixture of driver type is that the platform wants to make better use of each driver's time to provide more supply. Table B.1 shows the car make and service quality difference between the two types of drivers. Both the car price and review rate of high-tier drivers are significantly higher than that of low-tier drivers.

Though there is a quality mixture of the low- and high-quality cars for premium riders, it shouldn't be a big concern for analyzing vertical differentiation. Vertical differentiation can be interpreted from a different perspective. On the rider side, a premium ride can be

Table B.1: Quality Difference Between the Two Types

	High-tier Driver	Low-tier Driver
Car Make	Bentley, BMW etc.	Toyota, Honda, Ford etc.
Average Car Price	$62,445	$29,833
Review Rate (without 0 rate)	4.68	4.60
Review Rate (with 0 rate)	1.26	0.85

viewed as the potential to get a high-quality car, while a standard ride can be viewed as a type to get a low-quality car for sure. In this way, riders still get higher quality value from premium type, and the market can still be differentiated. On the supply side, the two products can be viewed as competing for drivers' available time instead of the number of drivers. The larger the demand size, the more time a mixed driver would spend serving that type because of the larger network value. In this way, this setting can still be a good setting to study vertical differentiation and quantify the network externalities of the two products by analyzing riders' and drivers' choices.

B.2 Data Preparation

B.2.1 Rider's Expected Waiting Time

I first construct rider's expected waiting time. Because of the distinct feature network externalities in two-sided markets, network size would add value to riders by affecting the time they have to wait for a car. Though riders cannot observe waiting time directly, they can observe the real traffic environment and would form expectations on the time they have to wait. In this data set, the time when the order is placed by rider and the time when it is confirmed by driver, if available, are observed by researchers. Once an order request is confirmed by a driver, the estimated driver arrival time is further observed. I use those time points to construct and approximate the rider's expected waiting time in a specific time and location.

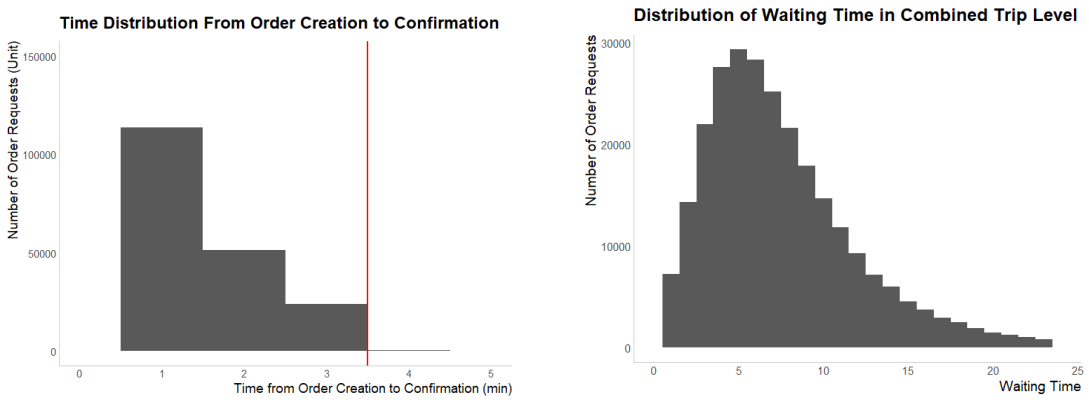
There are mainly three steps for the construction of rider's waiting time. First, there's a significant number of reorders occur in this platform, and I will first find a way to combine them. Reorder is defined as orders placed by the same rider in the same location within a very short period. There can be multiple reasons why those reorders occur. For example, the rider might fill in wrong information; the order request might expire without any response from the driver, etc. Since waiting time is a key quality dimension that differentiate riders to choose different type of rides, it's crucial to capture rider's actual waiting time for a ride. Figure B.1(a) shows the distribution of time from order creation to confirmation. There's a clear cutoff at 3 minutes, which means order requests would expire after 3 minutes and would be cancelled automatically if no response is received from drivers within this short period. Then riders might place an order again or leave. I observe that reorders account for 27% of all orders in this data set. If I just treat each order request as an independent trip, waiting time for the trip would be downward biased. To solve the problem, I combine those reorders into one trip and use the aggregate time from the creation of the first request to the final driver estimated arrival time as the rider's total waiting time. I set several criteria to define reorders. An order is defined as a reorder if

a. the rider's last order request was not realized and

b. the time interval from the last request is shorter than 15 minutes and

c. the pick-up location the rider provides is the same with last request.

After I change requests to trip-level requests, I calculate rider's waiting time for a trip as the aggregate time of all requests from order creation to driver estimated arrival time if the trip is finally confirmed. Figure B.1(b) shows the distribution of the total waiting time from order creation to driver estimated arrival in a combined trip. After combining to trip-level request, total number of orders decreases from 1.2 million to 766,471, but percentage of realized trips increases from 37.7% to 59.4%. Moreover, trips with more than one request



(a) Order-level Wait **(b)** Payment with Cruise Time

Figure B.1: Waiting Time Distribution.

only accounts for 27.3% of all trips, and among those trips with multiple requests, 84% of them don't include requests for both the types of requests, which means that most riders wouldn't change their idea of what type to choose when they make multiple requests. This gives us a clear setting to calculate aggregate waiting time for each trip.

Second, after the construction of waiting time above, another problem is that the waiting time of trips that are not confirmed by any driver is missing. I approximate and fill in those missing waiting time using trips with available waiting time in a similar environment. Below shows the steps how I construct a market and fill in the missing waiting time with the observed waiting time in the same market.

1. I first divide Manhattan area into 29 smaller locations. The 29 neighborhoods are divided according to a common demarcation of Manhattan, which is visualized in Figure 4.1;

2. I assume traffic environment in a specific hour in weekdays are similar, and define market in the location-week-hour level. That is, I treat observations in the same week and at the same location and hour from Monday to Thursday as in the same market. This is a reasonable assumption, as both the price patterns and people's behavior are

the same from Monday to Friday, while are very different on Friday, Saturday, and Sunday. For example, traffic in 7:00 am of Monday and Tuesday should be similar, and traffic at the same in Monday and Saturday should be quite different. Thus, to keep it clean, I define market in the above way and just use observations from Monday to Thursday for analysis. In the end, there are 5,672 markets in total;

3. Once I group together trips with similar environment, I fill in the missing waiting time of a trip with the maximum waiting time among same-type trips in the same market. After this step, all trips have waiting time;

Third, after we obtain waiting time for each trip, we further constructed rider's expected waiting time. The waiting time constructed above is ex post waiting time, which is obtained after the trip is fulfilled. However, when riders make order request, they cannot observe waiting time in advance, and thus I continue this step to further construct ex ante expected waiting time for each type of ride by averaging the waiting time of all trips with the same type in the same market.

B.2.2 Driver's Cruising Time

Once a driver receives an offer, he will observe the distance from the rider and thus can infer how long he has to cruise to pick up the rider. In this section, I focus on how to construct cruising time using observed distance.

What I can observe in data is drivers' distance from rider for all received offers, and driver's expected arrival time for offers that are accepted by drivers. I first get drivers' cruising time for those realized trip by calculate the time between driver's expected arrival time and the time he accepted the offer. Then I regress the cruising time on distance and market fixed effects such as week, hour and location, and then use the regression result to

predict cruising time for those unrealized trips. The construction of drivers' cruising time is simpler than rider's waiting time. For cruising time, I don't need to calculate market-level expected cruising time, as drivers can observe the location of the rider and the distance upon receiving each offer.

B.2.3 Construction for Outside Demand and Counterfactual Options

On this platform, I can only observe order requests from riders who choose this platform, and cannot observe options from those who choose outside options. In real situation, riders may consider more options besides the two options in this platform. Moreover, for those who choose this platform, we researchers can only observe the information of the chosen option, while the information of the one not chosen is not available. But riders can observe information for each type at the time when placing an order.

Thus, to better capture and model rider's behavior, I include taxi data in New York City as outside option for estimation to control market size, and estimate rider's choice decision taking outside option into consideration. The source of taxi data is public online [1] [2]. Meanwhile, I fill in the missing information of those options not chosen for all possible demand, including the outside demand. The taxi data in New York City is public online. It includes all realized trips with trip start time, trip fare, distance and geographic information. I first match taxi demand with the demand using geographic location to get demand size in each market, and then simulate the price of the two options in this platform for those outside demands. To construct the counterfactual choice for all demand, I fill in the missing order-level variable (e.g., price) of the option not chosen. According to the price structure listed in Table 4.1, price for standard is a fixed price $8.98, while price for premium depends on trip duration and trip distance. In the dataset, I can observe trip duration for realized trips,

[1] https://data.cityofnewyork.us/dataset/2016-Yellow-Taxi-Trip-Data/uacg-pexx;
[2] https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb.

and origin and destination for all trips, thus can get relationship between trip price, duration and distance using realized trip and then predict trip price for counterfactual option. I list the steps how I get outside demand in each specific market and how I predict and simulate price for counterfactual option below.

1. Download Green and Yellow taxi data in 2016 at New York City online, in which information such as price, trip duration and trip distance are listed for each realized trip;

2. Get the market index for each taxi trip using trip time and geographic information (latitude and longitude), and calculate taxi demand in each market (week-hour-location). To save time for MCMC code running, I randomly sample 10% taxi demand as the outside demand for estimation;

3. For unrealized trips in this platform, I approximate trip distance and duration using trips with same origin and destination in taxi data. I define two trips to have the same origin and destination if the rounded latitude and longitude with 2 decimal places are the same.

4. Regress price on trip distance and duration using the realized trips in this platform, and get the relationship between price and trip distance and duration;

5. Use the relationship in step 4 to predict price for the option not chosen or trips with missing price;

6. Fill in waiting time for the option not chosen with market-level waiting time of that type.

B.3 Bayesian MCMC Estimation

In my model, I have individual specific coefficients and unobserved demand and supply shocks. For model estimation, I use Bayesian method and treat those random coefficients and unobserved shocks as latent variables instead of using frequentist approach and integrating those parameters out. Analysis proceeds by iteratively sampling from the full conditional distributions of all model parameters, including those latent variables. The estimation process can be written in hierarchical form:

$$Y^d|X^d, Ewait_{km}, \theta_i, \zeta_{km}, \lambda_i, \mu_w, \varepsilon \quad \text{Observed demand,} \tag{B.1}$$

$$Y^s|X^s, D, S, \theta_j, \zeta_m, \lambda_j, \mu_c, \varepsilon \quad \text{Observed supply,} \tag{B.2}$$

$$Ewait_{km}|D_{km}, S_{km}, \gamma, \mu_w, \mu_c \quad \text{Market-average waiting time} \tag{B.3}$$

$$\theta_i, \theta_j|\bar{\theta}, I_i, I_j, \Sigma_\theta \quad \text{Heterogeneity,} \tag{B.4}$$

$$\varepsilon, \quad \text{Extreme value error} \tag{B.5}$$

where $\theta_i = (\beta_i^d)$, $\theta_j = (\beta_j^s)$ are rider and driver heterogeneous parameters, and I_i, I_j are their demographic information. Observed demand and supply are dependent on rider's and driver's coefficients (θ_i, θ_j), demand and supply shock ζ, unobserved error (ε) and explanatory variables. The conditional demand and supply density, given θ and explanatory variables is multinomial-logit probabilities (see Equation (4.7)). Rider and driver coefficients are specified as random coefficients and depend on demographics. The joint density of all model parameters is then

$$\begin{aligned} &f(\theta_i, \theta_j, \zeta, \bar{\theta}, \Sigma_\theta|Y_d, Y_s, X, I) \\ &\propto \Pi_i \Pi_j \Pi_t \Pi_m \text{prob}(Y_d, Y_s|X, \theta_i, \theta_j)\pi_1(\theta_i, \theta_j|I, \bar{\theta}, \Sigma_\theta)\pi_2(\bar{\theta}, \Sigma_\theta) \end{aligned} \tag{B.6}$$

Estimation is carried out using a Markov chain Monte Carlo procedure that involves generating a sequence of draws from the full conditional distributions of the model (see Gelfand

and Smith, 1990; Gelfand et al., 1990).

Estimation steps:

1. Generate θ_i, θ_j for (i = 1,2,..., I; j = 1,2,..., J)

$$[\theta_i, \theta_j | *] \propto \Pi_i \Pi_j \text{prob}(Y_d, Y_s | X, \theta_i, \theta_j, \zeta_m)(\theta_i \sim MVN(\bar{\theta}_i, \Sigma_{\theta i}))(\theta_j \sim MVN(\bar{\theta}_j, \Sigma_{\theta j}))$$

   I use Metropolis-Hastings algorithm with a random walk chain to generate draws of θ_i, θ_j.

2. Generate $\bar{\theta}, \Sigma_\theta$ using Gibbs sampling.

   Note that random-coefficients for rider and driver are generated and updated separately. Here I only update coefficients for rider, $\bar{\theta}_i \Sigma_{\theta i}$, as an example.

$$\begin{aligned}
[\bar{\theta}_i | \theta_i, D_i, \Sigma_{\theta i}] &\sim N((\widetilde{D}_i{}' \widetilde{D}_i + A)^{-1}(\widetilde{D}_i{}' \widetilde{\theta}_i + A\beta_0), (\widetilde{D}_i{}' \widetilde{D}_i + A)^{-1})) \\
\widetilde{D}_i &= U^{-1\prime} D_i \\
\widetilde{\theta}_i &= U^{-1\prime} \theta_i \\
\Sigma_{\theta i} &= U'U \\
[\Sigma_{\theta i} | \bar{\theta}_i, \theta_i] &\sim \text{Inverted Wishart}(\Sigma_{i=1}^{I}(\theta_i' - \bar{\theta}_i D_i)(\theta_i' - \bar{\theta}_i D_i)' + V_0, I + n_0)
\end{aligned}$$

   suppose the priors of $\bar{\theta}, \Sigma_\theta$ are $\bar{\theta} \sim N(\beta_0, A^{-1}), \Sigma_\theta \sim IW(V_0, n_0)$.

3. Generate γ using Gibbs sampling.