ESSAYS ON APPLIED ECONOMETRICS

A Dissertation

by

LI ZHENG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Qi Li
Committee Members,	Yonghong An
	Silvana Krasteva
	Ximing Wu
Head of Department,	Timothy Gronberg

May 2020

Major Subject: Economics

ABSTRACT

This dissertation consists of three essays on applied econometrics.

The first essay is entitled *A Structural Analysis on US Spectrum Auctions*. The spectrum auction allocates spectrum licenses to companies. This paper provides a structural analysis on US spectrum auctions to estimate bidders' values, which is essential for auction policy evaluations. I first perform a theoretical analysis, then construct a structural model to rationalize bidders' bidding behaviors as a bundle choice problem. I propose a multiple-step estimation to recover the parameters in bidders' value function from the model. In the estimation, I develop a framework to handle the high-dimensionality issue in the bundle choice model with individual-level data. This paper analyzes the 1995-1996 spectrum auction in the US. I find evidence of complementarity in this auction, as well as heterogeneity in the complementarity valuation across bidders.

The second essay is entitled *Optimal Model Averaging of Mixed-Data Kernel-Weighted Spline Regressions*, and it is coauthored with Qi Li and Jeffrey S. Racine. Model averaging has a rich history dating from its use for combining forecasts from time-series models and presents a compelling alternative to model selection methods. We propose a model average procedure defined over categorical regression splines. Theoretical underpinnings are provided, finite-sample performance is evaluated, and an empirical illustration reveals that the method is capable of outperforming its nonparametric peers in applied settings.

The third essay is entitled *Multivariate Density Forecast Combination*. Density forecasts are able to convey the uncertainty in addition to the point forecasts, and multivariate density forecasts further allow people to capture the interdependency among different variables of interest. This paper develops a class of combination schemes for multivariate density forecasts, in view of that the forecast combination could effectively improve the forecast performance upon single forecasts. I prove the asymptotic optimality of the estimated combination weight. Monte-Carlo simulations are provided to demonstrate the theoretical results.

DEDICATION

To my parents.

ACKNOWLEDGMENTS

I am deeply indebted to Qi Li for his invaluable guidance and assistance during my doctoral studies. I am forever grateful to Yonghong An for his support throughout the course of my research. I benefited greatly from the help and encouragement from Silvana Krasteva. I thank Zheng Fang, Tatevik Sekhposyan, and Ximing Wu for their helpful comments.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Qi Li (chair), Professor Yonghong An, and Professor Silvana Krasteva of the Department of Economics, and Professor Ximing Wu of the Department of Agricultural Economics.

The analyses depicted in Chapter Two were conducted in part by Professor Qi Li and Professor Jeffrey S. Racine.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by the Department of Economics of Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

This dissertation concerns developing econometric methods for studying economic issues. The first essay presents a structural approach to analyze US spectrum auctions. The second essay establishes model averaging in the mixed-data environment. The third essay provides a forecast combination scheme for multivariate density forecasts.

In the first essay, I provide a structural analysis on the US spectrum auctions. The spectrum auction allocates spectrum licenses to companies. The evaluation of its efficiency and revenue calls for a structural approach to recover bidders' values from the empirical data. The difficulty arises because the spectrum licenses are heterogeneous and complementary. In this article, I analyze the 1995-1996 C-block spectrum auction in the US. To begin with, I perform a theoretical analysis to motivate the subsequent modeling. Next, I construct a structural model to capture bidders' bidding behaviors in the auction, which are rationalized as a bundle choice problem. Specifically, bidders choose the bundle of licenses to bid by maximizing their current-round expected payoffs. I propose a multiple-step estimation procedure to recover the structural primitives in the model, the parameters in bidders' value function. In particular, I first estimate the winning probabilities for the purpose of obtaining the current-round expected payoffs. Then I estimate the high-dimensional bundle choice problem with individual-level data, where the high-dimensionality stems from the large number of bundles. I find strong evidence of complementarity in the auction. The complementarity of the nationwide bundle is worth 8 billion dollars for an average bidder, which equals 59.54% of the sum of final prices of all licenses. Moreover, I explore the bidder heterogeneity in the complementarity values and find that large bidders value the complementarity higher than medium and small bidders. The stand-alone values are mostly reflected by the license-characteristic, rather than the bidder-characteristic.

In the second essay, we study model averaging in the mixed-data environment where both continuous and categorical covariates are present, using regression spline models as candidates. Model averaging is an appealing tool to deal with model uncertainty and presents a compelling alternative

to model selection methods. Our target is estimating the conditional expectation. Practitioners who adopt model averaging often construct a weighted average defined over a set of parametric candidate models. Our approach adopts a nonparametric perspective that allows people to approximate a wide range of data generating processes. In addition, we admit both continuous and categorical predictors in the model for the full flexibility of empirical uses. We estimate each candidate model using a regression spline method. We combine the candidate model with a weighted average and estimate the optimal weight using the Mallows criterion. We demonstrate the asymptotic optimality of the selected weight, in the sense that we attain the minimized predictive loss as if we knew the infeasible optimal weight. We take into account the heteroskedasticity and consider both cases where the error structure is known and unknown. Simulation studies support the superiority of our method over the extant ones. We illustrate our approach using an empirical dataset.

In the third essay, I provide a forecast combination for the multivariate density forecasts. Density forecasts are more and more popular because they could communicate with users about the uncertainty around the point forecasts. Multivariate density forecasts achieve an extra advantage that they are able to reveal the interactions among variables of interest. This paper aims to improve the multivariate density forecasts via forecast combination. Specifically, I combine the different multivariate density forecasts via a weighted average. The estimation of optimal weights relies on the Probabilistic Integral Transformation and the Kullback-Leibler Information Criterion. I show that the selected weight is consistent, and use simulation to demonstrate the validity of my theoretical results.

2. A STRUCTURAL ANALYSIS ON US SPECTRUM AUCTIONS

2.1 Introduction

The wireless communication has deeply changed our lives, and will be continually influencing the world in the upcoming 5G era, and in the future. In many countries and regions, the *spectrum* usage of the wireless service is allocated to the telecommunication companies via *auctions* by the government. The spectrum auctions generate large revenue for the government, and due to its relevance in the economy and society, its allocation effectiveness also concerns the public [1]. In view of that, the choice of auction formats and design of auction rules are crucial. However, there exist different auction formats for spectrum auctions in different countries, for example, the Simultaneous Ascending Auction (SAA) conducted in US [2], and the Combinatorial Clock Auction (CCA) held by some other countries [3].[1] Even for a same format, for example, the SAA in US, there also exist various versions in terms of detailed auction rules. Therefore, quantifying the performance of spectrum auctions becomes an important issue for the government and policy makers.

To achieve the policy evaluation in the spectrum auction, which is an exquisitely designed market mechanism participated by sophisticated players [5], we need a structural analysis to recover *bidders' private values*. In this paper, I analyze the Simultaneous Ascending Auction format, which has been the baseline design for the US spectrum auction since its launch into practice in 1994. The auction sells many licenses simultaneously, where each license represents the right to transmit signals for this band of electromagnetic spectrum of a specific geographic area. SAA is a multiple-round process. As rounds evolve, bidders raise the prices on the licenses, and the auction stops until no new bids appear on any license. The licenses are *heterogeneous*, since licenses for different areas are valued differently by the bidders. In addition, licenses may be *complementary*, in the sense that for a bidder, the value for a bundle of licenses is higher than the sum of individual

[1] Countries adopting CCA for spectrum auctions include Australia, Austria, Canada, Denmark, Ireland, the Netherlands, Slovenia, Slovakia, Switzerland, and the United Kingdom [4].

values for each license in this bundle. This is because bidders might be more willing to obtain a group of licenses, potentially due to the saving in fixed costs and economies of scale [6].[2] The *heterogeneity* and *complementarity* of licenses, along with the *large number* of licenses, make it infeasible neither for the economists to solve for the equilibrium of the model [7], nor for the bidders to calculate their optimal strategy [5].[3] Accordingly, learning the bidders' private values from this complex mechanism is challenging.

This paper aims to solve this problem. So far, the evaluation of spectrum auction designs has mainly come from the economic theories and laboratory experiments. Explorations of this question using empirical data, in spite of its importance, remain relatively sparse. I study the US spectrum auction of the licenses for the C block of the 1900 MHz PCS band in 1995-1996, which has raised 13.43 billion dollars in total.[4] There are 493 licenses to be sold, with 255 bidders participating in the auction. It lasts for 184 rounds, from December 1995 to April 1996.

In this article, I first perform a theoretical analysis to guide the subsequent structural modeling. Due to the infeasibility of the full equilibrium for SAA, I study a multiple-object clock auction (MCA) model with complementarity, which could be viewed as an approximation of SAA [9]. On the one hand, I derive the Bayesian Nash equilibrium (BNE) of MCA, and prove some useful equilibrium properties in an attempt to understand the bidding strategies in SAA. I find that the BNE bidding strategy of MCA is non-decreasing in bidders' private values, and non-increasing in the number of remaining bidders in the auction, which corresponds to a piece of the current-round information in SAA. On the other hand, using numerical solutions for the BNE, this analysis demonstrates the existence of the *exposure problem*, which is widely documented for SAA [10]. The exposure problem means that when a bidder is bidding on a bundle of licenses, she is exposed to the risk of obtaining only a subset of the bidding set, and thus tends to bid lower than her true value. I find that in MCA, bidders' BNE strategy is indeed lower than their true values. Furthermore, the extent of the exposure problem is associated with the number of bidders. Such

[2]For example, the cost in building radio stations for two adjoining areas will be lower than for two distant areas. And mobile service companies with a wide range of coverage will attract more consumers.

[3]The spectrum auctions typically have a large number of licenses. See Section 2.2.

[4]It is also referred to Auction 5 by FCC. [8] analyzed the same auction as in this paper.

difference from the single-object clock auction (where bidding the true value is a weakly dominant strategy) results from the complementarity among objects.

Next, I develop a structural model to capture bidders' bidding behaviors revealed from the theoretical insights and preliminary data explorations. While maintaining tractability, the model is able to rationalize bidders' decisions at each round based on their private values and current-round information. Since in the data most submitted bids equal the minimum acceptable bids, we model the a bidder's decision at a given round as the choice of *bidding bundle*: the set of licenses she will place bids on. In particular, at each round, a bidder chooses the bundle that maximizes her *current-round expected payoffs*. The current-round expected payoff on a bundle are constructed using the bidder's beliefs about the probability of provisionally winning each license in this bundle at the current round, using the current-round price. The winning probabilities depend on their private values and the current-round information. This current-round expected payoff reflects the expected profit of a bidder if the auction ends at the current round. The bidding strategy is inherited from the Straightforward Bidding (SB) strategy in [2]. However, this model differs from SB in that I incorporate the current-round information into bidders' decision making, leading to the explicit underbidding behaviors which respond to the exposure problem, and is thus consistent with our theoretical results. I specify the equilibrium concept for the structural model, requiring that bidders' beliefs are consistent with the true winning probabilities. [11] use a similar way to represent the expected utility of bidders, whereas they consider bidders' beliefs of winning a license at the end of the auction.

Finally, I propose a four-step estimation scheme for the structural model and recover the parameters in bidders' private value function. Remarkably, within this procedure I provide a general framework for estimating the high-dimensional bundle choice problem with individual-level data. I parametrize the bidders' private value to be composed of the *stand-alone values* and the *complementarity values*. Firstly, since bidders' decisions are based on their current-round expected payoffs, we need to estimate the winning probabilities. Subsequently, we are faced with a bundle choice problem based on the current-round expected payoffs, where the total number of bundles

is extremely large. There are 480 licenses in the estimation, resulting in 2^{480} bundles in total.[5] I apply the semiparametric estimation method for multinomial discrete choice proposed by [12] into our bundle choice problem. There are two essential properties of this approach. On the one hand, it allows for arbitrary dependence across different choices, which particularly fits our case because bundles containing the same licenses are correlated. On the other hand, one can use a *random projection* technique to achieve dimension reduction on the number of choices, which is proposed by [13]. In addition, the method is designed for panel data, which corresponds to our data structure. However, [12]'s method demands the choice probabilities for each bidder at each round, while we only observe the realized chosen bundle. In view of that, we need to estimate the choice probabilities beforehand.

Below I provide an overview of the estimation steps. In step 1, I estimate the winning probabilities for each license, for the purpose of constructing the current-round expected payoff. I use the current-round provisionally winning results for the bidding bundle of a bidder, i.e. whether she wins the licenses she bids on, as the outcome variable, and use her private value as well as current-round information as covariates. A challenge is that, the size of bidding bundle is usually much smaller than the size of the full set of licenses, while we need the estimation of winning probabilities for all licenses.[6] This calls for a more accurate prediction approach than the traditional ones. To this end, I first fit six models, including the traditional logit model, and five machine learning classification methods: LASSO, Random Forest, Boosting, Support Vector Machine, and Neural Nets. Next, I combine these models using the model averaging method to further enhances the predictive performance. The combined estimation model improves the prediction accuracy by reducing the (cross-validation) misclassification rate by 2.6%.

In step 2, I estimate the choice probabilities for each *bundle of licenses*, which is an input for the estimation in the last step.[7] The major difficulty is that, if we regard each bundle as a discrete choice, then the direct application of the traditional multinomial discrete choice models would be

[5] Attention is limited to the licenses in the continental US.

[6] See Table 2.3. The average size of bidding bundle is 5.2.

[7] Note that the sequence of step 1 and step 2 does not matter for the final result.

computationally expensive or even infeasible. To address this issue, I view the choice of bidding bundle as a multivariate Bernoulli (MVB) variable, and instead of modeling the joint distribution of the MVB, I model the (univariate) conditional distributions in a logit form following [14]. Involving only the conditional probabilities in the likelihood function for MLE still produces consistent estimators, while greatly accelerates the computation. This model of conditional probabilities implies a multinomial logit model for bundles, and thus we can estimate the choice probabilities using the estimators from MLE. Notably, since the multinomial logit model corresponds to a random utility model, we obtain the *current-round decisional utility* (containing private values and current-round information) for the bidders, which could be interpreted as that bidders use it to decide which bundle to bid on at each round. Such an interpretation is also consistent with my theoretical analysis that bidding strategies are associated with current-round information.

In step 3, I proceed to deal with the high-dimensionality issue in the bundle choice problem. Now, our data is bidders' current-round expected payoff estimated from step 1, and their choice probabilities estimated from step 2, for each bundle. Since we have 480 licenses in total for estimation, the number of all possible bundles is 2^{480}. I first restrict the attention to the bundles that has once appeared during the whole course of the auction. This gives us 3998 different bundles of licenses. Nevertheless, this number is still far from feasible for the estimation approach by [12]. Next, I apply the method proposed by [13], which exploits the *Random Projection* (RP) technique from machine learning to reduce the dimensionality.[8] The RP method is to premultiply the original d-dimensional data by a $k \times d$ random matrix, and then obtain a k-dimensional projected-down data. The reason RP could be used to reduce dimensionality is that RP preserves the Euclidean distance between data vectors to the projected-down subspace with high probability, and the optimization objective function in [12] only involves Euclidean distance among data points.

In step 4, we use the projected-down data, and apply the semiparametric estimation of multinomial discrete choice with panel data proposed in [12]. This method makes use of the *cyclic monotonicity* property of the choice probability function to construct inequalities for estimation.

[8] If I skip the random projection step, the computational time is 150 times of that with random projection.

It allows arbitrary interdependence of the unobserved terms among different bundles and across different rounds, which is essential for the bundle choice problem. In addition, it does not need to assume any distributional form of the error term, and it allows for bidder fixed effect in the panel data environment. Notice that step 2 to step 4 is a complete procedure of estimating the high-dimensional bundle choice problem with individual-level data, where the high-dimensionality refers to the large number of single objects.

I analyze Auction 5 of the US spectrum auction, which is held by FCC in 1995-1996. From the structural estimation, I find a large and significant effect of the complementarity on bidders' private values. The complementarity of the nationwide bundle is worth 8 billion dollars for an average bidder, which equals 59.54% of the sum of final prices of all licenses. Suppose a bidder with average eligibility wins the nationwide bundle, the complementarity contributes 24.46% of the private value. For the bidders' stand-alone values, on the one hand, I document large and significant effect of the license-characteristic. For an average bidder, a license with 1 more million population, will be valued 98.71 million dollars higher. On the other hand, we find small and insignificant effect of the bidder-characteristics. Therefore, the variation of the stand-alone values is mostly generated by the licenses. Moreover, I explore the bidder heterogeneity in the complementarity values, and find that large bidders value the complementarity higher than medium and small bidders.

Structural analysis on spectrum auctions is sparse in the literature. [8] use a matching approach to estimate the value function parameters of bidders. They assume that the final outcome of matches between bidders and licenses is pairwise stable, and only use the data of final allocation and prices of the auction for estimation. I analyze the same auction as [8], while taking a different structural modeling and estimation method. Specifically, instead of focusing on the outcome at the end, I model bidders' bidding behaviors at each round, and make use of the information during the full course of the auction. [15] adapt [8]'s method to the Canadian spectrum auction to estimate the implicit cost. In terms of the structural model, this paper shares similar spirits with [11], where bidders also maximize their expected payoffs at each round. This paper differs from [11] in two folds. On the one hand, for the structural model, [11] assume that bidders form beliefs of winning a

license at the end of the auction, which requires that bidders can perfectly foresee the final results, potentially far from the present round. Instead, in my model bidders have consistent beliefs of the current-round winning probabilities, which is more realistic given the complex bidding environment, and the current-round expected payoff has an interpretation close to [2]. On the other hand, their estimation approach is based on several behavioral assumptions which are necessary conditions of the structural model, as in [16]. In contrast, my structural estimation comes directly from bidders' optimization problem over bundles of licenses, and thus makes use of the full information of bidders' decisions. This comes with the cost of the high dimensionality issue in the estimation.

The high dimensionality issue has attracted much attention in the multinomial discrete choice models. [13], which is our closest predecessor, leverage the random projection method to reduce the dimension of discrete choices using aggregate level data. Other approaches for dealing with high dimensionality in multinomial discrete choice models include [17] and [18] that rely on taking subsets of choices for estimation; [19] who use a Bayesian method to manage the large number of parameters; [20] who propose a machine learning model of demand for bundles with sequential search; and [21] with indirect inference estimation. For the high dimensional bundle choice problem, the literature is still at the growing stage, but starts to be more and more appealing. [22] use a novel demand inverse to estimate demand for bundles with a large number of goods. I add to the literature by providing an estimation approach of bundle choice models with many objects, where only individual-level choice data is observed. Similar economic problems and data structures also rise in, for instance, the retailing data.

The literature in the combination of machine learning and structural econometrics or industrial organization is rapidly growing. Our estimation of winning probabilities follows [23], who apply several machine learning methods to estimate consumer demand, accompanied with model averaging of all the considered methods. [24] estimate the consumer preferences using probabilistic models of matrix factorization. [25] use LASSO to handle the high dimensionality in BLP models where covariates are rich. Our practice shows that economists and econometricians can beneficially leverage the power of the machine learning in the structural model, especially when

(1) the estimation involves predictions as intermediate steps; and (2) we are faced with the high dimensionality issue.

The literature on theoretical analysis of spectrum auctions and SAA is rich. I summarizes the related literature and the contribution of my theory model in Section 2.3.1. There are also extensive exploration of spectrum auction from the perspective of lab experiments, including [26], [27], among others.

The rest of this essay proceeds as follows. Section 2.2 introduces the basics of spectrum auctions and the auction rules of the Simultaneous Ascending Auction design. In Section 2.3 I conduct a theoretical analysis and provide guidance for the following structural modeling. Section 2.4 describes the data we are studying, and perform preliminary analysis which motivates our structural estimation. Section 2.5 presents my structural model. Section 2.6 elaborates the estimation approach step by step, and shows the estimation results. Section 2.7 discusses the bidder heterogeneity. Section 2.8 concludes.

2.2 US Spectrum Auctions

In this section, I first review the history and current status of the spectrum auctions, along with a summary statistics of the US spectrum auctions. Next, I introduce the basic auction rules of the US spectrum auction, which uses the Simultaneous Ascending Auction (SAA) design.

2.2.1 An Overview of US Spectrum Auctions

The first spectrum auction in US was launched in 1994 held by the Federal Communications Commission (FCC). It not only pioneered the spectrum auctions all over the world, but also opened the era of putting auction theory to work [7]. Before, the spectrum rights in the US and many other countries were assigned using non-market mechanisms, such as comparative hearing (also known as “beauty contests") or lottery. In 1994, the Simultaneous Ascending Auction (SAA) designed by Preston McAfee, Robert Wilson, and Paul Milgrom, was adopted by FCC and created the first of the large modern auctions. Since then, SAA has been popularized worldwide for spectrum auctions

and earned the compliment “The Greatest Auction Ever".[9]

Nowadays, the world is entering the 5G era. The first 5G spectrum auction is just fulfilled in January 2019 in the 28 GHz band (Auction 101), with more auctions for other bands taking place or being prepared for auctioning. Other countries, for example, German and China, have also started the allocation of 5G spectrum rights.

So far, the FCC has conducted 89 spectrum auctions since 1994 until 2018, with total revenue being over 120 billion dollars. I summarize the primary statistics for the auctions in Table 2.1, where the data is from FCC.[10] We see that on average, the number of licenses is large, where for some auctions it is extremely large. Therefore, oftentimes analysts are faced with an auction with *many heterogeneous* licenses sold simultaneously, which is quite different from the traditional single-object auction. In addition, the number of rounds and number of bidders are also large, making the game-theoretical analysis of this auction considerably hard, and the bidding strategy for the bidders very complicated. Consequently, people are demanding an appropriate framework to conduct empirical analysis on the spectrum auction.

Table 2.1: US Spectrum Auctions: 1994-2018

	Average	Median	Min	Max	Std	Obs
Gross Bids ($\times 10^9$)	1.4610	0.0148	2.50e-5	44.8994	5.7819	89
Net Bids ($\times 10^9$)	1.3592	0.0136	2.50e-5	41.32	5.3743	89
Winners	31.47	12	1	182	40.28	89
Bidders	55.75	25	2	456	77.68	89
Licenses Won	519.75	90	1	5323	1084.04	89
Licenses FCC	114.38	0	0	4889	612.42	89
Licenses Total	634.13	96	1	9603	1481.80	89
Rounds	72.56	44	1	341	74.13	88

[9] William Safire, “The Greatest Auction Ever", *New York Times*, March 16, 1995.

[10] The observation of Rounds is 88, because Auction 2 used the “Oral Outcry" design.

2.2.2 Auction Rules

The spectrum auction sells multiple objects simultaneously, where an object is the license of spectrum for a geographic area. The auction format is referred as Simultaneous Ascending Auction, which is a multiple-round process. At each round, bidders simultaneously submits sealed bids on the objects they are interested in. After the end of a round, the highest bid for each license is referred to *standing high bid* [2]. The corresponding bidder is called *standing high bidder*.[11] If nobody has ever bid on an object, then the standing high bidder is defined to be the auctioneer. The minimum acceptable bids at next round are computed as the standing high bids plus some smallest increment.[12] The auction proceeds in an ascending way in the sense that at each round any submitted bids should be higher than the minimum acceptable bids.

At the end of each round, round results are posted. These results include all new bids and the corresponding bidder identities, the standing high bids and the standing high bidder identities, as well as the minimum acceptable bids. The auction stops at a round when no new bids are submitted for any objects in this round. The licenses are allocated to their standing high bidders, at the standing high bids.

The standing high bidder is allowed to withdraw her standing high bid for an object. If the standing high bidder withdrawed at an object, then the high bid of this object becomes the second highest bid, and the standing high bidder becomes the corresponding bidder. Such permission of “regret" serves an alleviation of the widely discussed exposure problem of SAA. Meanwhile, it risks generating irrational tentative bids and strategic collusive behavior (intimidating bids). Therefore, a penalty is designed to associate the withdrawls: if the final price of the object is less than the withdrawn bid, the withdrawing bidder must pay the difference (otherwise the penalty is just zero). In the empirical analsis, I consider the after-withdrawl bidding data so as to ignore the strategic uses of withdrawls.

[11] In the event of tie bids, the standing high bidder will be identified by the order of the bids received by FCC, starting from the earliest bid

[12] In Auction 5, the smallest increment is equal to the greater of $0.02 per bidding unit, or 5% of the standing high bid.

Before the auction, the auctioneer posts the “bidding unit" of every license, roughly measuring the value of it. In practice, the bidding unit is the population in this area. The eligibility, measured by bidding units, is to provide an upper bound of each bidder's bidding activity at each round. Before the first round, each bidder is required to submit its *initial eligibility* by making a corresponding deposit. In the C block of Auction 5, the eligibility payments were 1.5 cents per MHz-individual. Afterwards, a bidder's *current-round eligibility* at each round is evolved according to the *activity rule*.

To accelerate the auction and ensure that it ends in a reasonable amount of time, an *activity rule* is introduced to SAA by Paul Milgrom. At the end of round, a bidder is considered *active* in an object if she places a bid on this object at round t, or she was the standing high bidder of this object at last round. A bidder's *activity* at each round is the sum of bidding units of the objects in which she is active. In a round, a bidder's activity cannot exceed her eligibility at this round. What's more, the eligibility at the next round depends on the current round's activity: if a bidder's current activity is no less than a prespecified fraction of her current eligibility, then her eligibility at next round remains unchanged; otherwise, the eligibility is reduced by a proportion, until she is no longer eligible for bidding any license.[13]

In Auction 5, FCC designed different levels of activity rules in three different stages during the auction. The transition from Stage 1 to Stage 2 and finally to Stage 3 was determined by the aggregate level of the bidding activity, subject to FCC's discretion.[14]. The transition is irreversible. In Stage 1, a bidder who wishes to maintain its current eligibility is required to be active on licenses encompassing at least 60% of the bidding units for which it is currently eligible. Failure to maintain the requisite activity level will result in the reduction of eligibility for the next round, which amounts to 5/3 of the current round activity. For Stage 2 and 3, these two numbers become 80%, 5/4, and 95%, 20/19, respectively. In other words, the activity rule would be more and more

[13]In the design described in [2], the reduction proportion is different for different stages during the auction, which is determined before the auction begins.

[14]The transition rules may depend on several measures, which are not informed to the bidders. These measure may include, for example, the auction activity level, which is the sum of activity units of those licenses whose high bid increased in the current round. Other measures could include, but not limited to, the percentage of licenses (measured in terms of activity units) on which there are new bids, the number of new bids, and the percentage increase in revenue.

strict, as stage evolves. In the reality, Stage 2 started at round 58, and Stage 3 started at round 70.

The auction provides several *waivers* of the activity rule for each bidder. The waivers are introduced to prevent errors in the bid submission process from causing unintended reduction is a bidder's eligibility [2]. In Auction 5, bidders are offered 5 activity waivers. In our model, we do not consider strategic uses of waivers.

To prevent collusion, the auctioneer (FCC) prohibits communications during the course of the auction among applicants for the same geographic license areas when such communications concern bids, bidding strategies or settlements.

2.3 Theoretical Analysis

In this section I study the Bayesian Nash equilibrium (BNE) and its properties in a multiple-object clock auction, where objects are heterogeneous and complementary. It has been widely discussed in the literature that solving the full equilibrium of the Simultaneous Ascending Auction (SAA) would be infeasible. However, conducting a theoretical analysis on a clock auction, which is an approximation of the SAA, would be necessary for us to deeply understand the complex auction mechanism and bidding strategies in the realistic SAA. The insights from the theory model can guide the construction of the structural model and the interpretation of the estimation results. Aside, the BNE for multiple-object clock auction with heterogeneity and complementarity and the equilibrium properties also add to the economic theory literature, and could be of independent interest.

In this section, I first review the literature of theoretical analysis of SAA and related auction models, which has a long tradition since the seminal paper by [2]. Next, I present the setup of my multiple-object clock auction model. Finally, I provide a proposition characterizing important properties of the Bayesian Nash equilibrium, along with discussions of the heuristics for the subsequent structural analysis. The full BNE will be given in Appendix A.1.

2.3.1 Related Literature

There is a rich literature concerning SAA and its related auction mechanisms, due to the momentous impact of its application into the spectrum auctions. [2] provides one of the earliest comprehensive theoretical analysis on SAA, and served as the fundamental for the following researches. He proposes the Straightforward Bidding (SB) strategy in SAA, meaning that bidders make decisions on which bundle of licenses to place bids on, based on their current-round payoffs. In other words, at each round, bidders behave as if it is the last round of the auction, and truly reveal their preference in the bidding. Milgrom showed that if licenses are substitutes, then SB is individually rational and consists of an efficient competitive equilibrium. However, the difficulty arises in the presence of complementarity, under which the equilibrium does not always exists. Afterwards, on the one hand, there are more theoretical results achieved under substitution. For example, [28] studies the multiple-object clock auction where preferences satisfy the "substitution condition", and proves that SB consists of an ex post perfect equilibrium, whose outcome is a modified VCG mechanism; [29] and [30] also focus on multiple-object clock auction under no or large complementarities, and showed the existence of the collusive equilibrium. [31] point out an intrinsic link between multiple-object auctions and matching models, when preferences satisfy substitution and law of aggregate demand. On the other hand, the study of the case with complementary objects remains challenging.

Of the literature my theory model is closest to our predecessors [9] and [32], both providing Bayesian Nash equilibrium analysis on the multiple-object clock auction with complementarity, using different settings. My model differs from them in the following ways. [9] consider n objects, each of which has one local bidder, and there are multiple global bidders. In their setting, the global bidders treat all objects *homogeneously*, in the sense that they assign the same values to two bundles with the same number of objects: $V_i(L) = \alpha(|L|) \cdot V_i$, where $V_i \sim F$. This assumption does not take the heterogeneity among objects into account in bidders' preferences, which is unlikely to be true in reality.[15] In addition, such assumption implies that the a global bidder has a single drop-

[15] Our structural estimation indeed shows that bidders' private values are indeed heterogeneous in objects.

out level b: when $p \leq b$, she stays in all objects; when $p > b$, she drops out at all objects and quits the auction. However, in the data we observe that bidders have different drop-out levels on different objects. In comparison, I consider *heterogeneous* objects in my model. [32] consider two objects, one global bidder, and multiple local bidders. Similar to our model, they assume that bidders will order the objects in the same way (in terms of stand-alone values). We consider a more general setting with multiple global bidders, which gives a more realistic insight to the practical spectrum auction. Moreover, I consider the case with multiple heterogeneous objects, and characterize the BNE using a recursive representation. [33] presents an impossibility result, stating that when objects are not substitutes, there is no standard ascending auction that yields a bidder-optimal competitive equilibrium under truthful bidding.

For other related theoretical researches about multiple-object auctions, [34] derive the BNE for simultaneous second-price sealed-bid auction with complementary goods. [35] characterizes the BNE in the simultaneous first-price sealed-bid auction. [36] provide a unified framework for simultaneous standard auctions where objects are complementary and sold separately, and proved the existence of a equilibrium which is monotonic in the sense of a suitable partial order. From the mechanism design perspective, [37] propose a new auction design to implement SB in the ex post perfect equilibrium for multiple-object auctions with complementarity. The idea is that the auctioneer provides prices for each bundle of objects, instead of for each individual object.

2.3.2 Setup

We consider a Multiple-object Clock Auction (MCA), which is regarded as a stylized approximation of the Simultaneous Ascending Auction (SAA). There are m heterogeneous objects indexed by $j \in J = \{1, \cdots, m\}$, to be allocated among n bidders, indexed by $i \in N = \{1, \cdots, n\}$.

First, I present the auction rule of MCA. The time is continuous: $t \in [0, \infty)$. At every time t, there is a price vector $p^t = (p^t(1), \ldots, p^t(m))$. Set $p^0 = (0, \ldots, 0)$. At every time t, each bidder makes a decision on each object: whether to stay or drop out. For an object j, if more than one bidder stays, then its price $p^t(j)$ will increase, while if only one stays or all drop out, the price will remain the same. The prices of all objects will increase (if any) continuously at the same rate. The

auction ends when all prices stop at some time T.

We impose an *activity rule* in the auction: bidders are not allowed to place bid on an object which she has already dropped out before. This is a simplified version of the activity rule used in the practical spectrum auctions. It will be clear that, in the models we are looking at in this section, the real-world activity rules (such as the ones used in US spectrum auction by FCC) will imply our activity rule.

Next, we specify the model in the language of game theory. Generally, we use upper case letters to denote random variables, and lower case their realizations. For distribution functions, we use upper case to denote CDF, and lower case the corresponding pdf (assume it exists).

We consider independent private value (IPV) paradigm. Before the auction begins, each bidder independently draws her value structure $\mathbf{V}$ from a commonly known distribution F with support Λ, where $\mathbf{V}$ is a 2^m- dimensional random vector, with every entry the value on a subset of J. Let $V : 2^J \to \mathbb{R}$ be the value mapping. Without loss of generality, set $V(\phi) = 0$ with probability one, where ϕ is the empty set. Let $v_i(L)$ be bidder i's private value on a set of object $L \subseteq J$.

The information structure is as follows. At every time t, (before bidding) a bidder observes: the current price vector p^t, the number of remaining bidders on each objects $N^t(j)$, and her own bidding history w_i^t. Therefore, the full history is $h^t = (p^\tau, (N^\tau(j))_{j=1}^m, w_i^\tau)_{\tau\in[0,t]} \in H^t$.

Let the final price and allocation of the objects be $((p(j))_{j=1}^m, (J_i)_{i=1}^n)$, where $(J_i)_{i=1}^n$ is a partition of J, with J_i the set of objects that bidder i wins finally. Then bidder i's utility is $\pi_i = v_i(J_i) - \sum_{j\in J_i} p(j)$, which is in a quasi-linear form. Bidder i's action space at each time is $\{1, 0\}^m$, where 1 denotes action stay, and 0 denotes out. Bidder i's strategy at time t is a function mapping her private and the current history to the action space, $s_i^t : \Lambda \times H_t \to \{1, 0\}^m$.

We use Bayesian Nash equilibrium as the solution concept of this dynamic game with incomplete information. We do not consider the sub-games on the off-equilibrium path.

2.3.3 Equilibrium Properties

In this subsection, I analyze the two-object clock auction, and provide the relevant equilibrium properties to shed lights on the general bidding behaviors in SAA. The characterizations of BNE

for the two-object and multiple-object case are provided in details in Appendix A.1, with related proofs in Appendix A.2.

We denote the two objects by A and B. Let bidder i's stand-alone value of A and B be $v_i^A \equiv v_i(\{A\})$ and $v_i^B \equiv v_i(\{A\})$, with distribution F^A and F^B, respectively. Let the support of F^A and F^B be $[\underline{v}^A, \overline{v}^A]$ and $[\underline{v}^B, \overline{v}^B]$, respectively. Denote the complementarity between the two objects to be $\theta_i \equiv v_i(\{A, B\}) - v_i^A - v_i^B$. Let $\theta_i \sim F^\theta$ with support $[\underline{\theta}, \overline{\theta}]$. Therefore, bidder i's value structure is $\mathbf{v}_i = (v_i^A, v_i^B, \theta_i)$, with the commonly known distribution $F = (F^A, F^B, F^\theta)$. Bidders decide when (facing what price vector) to drop out on objects, conditional on their private value structure and current history. Let the drop-out price of bidder i on object A and B be s_i^A and s_i^B, respectively.

Suppose there are $n > 1$ bidders in the auction, who demand both objects.[16] Next, we make an assumption on the value structure distribution of the bidders.

Assumption 1 (Weak Value Ordering)**.** *The value structure distribution F satisfies:*

$$\mathbb{P}(V_i^A > V_i^B) = 1. \tag{2.1}$$

Assumption 1 says that for any bidder, object A is more valuable than object B. Although bidders' values are private information, this order between objects is common knowledge. This is reasonable because in practice, bidders' values are affected by some publicly observed characteristics of the objects. For example, in the US spectrum auction, the objects to be auctioned are the spectrum licenses in geometric areas. It is natural to assume that bidders all agree that the license in Houston area is more valuable than the one in College Station-Bryan area, due to the population difference.

With Assumption 1 and the activity rule, it is clear to see that a bidder will drop out on object B first, and then (or simultaneously) drop out on object A. That is, $s_i^A \geq s_i^B$. Moreover, the drop-out price on B will not fall below the stand-alone value on B (see Lemma 1 in Appendix A.1). This

[16] In [34] and [9], these are called *global bidders*. Without loss of generality, we do not independently consider local bidders.

property greatly simplifies the equilibrium strategy analysis. The full characterization of BNE is given in Appendix A.1. From Lemma 2 of Appendix A.1, we see that after winning or dropping out at B, the bidding strategy on A is just the same as single-object English auction. Therefore, the equilibrium drop-out level on B, s_i^B, is going to capture the unique feature of the multiple-object auction with complementarity. The next proposition states the comparative static properties of s_i^B. The proof is given in Appendix A.3.

Proposition 1. *Let Assumption 1 hold. For a two-object clock auction, the Bayesian Nash equilibrium strategy s_i^B, which is the drop-out level on object B, satisfies the following:*

1. *s_i^B is non-decreasing in the stand-alone values v_i^A and v_i^B.*
2. *s_i^B is non-decreasing in the complementarity value θ_i.*
3. *s_i^B is non-increasing in the number of bidders remaining on object A, denoted as n^A.*

Since there is no analytical solution for the equilibrium, we find numerical solutions using MATLAB. We plot s_i^B with respect to varying v_i^A, v_i^B, θ_i, and n^A. I use (truncated) uniform distribution to be the common distribution of v_i^A, and explore the case with (truncated) normal distribution for robustness check in Appendix A.4. The range of F^A is $\underline{v}^A = 0.3, \overline{v}^A = 1$.

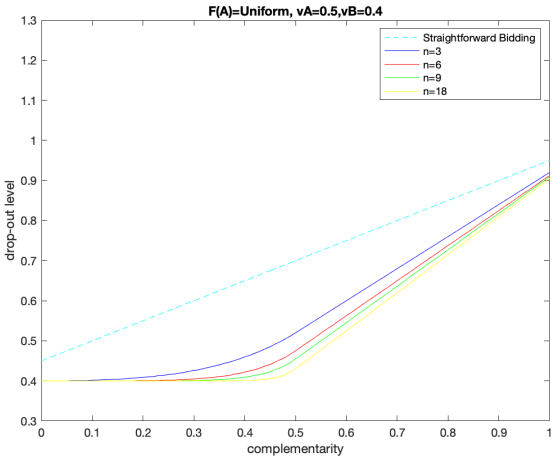
First, the graphs demonstrate Proposition 1. We see from Figure 2.1 that s_i^B is non-decreasing in the complementarity value θ_i and non-increasing in the number of bidders n^A, given different values of fixed v_i^A and v_i^B. Figure 2.2 use the same parameter settings with Figure 2.1, and show that s_i^B is non-decreasing in v_i^A and v_i^B.

Moreover, from Figure 2.1, we see that the equilibrium bidding strategy (drop-out level) s_i^B does not exceed their true value (which corresponds to SB), i.e. $s_i^B = (v_i^A + v_i^B + \theta_i)/2$.[17] This is obvious because bidders will have negative profit for bidding something higher than her true value. When bidders' value $(v_i^A + v_i^B + \theta_i)$ is not large enough (see Figure 2.1a and 2.1c), the BNE bidding strategy is lower than the true value, which corresponds to the *exposure problem* for the

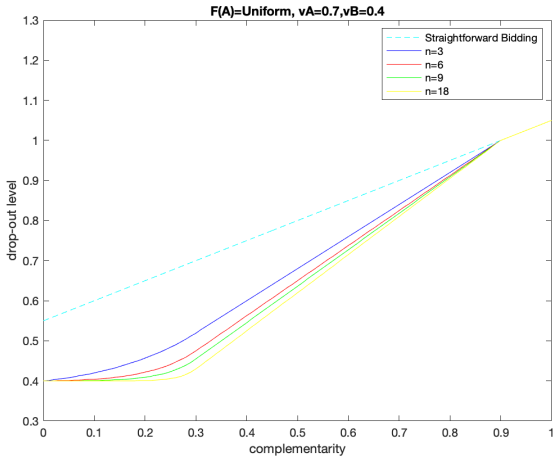
[17]Because by Lemma 1, $s_i^A \geq s_i^B$, then when bidder i drops out at object B, the price for her at object A is also equal to s_i^B. Therefore, at that moment, the total price is $2s_i^B$.

Figure 2.1: Equilibrium Bidding Strategy s_i^B and Complementarity θ_i.
Property: s_i^B Non-increasing in n^A and Non-decreasing in θ_i

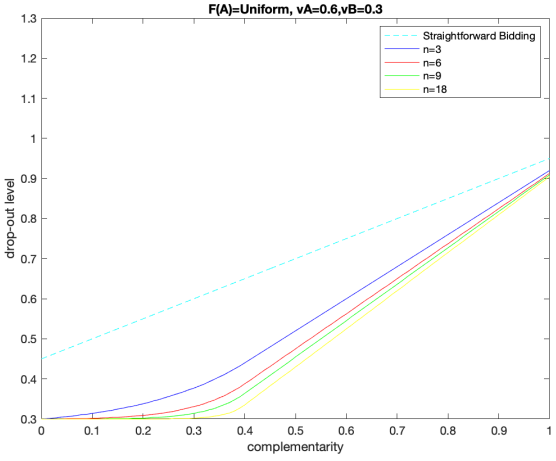
(a) $v_i^A = 0.5, v_i^B = 0.4$



(b) $v_i^A = 0.7, v_i^B = 0.4$



(c) $v_i^B = 0.3, v_i^A = 0.6$



(d) $v_i^B = 0.5, v_i^A = 0.6$

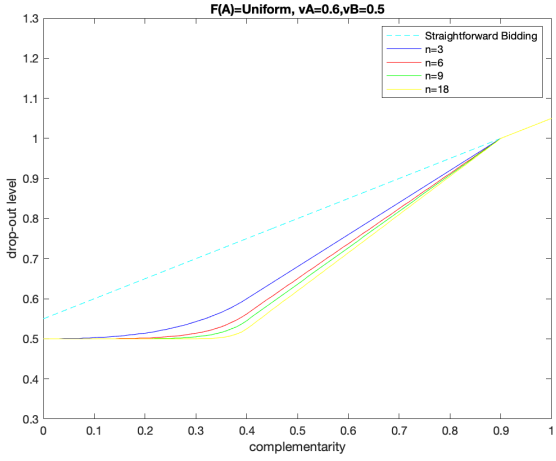
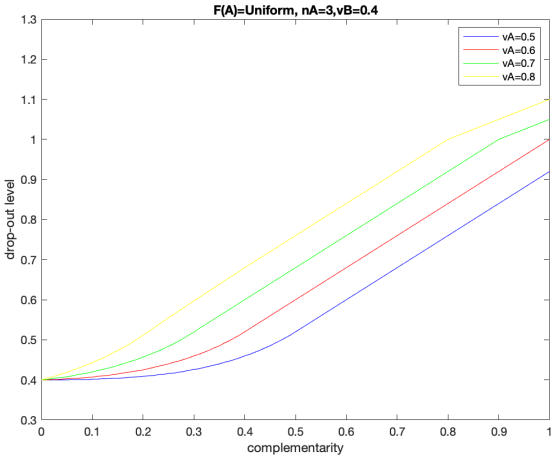
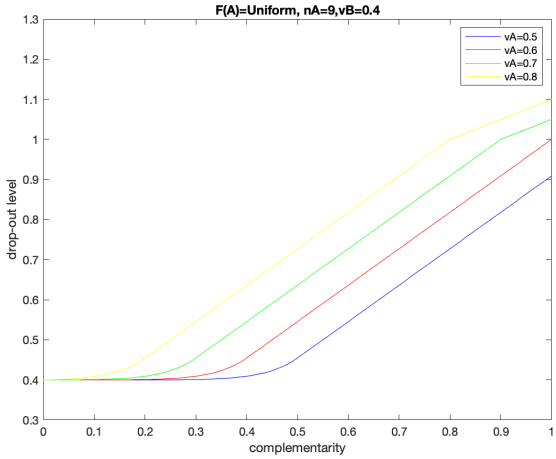


Figure 2.2: Equilibrium Bidding Strategy s_i^B and Complementarity θ_i.
Property: s_i^B Non-decreasing in v_i^A, v_i^B, and θ_i.

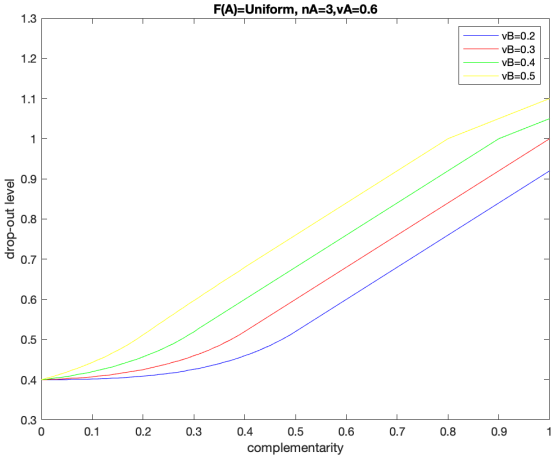
(a) $n^A = 3, v_i^B = 0.4$



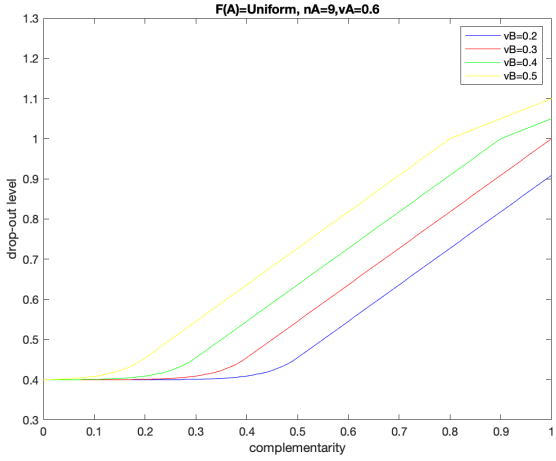
(b) $n^A = 9, v_i^B = 0.4$



(c) $n^A = 3, v_i^A = 0.6$



(d) $n^A = 9, v_i^A = 0.6$



multiple-object auction. Bidders will discount their true values in evaluating their optimal drop-out level, or put in another way, use a *discounted* version of private value in the SB. The reason is, as mentioned previously, that bidders will consider the risk of winning only a part of the bundle of objects (in this case, only winning object B), and obtaining the complementarity θ_i only if she wins both A and B. However, when a bidder value is large enough (see Figure 2.1b and 2.1d), the BNE strategy converges to their true value. In other words, large bidders will reveal their true preference in equilibrium, i.e. use straightforward bidding, and will not exhibit the exposure problem. This is because for a large bidder, she has either strong enough belief that the bundle will belong to her as a whole, or high enough complementarity such that the expected payoff of bidding exceeds not-bidding.

2.4 Data and Preliminary Analysis

This paper analyzes the 1995-1996 US spectrum auction held by FCC, which is also referred to Auction 5. I focus on the C block auction of 1900 MHz PCS band.[18] There are 493 licenses for sale, and 255 bidders in this auction. It lasts for 183 rounds, starting at 1995/12/18, ending at 1996/5/3. The auction has a gross total bids of 13 billion dollars, and a net total bids of 10 billion dollars.[19]

In this section, I provide some preliminary analysis to show the data pattern, which motivates our subsequent structural modeling. Firstly, I present evidence of Straightforward Bidding (SB) for the bidders, meaning that: (1) they bid on the minimum acceptable bids; and (2) they behave myopically and non-strategically. Secondly, I observe substantial variation of bidders' decision of bidding bundles across rounds. Finally, I look at the round-specific statistics to reveal the trend of the auction across rounds.

2.4.1 Straightforward Bidding

First, I observe that most of the submitted bids are following the minimum acceptable bid (MAB). Of all the submitted bids, 58.75% are exactly equal to the MAB, and 96.57% exceed

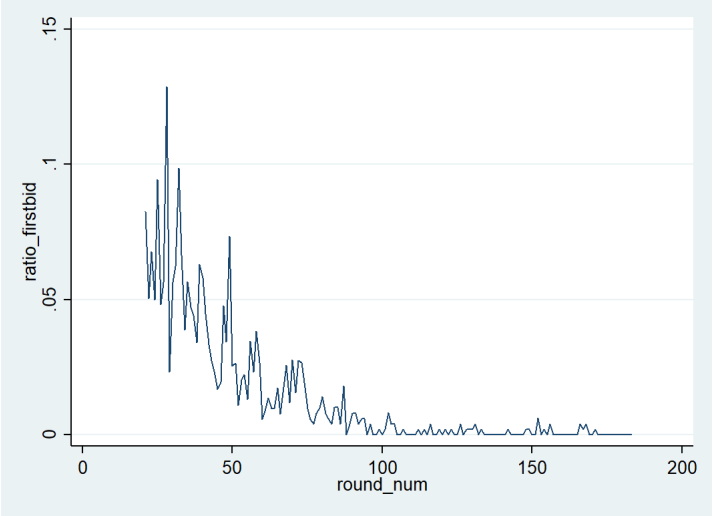
[18]The dataset is from [18].

[19]Some winners failed to fulfill the payments.

the minimum acceptable price by less than 1% of the MAB. Therefore, jump bidding is hardly a concern here. Bidders' bidding strategy could be simplified as choosing the bundle of licenses to place bids on (after the determination of the bundle, she just submits the MAB).

My second observation is that there is only a small proportion of first time bidders at each round. This suggests that bidders are not likely to bid strategically based on other bidders' action. They may just make use of their own valuation and the current round information on objects to make bidding decision. This motivates our non-strategic bidding strategy. See Figure 2.3.

Figure 2.3: Percentage of First Time Bidder in Each Round.

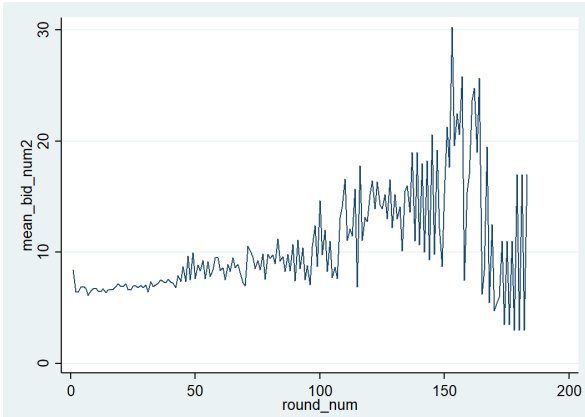


2.4.2 Variation in Bidders' Choices

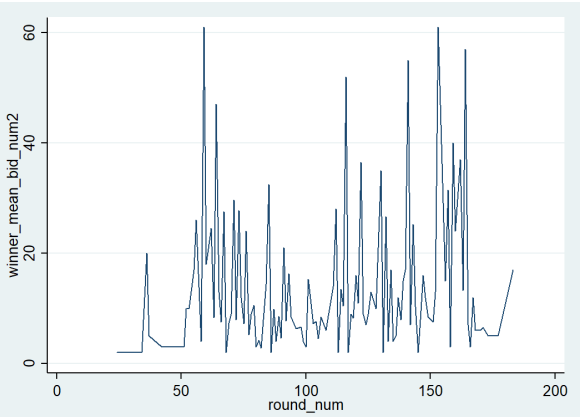
We observe substantive variation of bidders' choices across rounds. See Figure 2.4, where I plot the size of bidding set with respect to the round number, for the average bidder, average winner, and two specific bidders. Such variation may provide rich information for us to recover the bidders' preferences. This motivates us to model bidders' choices of bidding set across rounds.

Figure 2.4: Variation of Choice Sets over Rounds

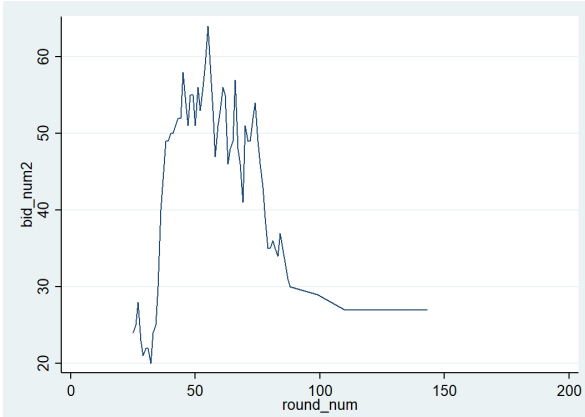
(a) All Bidders



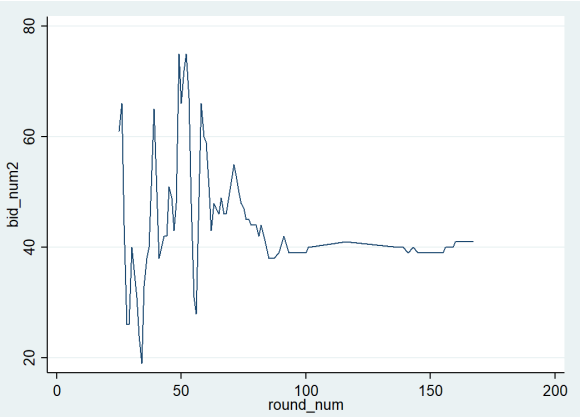
(b) Winners



(c) Bidder #2332



(d) Bidder #2358



2.4.3 Round-Specific Statistics

I summarize the bidding statistics for every 20 rounds. See Table 2.2. I construct several quantities, where choice_var is the number that bidders change their bidding set at the next round, summing over the 20 rounds; mean_bid is the mean of total bids for a round, measured in thousand dollars, summing over the 20 rounds; newbids is the number of new submitted bids compared to last round, summing over the 20 rounds; num_bidder is the number of active bidders during the 20 rounds; bid_bidder_ratio is equal to newbids divided by num_bidder, which measures that during the 20 rounds, how many new bids are submitted for an active bidder; license_win is the number of licenses that are won (has stopped the price ascending thereafter) during the 20 rounds.

Firstly, we see that the first 20 rounds are chaotic and tentative. The variation of choice bundle is large, and the number of new bids per bidder is very high. In addition, no license is won at this stage. That is because bidders might be learning their private values and the optimal bidding strategy at the beginning, knowing that the auction will not end this early. Secondly, for the rounds after 120, we see that the bidding behaviors become stable. When the auction approaches to the end, there are only small number of licenses subject to bidding, so it becomes feasible for bidders to consider their competitors in specific licenses.[20] Therefore, the bidding behavior tends to be more strategic and forward-looking, and may violate the straightforward bidding style in our structural model. In view of the above considerations, I select rounds $21-120$ as our sample in the estimation in Section 2.6.

2.5 Structural Model

In this section, I present the structural model. We model the bidders' bidding strategy as choosing the bundle of licenses that maximizes their *current-round expected payoffs*. Along the auction, a bidder has private values for each bundle of licenses, which is fixed across rounds. At each round, she forms the current-round expected payoffs based on her beliefs on the current-round winning probabilities on each license on her bidding bundle. The winning probability is the

[20] For the Auction 5 that we analyze, bidders could observe the identities of all other bidders for all licenses at each round.

Table 2.2: Bidding Statistics by Rounds

round	choice_var	mean_bid	newbids	num_bidder	bid_bidder_ratio	license_win
1-20	1825	9761.58	13389	253	52.92	0
21-40	1217	16237.20	7726	199	38.82	16
41-60	915	18586.41	4510	156	28.91	49
61-80	647	21930.57	2571	124	20.73	183
81-100	349	24802.28	895	106	8.44	108
101-120	199	26215.82	321	97	3.31	53
121-140	157	26623.58	239	93	2.57	32
141-160	159	26961.47	141	89	1.58	31
161-183	120	27120.83	73	89	0.82	21

probability that she becomes the provisional winner on a license at the end of the current round, which depends on the current-round information. Therefore, bidders make decisions at each round based on their private values and current-round information, as if this would be the final round. The equilibrium is such that bidders' decision maximize their current-round expected payoffs, and their beliefs are consistent with the realized outcome. Finally, I describe the specification of bidders' private value function, which includes the stand-alone value and complementarity.

[11] use a similar model, while they assume that bidders form beliefs about the final winning probabilities, which is hard for bidders to consistently predict, especially in the middle of the auction.

2.5.1 Setup

There are m heterogeneous licenses indexed by $j \in J = \{1, \cdots, m\}$, to be allocated among n bidders, indexed by $i \in N = \{1, \cdots, n\}$. I formalize the setup of Simultaneous Ascending Auction (SAA) as follows.

The SAA is a multiple-round process. Denote the rounds by $t = 1, ..., T$. During each round t, each bidder i could submit sealed bids for (possibly) multiple licenses. Denote the bidding amount of bidder i on license j by $b_i^t(j)$, and the bidding set L_i^t. After the end of round t, the highest bid for object j is referred to "standing high bid", denoted as $r^t(j) = \max_{i \in N} b_i^t(j)$. The corresponding

bidder is the "standing high bidder".[21] The minimum acceptable bids at round $t+1$, denoted as $p^{t+1}(j)$ are computed as the standing high bids $r^t(j)$ plus some smallest increment $d^t(j)$, i.e. $p^{t+1}(j) = r^t(j) + d^t(j)$.[22] The auction proceeds in an ascending way in the sense that at each round any submitted bids should be higher than the minimum acceptable bids. The auction stops at round T when no new bids are submitted for any objects in this round. The objects are allocated to their high bidders at round $T-1$, at the standing high bids.

We consider the independent private value (IPV) paradigm. Before the auction begins, bidder i forms her private value $v_i(L)$ on each set of license $L \in \mathcal{L}$, where $\mathcal{L}$ is all potential combinations of licenses that bidders will consider. In total there are 2^m subsets of $J = \{1, ..., m\}$, but not all combinations are beneficial for the bidders. In particular, $v_i(L)$ is composed of the *stand-alone values* and *complementarity*:

$$v_i(L) = \sum_{j \in L} v_i(j) + \frac{1}{2} \sum_{j \in L} \sum_{j' \neq j, j' \in L} \eta(j, j'), \tag{2.2}$$

where $v_i(j)$ is the value of obtaining license j alone, and $\eta(j, j')$ is the complementarity between the two licenses j and j', i.e. $\eta(j, j') = v_i(\{j, j'\}) - v_i(j) - v_i(j')$. The private value is invariant across rounds. Note that We assume the complementarity among the set L is composed of all pairwise complementarities of pairs $\{j, j'\} \in L$. That is, we assume away the higher order complementarities among licenses (extra complementarities generated from triples, quadruples, etc). The payoff of obtaining a set of license L is $u_i(L) = v_i(L) - \sum_{j \in L} p(j)$, where $p(j)$ is the final price on license j.

At the end of a round t, round results will be posted. These include the minimum acceptable price for each license at next round $t+1$, $\boldsymbol{p}^{t+1} = (p^{t+1}(j))_{j \in J}$, and all new submitted bids along with bidder identities on each license. Denote the round results up to t that bidder i could observe

[21] If nobody has ever bid on an object, or the standing high bidder withdraws, then the standing high bidder is defined to be the auctioneer. Ties are broken randomly.

[22] The initial minimum acceptable bid for each license is zero. Once a bid has been received on a license, the minimum bid increment $d^t(j)$ for that license will be the greater of 5% of the previous high bid or $\$0.02$ per bidding unit.

as h_i^t.

Before the auction, bidders are informed of the characteristics of all licenses, and their private values. During each round t, bidder i's information set $\mathcal{I}_i^t$ contains all historical round results h_i^t, and the before-auction information, which is her private value $\boldsymbol{v}_i$.

2.5.2 Strategy and Decision Making

From the previous discussion, I assume that bidders are bidding in a similar manner of Straightforward Bidding as in [2]. At each round, bidders make decisions on which licenses to bid, based on their current information set. Therefore, the action space for the bidders is the collection of all combinations of licenses to be considered during the auction, denoted by $\mathcal{L}$. Let $\boldsymbol{Y}_i^t = (Y_i^t(j))_{j=1}^m$, where $Y_i^t(j)$ is a binary variable indicating whether bidder i bids on license j at round t, and with slight abuse of notation, I sometimes use $\boldsymbol{Y}_i^t = L$ to denote the bidding set. Moreover, bidders do not take into account neither the strategic interactions with other bidders, nor do they focus on the current round payoff. Hence, bidder i's strategy at round t is $s_i^t : \mathcal{H} \to \mathcal{L}$. That is, a bidder chooses a set of licenses $L \in \mathcal{L}$ to place bids on, based on her information set $\mathcal{I}_i^t \in \mathcal{H}$.

In [2], bidders choose the bidding set to maximize their current round payoff. To address the exposure problem in the Simultaneous Ascending Auction, we modify the original version of Straightforward Bidding, such that the bidders are maximizing their *expected* current round payoff. To calculate their expected current round payoff, bidders need to form beliefs about their winning probabilities on their bidding licenses at this round. We discuss the beliefs in the next subsection.

2.5.3 Belief and Current-Round Expected Payoff

At a round t, bidder i's *current-round* payoff for set L is the difference between her private value on L and the current-round personalized price on L:

$$u_i^t(L) = v_i(L) - \sum_{j \in L} p_i^t(j) \tag{2.3}$$

Bidders form beliefs of the current-round winning probabilities on each license she bids or provisionally wins, given her information set at the current round. Specifically, if bidder i places

a bid on license j at round t or she is the provisional winner on license j at round $t-1$, then $q_i^t(j) = q(j|\mathcal{I}_i^t)$ is bidder i's belief that she will (provisionally) win license j at round t, given her information set at the beginning of round t, $\mathcal{I}_i^t$.

Assumption 2 (Independence)**.** $q_i^t(\{j, j'\}) = q_i^t(j)q_i^t(j')$.

This assumes that all complementarity effects in the current round winning probabilities have been revealed in the information set.

Given the current-round payoff function and the belief, bidder i's current-round *expected* payoff at round t on a set of licenses L is

$$\mathbb{E}[u_i^t(L)|\boldsymbol{q}_i^t, \mathcal{I}_i^t] = \sum_{j\in L} v_i(j)q_i^t(j) + \frac{1}{2}\sum_{j'\in L}\sum_{j'\neq j, j'\in L} \eta(j, j')q_i^t(j)q_i^t(j') - \sum_{j\in L} p_i^t(j)q_i^t(j), \quad (2.4)$$

where the expectation is with respect to the winning probabilities $\boldsymbol{q}_i^t$, and the price for a license is the current-round personalized price of bidder i, $p_i^t(j)$.

Therefore, at round t, bidder i places bids on L_i^t if and only if

$$L_i^t = \arg\max_{L\in\mathcal{L}} \mathbb{E}[u_i^t(L)|\boldsymbol{q}_i^t, \mathcal{I}_i^t]. \quad (2.5)$$

2.5.4 Equilibrium

The equilibrium concept of the structural model is the Bayesian Nash equilibrium, defined as follows. In the equilibrium, bidders are choosing the bundle of licenses to place bids on by maximizing their current-round expected payoffs, and their beliefs about the current-round winning probabilities are consistent with the realized outcome.

Definition 1 (Bayesian Nash Equilibrium)**.** *Let a bidding strategy of bidder i at round t be $\sigma_i^t(\mathcal{I}_i^t, \boldsymbol{q}_i^t)$, and a belief system be $\boldsymbol{q}_i^t$. Then $(\sigma_i^t, \boldsymbol{q}_i^t)_{i=1,t=1}^{n,T}$ is a Bayesian Nash Equilibrium such that $\sigma_i^t(\mathcal{I}_i^t, \boldsymbol{q}_i^t) = L_i^t$, if and only if for any $L' \in \mathcal{L}$,*

$$\mathbb{E}[u_i^t(L_i^t)|\boldsymbol{q}_i^t, \mathcal{I}_i^t] \geq \mathbb{E}[u_i^t(L')|\boldsymbol{q}_i^t, \mathcal{I}_i^t], \quad (2.6)$$

and the true winning probabilities are consistent with bidders' belief system.

2.5.5 Parametrization

We parametrize the deterministic part of bidders' stand-alone value to be

$$\overline{v}_i(j) = \alpha_0 + \alpha_1 \operatorname{elig}_i + \alpha_2 \operatorname{pop}(j) + \alpha_3 \operatorname{elig}_i \operatorname{pop}(j). \tag{2.7}$$

Here $\operatorname{pop}(j)$ is the population in the area of license j from the 1990 census data, normalized in the fraction of the total US population. elig_i is bidder i's *initial eligibility*. This initial eligibility is measured in population, meaning that at any round, bidder i cannot bid on a set of licenses with total population larger than elig_i. This number is submitted by to bidders to the auctioneer before the auction begins, and is associated with an upfront payment. Therefore it could represent a bidder's maximum willingness to pay during the auction, and could serve as the bidders' characteristic.

The deterministic part of pairwise complementarity is modeled as

$$\overline{\eta}(j, j') = \beta \tau(j, j'), \tag{2.8}$$

where $\tau(j, j')$ is the complementarity measure between license j and j', which is associated with their populations, and the geographic distance between them:

$$\tau(j, j') = \operatorname{pop}(j) \frac{\frac{\operatorname{pop}(j)\operatorname{pop}(j')}{\operatorname{dist}^{\delta}(j,j')}}{\sum_{k \in J, k \neq j} \frac{\operatorname{pop}(j)\operatorname{pop}(k)}{\operatorname{dist}^{\delta}(j,k)}} + \operatorname{pop}(j') \frac{\frac{\operatorname{pop}(j)\operatorname{pop}(j')}{\operatorname{dist}^{\delta}(j,j')}}{\sum_{k \in J, k \neq j'} \frac{\operatorname{pop}(j')\operatorname{pop}(k)}{\operatorname{dist}^{\delta}(j',k)}}. \tag{2.9}$$

This complementarity measure follows from [11] and [8]. We also set $\delta = 2$ following the specification of the above two papers. Notice that the nationwide complementarity constructed from this pairwise complementarity measure is $\sum_{j,j' \in J, j \neq j'} \tau(j, j') = 1$, making us easier to interpret the estimated complementarity effect.

Hence, bidder i's deterministic private value on a set of license L is

$$\overline{v}_i(L;\theta) = \alpha_0 + \alpha_1 \operatorname{elig}_i + \alpha_2 \operatorname{pop}(j) + \alpha_3 \operatorname{elig}_i \operatorname{pop}(j) + \frac{1}{2}\beta \sum_{j \neq j';j,j' \in L} \tau(j,j'), \tag{2.10}$$

where $\theta = (\alpha', \beta)'$ is the structural parameter we are interested in.

2.6 Estimation and Results

In this section, I elaborate on the four-step estimation approach of the structural model for the parameters in bidders' private values. I first describe the sample to be used. Next, I provide an overview of the four steps that consisting the estimation. Finally, I illustrate the four estimation steps in detail, along with the results.

2.6.1 Data Preparation

We focus on the 480 licenses in the continental United States, and assume that the complementarity between the oversea licenses and any continental licenses is always 0. Therefore, it can be regarded that the oversea licenses are held in another separate auction.

There exist only a few withdrawls from the high bidder of a current round. We drop the observation if there is a withdrawl. We treat the withdrawing as a mistake of bidding and is unintentional. That is to say, assuming away the penalty, the remaining effective bidding set maximizes her expected utility at the current round, although she realized that afterwards. During the auction, 184 withdrawls are observed, amounting to 0.6% of the total 29865 submitted bids.

In view of Section 2.4.3, I select round 21-120 for the sample in the structural estimation. The middle stage of the auction avoids the noisy information (e.g. bidders are learning their private values and optimal bidding strategy) at the beginning stage as well as the strategic behaviors at the ending stage, so it is more suitable for our proposed structural model.

2.6.2 Overview of the Estimation Procedure

In this subsection, I provide an overview of our estimation method for recovering bidders' private values. Firstly, since bidders' decisions are based on their current-round expected payoffs,

we need to estimate the winning probabilities (Step 1). Subsequently, we are faced with a bundle choice problem based on the current-round expected payoffs, where the total number of bundles is extremely large. There are 480 licenses in the estimation, resulting in 2^{480} bundles in total. I apply the semiparametric estimation method for multinomial discrete choice proposed by [12], into our bundle choice problem (Step 4). This method is based on the *cyclic monotonicity* of the choice probability function. There are two essential properties of this approach. It allows for arbitrary dependence across different choices, which particularly fits our case because bundles containing the same licenses are correlated. To deal with the high-dimensionality issue, I leverage the random projection technique in machine learning to achieve dimension reduction on the number of choices, which is proposed by [13] (Step 3). However, [12]'s method demands the choice probabilities for each bidder at each round, while we only observe the realized chosen bundle. In view of that, I need to estimate the choice probabilities beforehand (Step 2). Therefore, in the proposed estimation procedure, I first estimate the winning probabilities and calculate the current-round expected payoffs. Secondly, I estimate the choice probabilities for each bundle. Third, I apply the random projection to the quantities obtained from Step 1 and 2. Finally, I carry the projected down data to the multinomial discrete choice estimation to recover the structural parameters.

2.6.3 Estimating the Winning Probabilities

In this subsection, we discuss the estimation of bidders' current-round winning probabilities, i.e. the probability of provisionally winning a license she bids on at the current round. In particular, we want to estimate $q_i^t(j), \forall j \in J$, that is, for all the 480 licenses. However, we only observe the outcome of on the *bidding set* L_i^t of bidder i at round t: whether she provisionally wins license $j \in L_i^t$ or not. Table 2.3 summarizes the number of licenses in bidders' bidding sets.

Table 2.3: Size of Bidding Set

	Min.	1st Qu.	Median	Mean	3rd Qu.	95 Qu.	Max.
Round 1-183	1	1	3	5.29	6	16	483
Round 21-120	1	1	3	5.21	6	17	103

We see that most bidders in most rounds have a small size of bidding sets. 95% of the bidding sets contain less than 17 licenses, which is about 3.5% of all the licenses. In other words, we are going to fit our prediction model in a relatively small training set and make predictions in a relatively large test set. This calls for a powerful prediction model with good predictive accuracy. On the other hand, we only need the estimated winning probabilities in the next steps for calculating the expected current round payoff of the bidders. Thus, in this step what we only care about is the predictive performance, rather than model parameters. In view of that, we look for a method which could provide good prediction accuracy for the probability of a binary classification problem.

We consider the traditional linear logit model, as well as five off-the-shelf machine learning classification methods: LASSO, Random Forest, Boosting, Support Vector Machine, and Neural Nets. In general, the machine learning algorithms haven been demonstrated to possess better predictive power than the traditional parametric models, because they are all trying to balance the bias and variance of the model, and thus avoid overfitting and improve predictive performance in the test set. Except for LASSO, which is a penalized linear regression model, all other four methods are nonparametric. See [38] for details of these machine learning methods.

Moreover, we apply model averaging to the above mentioned six models, in an attempt to further improve the predictive performance compared to any single method. Model averaging is a general idea of combining several single models to alleviate model uncertainty. Specifically, when all existing models are misspecified, model averaging could improve the prediction accuracy upon each single model [39]. We adopt the model averaging scheme in [23], whose final prediction is a weighted average of each individual model prediction.

The covariates we use to estimate the winning probabilities are:

1. t: the round number

2. elig_i: bidder i's initial eligibility.

3. $\mathrm{pop}(j)$: license j's population.

4. nbid_i^t: number of bids that bidder i places at round $t-1$.

5. $\text{rwin}_i^t(j)$: number of rounds that bidder i is the provisional winner on license j before round t.

6. $n_i^t(j)$: number of bidders on license j at last round $t-1$ except bidder i.

7. $p_i^t(j)$: bidder i's personalized price on license j at round t. Particularly, if bidder i is the provisional winner on license j at round $t-1$, then $p_i^t(j)$ is the bidding amount she submitted at round $t-1$; if bidder i is not the provisionally winner on license j at round $t-1$, then $p_i^t(j)$ is the minimum acceptable price, which equals to the standing high bid at last round plus a small increment.

8. $c_i^t(j) = \sum_{j' \neq j} \tau(j, j') \mathbb{1}(j' \in W_{i,t-1})$: the contribution of license j to the complementarity effect in the provisional winning set of bidder i.

These covariates are all in the bidders' current-round information set $\mathcal{I}_i^t$. Next, we present prediction results from the six models, as well as the model averaging scheme.

2.6.3.1 Logit

I first estimate the traditional logit model for the winning probabilities. The results are shown in Table 2.4. We see that $n_i^t(j)$ and $c_i^t(j)$ have the largest explanatory and predictive power to winning probabilities. On the one hand, holding all else constant, when there are more bidders on a license at the previous round, a bidder may anticipate higher competition on this license, so that her chance of winning this license is smaller. This is consistent with our theoretical analysis in Section 2.3, as well as other general results in the auction theory. On the other hand, a strong and positive coefficient on $c_i^t(j)$ demonstrates the existence of complementarity among licenses in bidders' values. Furthermore, population has a negative effect on winning probabilities, meaning that in bidders beliefs, a license with higher value is more attractive and more difficult to win. The positive sign on the number of rounds is intuitive, because as rounds increase, more competitors will drop out, and if a bidder stays, she will be more confident that she could outbid other bidders. The positive sign on $\text{rwin}_i^t(j)$ is also reasonable, since the more times a bidder being the winner

on a license, the more confident she feels for winning this license. However, the negative sign on nbid_i^t is counter-intuitive: if a bidder placed at last round, she will have less probability of winning a license at the current round.

Table 2.4: Logit

Winning Prob	**Estimate**	**Std. Error**	**t value**	**Pr**(>|z|)
(Intercept)	0.2369	0.0355	6.6680	0.0000
Number of Rounds: t	0.0122	0.0005	22.9106	0.0000
Eligibility: elig_i	0.0045	0.0009	5.2710	0.0000
Population: $\mathrm{pop}(j)$	-0.5304	0.1391	-3.8144	0.0001
Number of Bidders: $n_i^t(j)$	-0.3469	0.0084	-41.3120	0.0000
Complementarity Effect: $c_i^t(j)$	0.4974	0.1114	4.4639	0.0000
Price: $p_i^t(j)$	0.0056	0.0010	5.5346	0.0000
Number of Bids: nbid_i^t	-0.0048	0.0008	-6.1881	0.0000
Winning Rounds: $\mathrm{rwin}_i^t(j)$	0.0338	0.0009	38.5801	0.0000
Adj R Squared	0.213148			

2.6.3.2 LASSO

LASSO (Least Absolute Shrinkage And Selection Operator) is a penalized or regularized regression method that would achieve variable selection and improve prediction accuracy compared to the traditional linear regression. The regression coefficients are estimated by

$$\widehat{\beta}_\lambda^L = \arg\min_\beta (y_i - x_i'\beta)^2 + \lambda\|\beta\|_1, \tag{2.11}$$

where the first part of the objective function is the squared loss which corresponds to the optimization objective of OLS, and the second part is the penalty term that shrinks the size of estimator $\widehat{\beta}$. Due to the introduction of the penalty, the LASSO estimator will shrink towards zero, and thus will be biased. However, the shrinkage also serves to reduce the variance of $\widehat{\beta}$. The tunning parameter λ controls for the amount of penalization and therefore balance the bias-variance trade-off. An appropriate value of λ can reduce the variance substantially, at the cost of a little increase in bias,

and thus achieve lower test MSE. Hence, LASSO deals with the overfitting problem that may occur in linear regression, especially when there are many regressors. Moreover, LASSO is using the ℓ_1 penalty of the parameter. This results in that some estimated coefficients will be shrunk to exactly zero, leading to a variable selection in the regression.

Table 2.5 shows the estimation results of LASSO. We see that the magnitude and sign of logit and LASSO estimation are quite consistent. In particular, LASSO drops the predictor $\mathrm{pop}(j)$ and nbid_i^t, where in the logit model, $\mathrm{pop}(j)$ is the least significant variable and nbid_i^t has the smallest coefficient. In addition, both logit and LASSO give very small estimates on elig_i and $p_i^t(j)$, and the signs are consistent. More importantly, $n_i^t(j)$ and $c_i^t(j)$ obtain the largest coefficients, which is the same as the logit estimation. The sign and magnitude of the estimates on $\mathrm{rwin}_i^t(j)$ and t are comparable to the logit model.

I implement LASSO using the **glmnet** package in R. The tuning parameter λ is chosen by cross validation.

Table 2.5: LASSO

Winning Prob	**Estimate**
(Intercept)	0.2041082
Number of Rounds: t	0.0122
Eligibility: elig_i	2.0720e-05
Population: $\mathrm{pop}(j)$	.
Number of Bidders: $n_i^t(j)$	-0.3351
Complementarity Effect: $c_i^t(j)$	0.2319
Price: $p_i^t(j)$	1.2968e-03
Number of Bids: nbid_i^t	.
Winning Rounds: $\mathrm{rwin}_i^t(j)$	0.0316

2.6.3.3 Random Forest

Random Forest (RF) is an ensembling method (which shares similar spirits with model averaging) based on the decision tree model. It is a nonparametric estimation method for the conditional

expectation, which is typically useful when there are high-dimensional predictors. When there are many covariates, the traditional nonparametric estimation methods like k-nn, kernel, series or splines, will suffer from curse of dimensionality. However, random forest has been demonstrated to work very well when the number of features is large. We first introduce the regression tree and classification tree, and then turn to the algorithm of random forest.

Decision tree is based on partitioning the feature space into regions. Then we simply make predictions for a given test sample using the mean (for regression problem) or mode (for classification problem) of the region that the test sample falls into. This idea is similar to the k-nn or kernel estimation, while the tree-based models determine the distance among observations via *trees*. The tree-based models could also be viewed as the generalized fixed effect models, where the fixed effects are allowed to vary over the characterstic space.

Suppose we have covariates X_{ij} and response Y_i, we select a predictor X_j and a cutpoint s splitting the predictor space into two regions $R_1(s,j) = \{X|X_j < s\}$ and $R_2(s,j) = \{X|X_j \geq s\}$, which leads to the greatest possible reduction in RSS (regression tree) or Gini index (classification tree). We repeat the same procedure on each obtained region, and find new splits that results in the greatest reduction is RSS or Gini index.[23] We stop until some stopping criterion is reached, e.g. no region contains more than a certain number observations, or until certain times of splits are made. Such algorithm is called Recursive Binary Splitting. Finally, the predicted response for a test sample x is the mean (regression) or mode (classification) of the response in the region where x falls into.

RF is originated from bagging, which has the similar idea of model averaging. In bagging, we grow B trees where each tree uses a bootstrap sample from the training data. For regression, the final prediction is the average of the B predictions: $\widehat{f}^{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f}_b(x)$, where $\widehat{f}_b(x)$ is the prediction from tree b on the test sample with covariate x. And for classification, the final prediction is made from the majority vote of the B trees. Since bagging is the mean of many different estimators, it can reduce the variance and enhance the out-of-sample prediction compared

[23] It captures the node purity: if a node contains predominant observations from a single class, then the Gini index tends to be close to zero.

to a single-tree model. Moreover, in bagging, each tree is grown deep, in the sense that we do not need to prune the trees to avoid overfitting, because via bagging, we can ensemble many weak learners and generate a strong learner.

RF generalizes the idea of bagging by bringing randomness to the set of predictors in each tree. In particular, we randomly select m predictors out of the total set of p predictors in each tree. Then like bagging, we make predictions for each tree, and the final prediction is the average of each single tree prediction. The motivation is that there might be (potentially strong) correlation between the B trees. For example, when there is a strong predictor, then most trees will use this predictor to do the first split, and finally most trees will look similar. RF aims at decorrelating the trees and reducing the variance of the final prediction. A recommendation for number of features to be used in each tree is $m = p/3$ for regression and $m = \sqrt{p}$ for classification [40].

I implement RF using function *randomForest* in package **randomForest** in R. I use *ntree*=1000 trees.

2.6.3.4 Boosting

Boosting is a general ensembling method to overcome the overfitting problem. Here I consider boosting based on the decision trees. The idea is to grow trees sequentially, instead of independently as in bagging. Therefore, boosting learns slowly. Specifically, we try to fit the current residuals from a tree in each step, instead of the outcome. Each tree could be rather small, so we are improving the prediction slowly in areas where it does not perform well. The shrinkage parameter λ even slows the process, allowing more and different shapes of trees to attack the residuals. In general, statistical learning approaches that learn slowly tend to perform well. The typical algorithm of boosting is as follows:

1. Set $\widehat{f}(x) = 0$ and $r_i = y_i$ for all i.

2. For $b = 1, 2, ..., B$, repeat:

   (a) Fit a tree $\widehat{f}^b$ with d splits (base learner) to the training data (X, r);

(b) Update $\widehat{f}$ by adding in a shrunken version of the new tree:

$$\widehat{f}(x) = \widehat{f}(x) + \lambda \widehat{f}^b(x). \tag{2.12}$$

(c) Update the residuals:

$$r_i = r_i - \lambda \widehat{f}^b(x_i). \tag{2.13}$$

3. The final prediction of the boosting model is $\widehat{f}(x) = \sum_{b=1}^{B} \lambda \widehat{f}^b(x)$.

I implement boosting using function *boosting* in package **adabag** in R.

2.6.3.5 Support Vector Machine

Support Vector Machine (SVM) is often considered one of the best out of the box classifiers [38]. Consider a binary classification problem where the response $y_i \in \{-1, 1\}$. The idea is to find a hyperplane $f(x)$ which could separate the data into two classes according to the response y, such that the *margin* (distance from the hyperplane to the closest data point) is maximized. The SVM allows some training data points to be on the wrong side of the hyperplane, even the wrong side of the margin, in order to overcome the overfitting problem and obtain a more robust classification. This is known as *soft margin*. The optimization problem is:

$$\begin{aligned} &\max_{\beta,\epsilon} M, \;\; s.t. \;\; y_i f(x_i) > M(1-\epsilon_i), \\ &\sum_{k=1}^{p} \beta_k^2 = 1, \;\; \epsilon_i \geq 0, \;\; \sum_{i=1}^{n} \epsilon_i \leq C. \end{aligned} \tag{2.14}$$

C is a non-negative tuning parameter, which determines the severity of violations to the margin (and the hyperplane). Think of a budget for the violations. As C increases, our tolerance increases, so the margin widens. C controls the bias-variance trade-off of this statistical learning process. When C is too small, we tend to overfit the data and lead to high variance. M is the width of the margin, which we want to maximize. ϵ's are the slack variables that allow individuals to be on the wrong side of the margin or the hyperplane. We make the prediction of a test sample x via the sign

of $f(x)$.

Observations that lie directly on the margin, or on the wrong side of the margin, are called *support vectors*. They are the only data that affect the hyperplane, and hence the classifier. Such robustness for observations that are far away from the hyperplane, is distinct from other classification methods.

SVM also allows for the nonlinear boundary: $f(x) = \alpha + h(x)'\beta$. It turns out that the algorithm involves $h(\cdot)$ only through the inner product $\langle h(x), h(x')\rangle$, instead of specifying h, so we could just provide the kernel function $K(x, x') = \langle h(x), h(x')\rangle$.

It can be shown that SVM could be solved with the optimization problem, which is a penalized regression with hinge loss and L_2 penalty.

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \max\{0, 1 - y_i f(x_i)\} + C \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{2.15}$$

I implement SVM using the **rminer** package, which calls function *ksvm* in package **kernlab**.

2.6.3.6 Neural Nets

The Neural Nets (NN) classification model could be regarded as a non-linear generalization of the multinomial logit model. Here we present a basic NN model, which is a single hidden layer feed-forward NN. Suppose we have response $Y \in \{0, 1\}$ and $X \in \mathbb{R}^p$. The algorithm is:

1. $Z_m = \sigma(X'\alpha), m = 1, ..., M$, where Z_m is a *neuron* in the hidden layer, which is the non-linear transformation of a linear combination of the features X. The *activation function* σ is often chosen to be the sigmoid function $\sigma(v) = 1/(1 + e^{-v})$.

2. $T_k = Z'\beta_k, k = 0, 1$.

3. $f_k(X) = g_k(T), k = 0, 1$, where g_k is the soft-max function $g_k(V) = e^{V_k}/(e^{V_1} + e^{V_2})$. The predicted probability is $\widehat{\mathbb{P}}[Y = 1|X] = f_1(X)$.

We see that if σ is the identity function, then NN collapses to the linear logit model. We can think of the hidden layer as a data-driven way of selecting basis function in the nonparametric series

or spline regression, while in standard series and spline methods, the basis functions are chosen ex-ante. The estimation of NN is usually conducted via *back-propagation* and *stochastic gradient descent*.

I implement NN using the **rminer** package, which calls function *nnet* in the **nnet** package.

Table 2.6 shows the variable importance for the four machine learning methods.

Table 2.6: Variable Importance of Machine Learning Methods

Variable Importance	RF	Boosting	SVM	NN
Number of Rounds: t	3602.53	30.0646	0.2313	0.0707
Eligibility: elig_i	1647.16	0.0999	0.1129	0.1250
Population: $\mathrm{pop}(j)$	2424.56	0.2285	0.0323	0.0950
Number of Bidders: $n_i^t(j)$	4117.50	58.6597	0.3015	0.3729
Complementarity Effect: $c_i^t(j)$	3221.71	0.0999	0.0782	0.1288
Price: $p_i^t(j)$	3176.05	1.2154	0.0227	0.1136
Number of Bids: nbid_i^t	2044.16	0.0000	0.1557	0.0235
Winning Rounds: $\mathrm{rwin}_i^t(j)$	3464.92	9.6320	0.0653	0.0705

For RF, the variable importance is measured by the mean decrease in Gini index. For Boosting, the measure of importance takes into account the gain of the Gini index given by a variable in a tree and the weight of this tree in the case of boosting. For SVM and NN, the variable importance is calculated via the sensitivity analysis.

From Table 2.6, we see that across all methods, $n_i^t(j)$ is the most important predictor for the winning probability, which coincides with the logit and LASSO estimation, as well as our theoretical analysis. In addition, most methods agree that $\mathrm{pop}(j)$ and nbid_i^t are of the least crucial to the prediction, which is also concluded from the previous two linear models.

2.6.3.7 Model Averaging

Following [23], our model averaging scheme is:

1. For each sample observation i, make predicted probabilities using the six models proposed

above, denoted as $(\widehat{y}_{1i}, ..., \widehat{y}_{6i})$. We use a 5-fold cross-validation to train the models and make predictions.

2. Run a linear regression the observed response y_i on the predicted probabilities $(\widehat{y}_{1i}, ..., \widehat{y}_{6i})$, with the constraint that all estimates are non-negative and sum up to 1. Obtain estimated coefficients $(\widehat{w}_1, ..., \widehat{w}_6)$.

3. The model averaging predicted probability on a test point is the weighted average of the six predicted probabilities, using $(\widehat{w}_1, ..., \widehat{w}_6)$ as the weights:

$$\widehat{q}_i^t(j) = \widehat{q}(j|\mathcal{I}_i^t) = \sum_{k=1}^{6} \widehat{w}_k \widehat{q}_k(j|\mathcal{I}_i^t), \ \ \forall j \in J. \tag{2.16}$$

I implement this procedure using **ForecastComb** package. Table 2.7 shows the 5-fold cross-validated predictive performance of the six individual methods and the model averaging prediction, as well as the estimated weight in the model averaging. The measure of model performance is the Root Mean Square Eroor (RMSE) and Misclassificiation Rate (MCR) on the (cross-validation) test set.

Table 2.7: Cross Validation and Model Averaging

Model	RMSE	MCR	Weight (%)
Logit	0.3804	0.2174	0.00
LASSO	0.3807	0.2172	0.00
RF	0.3629	0.2002	50.11
Boosting	0.3680	0.2068	8.62
SVM	0.3807	0.2028	11.84
NN	0.3636	0.2057	29.43
Combined	0.3572	0.1915	100

We see that in terms of RMSE and MCR, logit and LASSO have similar performances, and are beaten by the other four nonparametric machine learning methods. Within the last four machine

learning algorithms, we see that RF turns to the best model. In the weight of the model averaging estimation, we find that logit and LASSO receive zero weights, meaning that they are universally outperformed by the rest four predictive models. RF obtain the largest weight, which is consistent with its superiority over all other algorithms. NN turns out to be the second best model, followed by SVM and Boosting. The final model averaging prediction further improves the RMSE and MCR upon each single method. Compared to the traditional linear logit model, the combined model has a 2.32% decrease in RMSE, and 2.59% decrease in MCR.

We also explore several existing model averaging schemes. In terms of RSME, the best scheme is the OLS Model Averaging (OLS-MA), which yields the RSME of 0.3571. However, the OLS-MA can generate arbitrary weights, so that the combined predictive winning probabilities will no longer be a well-defined probability distribution. In this case, the weights selected by OLS-MA is $w = (-0.0467, -0.0051, 0.4948, 0.1193, 0.1505, 0.3169)$. While the heavy weight on RF, NN, and SVM are consistent with the constrained least square model averaging (CLS-MA), OLS-MA assigns negative weights to Logit and Lasso models. On the other hand, the RMSE from CLS-MA is very close the that of OLS-MA. The RMSE and MCR of OLS-MA are 0.3570 and 0.1914, respectively, which are just slightly lower than CLS-MA. Therefore, we use the CLS-MA scheme to combine our six models.

2.6.4 Estimating the Choice Probabilities

2.6.4.1 The High Dimensionality Issue

In this subsection, we discuss the estimation of the choice probabilities over different set of licenses for each bidder at a round. In particular, We are going to estimate $\mathbb{P}[\boldsymbol{Y}_i^t = L|\mathcal{I}_i^t]$, that is, the probability that, at round t bidder i chooses the set of licenses $L \in \mathcal{L}$ to bid, where $\mathcal{L}$ is all the combinations of licenses that bidders could consider.

The number of all possible sets of licenses would be huge. Since we have 480 licenses, then the total number of subsets of the 480 licenses would be 2^{480}, which far exceeds the computation power and storage of any ordinary computer. First, we need to restrict our attention to the sets of licenses

that bidders would really consider. Having a consideration set with the size 2^{480} is infeasible for the bidders, both intuitively and computationally. Instead, there must be only a part of the bundles that bidders are interested in. One of the distinct features of Simultaneous Ascending Auction (SAA) compared to Combinatorial Auction (CA), is that SAA allows the bidders to construct bundles of objects by themselves, so that the potentially beneficial combinations are revealed during the auction. So we set the considerations set $\mathcal{L}$ to be the collection of sets of licenses L_i^t that have been constructed during the course of the auction (including the empty set). That is, we believe that a combination of licenses that has ever been proposed by any bidder, is potentially acceptable by the bidders, and during the long course of this auction (183 rounds), the underlying possible combinations have all been raised. From that, we obtain our consideration set $\mathcal{L}$ with $d = |\mathcal{L}| = 3998$ different bundle of licenses.

However, 3998 choices is still very high-dimensional for a multinomial discrete choice problem. The computation for a multiclass classification algorithm a multinomial logit with this large number of classes will still be extremely expensive. In regard of that, we leverage the multivariate logit model in [14], which could provide a consistent estimator of the choice probabilities, while saving a significant time of computation. The idea is that, instead of modeling the joint distributions of the 480-dimensional multivariate Bernoulli variable and take it to a traditional Maximum Likelihood Estimation, we only use the conditional distributions and conduct a Composite Conditional Likelihood Estimation. Although we lose part of the information by only including the conditional distributions and therefore lose some efficiency, the estimators are still consistent.

2.6.4.2 The Model and Interpretation

Specifically, let $\mathbf{Y}_i^t = (Y_i^t(j))_{j \in J}$ be a multivariate Bernoulli variable. If $Y_i^t(j) = 1$, then bidder i has chosen license j at round t. We model the conditional probabilities of choosing j, given the selection outcome of the rest licenses being $\boldsymbol{y}_i^t(-j)$, as

$$\mathbb{P}[Y_i^t(j) = 1 | \mathbf{Y}_i^t(-j) = \boldsymbol{y}_i^t(-j), \mathcal{I}_i^t] = \frac{\exp(Z_i^t(j))}{1 + \exp(Z_i^t(j))}, \tag{2.17}$$

with

$$Z_i^t(j) = W_i^t(j)'\gamma + \phi \sum_{j' \neq j} y_i^t(j')\tau(j, j'). \tag{2.18}$$

Here $W_i^t(j)$ is a vector of covariates that only concern license j, and $\tau(j, j')$ is the complementarity measure defined in (2.9). We use the same covariates for $W_i^t(j)$ (except for the complementarity effect $c_i^t(j)$) in Step 1. The parameter ϕ captures the (pairwise) complementarity effect between j and other chosen licenses in the set L. This is consistent with our pairwise complementarity specification in bidders' utility function.

[41] show that the conditional distribution implies that the joint distribution of $\boldsymbol{Y}_i^t$ is

$$\pi_i^t(L) = \pi(L|\mathcal{I}_i^t) = \mathbb{P}[\boldsymbol{Y}_i^t = L|\mathcal{I}_i^t] = \frac{\exp(\mu_i^t(L))}{1 + \sum_{L \in \mathcal{L}} \exp(\mu_i^t(L))}, \tag{2.19}$$

where

$$\mu_i^t(L) = \sum_{j \in L} W_i^t(j)'\gamma + \phi \sum_{j \neq j'; j, j' \in L} \tau(j, j'). \tag{2.20}$$

It is consistent with the multinomial logit model, where bidder i's "current-round utility" for making decisions at round t on set L is $u_i^t(L) = \mu_i^t(L) + \epsilon_i^t(L)$, with $\epsilon_i^t(L)$ being i.i.d. type I extreme value distributed.[24] The deterministic part of "current-round utility" is composed of two parts. The first part $\sum_{j \in L} Wi^t(j)'\gamma$ is the sum of "current-round" stand-alone values of each license in set L, with covariates including $\mathrm{elig}_i, \mathrm{pop}(j)$, which consists of bidders' static stand-alone value defined in (2.7), and $t, n_i^t(j), \mathrm{nbid}_i^t, \mathrm{rwin}_i^t(j), p_i^t(j)$, which are the current-round information. The second part $\phi \sum_{j \neq j'; j, j' \in L} \tau(j, j')$ is the complementarity value among set L, which is of the same form as (2.2) and (2.8).

Therefore, we can also think of (2.20) a *penalized* utility:

$$\mu_i^t(L) = v_i(L) - \sum_{j \in L} p_i^t(L) + g(h_i^t), \tag{2.21}$$

where $v_i(L)$ is the private value, and $g(h_i^t)$ is the penalty term on the current-round utility based

[24] This is consistent with the Multivariate Bernoulli model by [42] with second order interactions.

on the current-round observed bidding history $h_i^t = (t, n_i^t, \mathrm{nbid}_i^t, \mathrm{rwin}_i^t(j))$. Here we normalize the coefficient on price to be -1, so the other coefficients are measured in the unit of (million) dollars. By our theoretical analysis Section 2.3, we expect that with more competitors, bidders may discount more on their "current-round utility", and thus the coefficient on $n_i^t(j)$ should be negative.

2.6.4.3 The Estimation Method and Results

As we have mentioned, using the joint likelihood function (2.19) for Maximum Likelihood Estimation would involve a large amount of computation. Instead, we focus on the conditional distributions (2.17), and conduct Composite Conditional Likelihood (CCL) estimation [43]. Specifically, our composite conditional log-likelihood function is

$$\ell(\gamma, \phi) = \sum_{i,t} \sum_{j=1}^{m} \log \mathbb{P}[Y_i^t(j) = y_i^t(j) | \boldsymbol{Y}_i^t(-j) = \boldsymbol{y}_i^t(-j), \mathcal{I}_i^t]. \tag{2.22}$$

[43] show that the estimator $(\widehat{\gamma}, \widehat{\phi})$ that maximizes $\ell(\gamma, \phi)$ is consistent. Such optimization could be processed very fast, and the standard errors can be computed from the standard way using information matrix.

Because a bidder at one round only places bids on a small subset of J, there are much more zeros (not chosen licenses, $y_i^t(j) = 0$) than ones (chosen licenses, $y_i^t(j) = 1$) in the response of the likelihood function. Such *rare event* data will dramatically slow the computation and occupy the storage of the machine, while in fact the rare events (not chosen licenses) are much less informative than the events (chosen licenses). In view of that, we randomly sample the data of "not chosen license", and keep all the data of "chosen license", and make the proportion of zeros of $y_i^t(j)$ around 90%. This is referred to as *choice-based sampling* in [44], which also pointed out that such sampling may cause bias in the estimated intercept. I use [45]'s method to correct for this bias, which is implemented in the **Zelig** package in R. The result of the choice probability estimation is in Table 2.8.

We see that the coefficient on $n_i^t(j)$ is significantly negative, which coincides with what our

Table 2.8: Choice Probabilities

Choice Prob	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|z\|)$
(Intercept)	-8.3596	0.3711	-22.53	0.0000
Number of Rounds: t	-0.0025	0.0048	-0.52	0.6057
Eligibility: elig_i	-0.3310	0.0181	-18.24	0.0000
Population: $\mathrm{pop}(j)$	61.1494	1.5588	39.23	0.0000
$\mathrm{elig}_i \times \mathrm{pop}(j)$	0.8548	0.0258	33.08	0.0000
Number of Bidders: $n_i^t(j)$	-1.0619	0.1284	-8.27	0.0000
Number of Bids: nbid_i^t	0.0270	0.0093	2.92	0.0035
Winning Rounds: $\mathrm{rwin}_i^t(j)$	0.3925	0.0195	28.558	0.0000
Complementarity: $\sum_{j \geq j';j,j' \in L} \tau(j,j')$	28.4948	2.8042	10.16	0.0000

theory expects. For the other covariates in h_i^t, we observe that nbid_i^t and $\mathrm{rwin}_i^t(j)$ both have significantly positive effect, meaning that a more aggressive and successful bidding history may encourage bidders to bid at the current round. The number of rounds has insignificant and negligible effect. We find strong complementarity effect in bidders' private value between licenses. For bidders' stand-alone values, the license's population plays a crucial role, while the bidder characteristic elig_i and the interaction term only have relatively small effects.

We then use the estimated parameters to calculate the predicted choice probabilities $\widehat{\pi}_i^t(L) = \widehat{\pi}(L|\mathcal{I}_i^t) = \widehat{\mathbb{P}}[\boldsymbol{Y}_i^t = L|\mathcal{I}_i^t]$ for all sets $L \in \mathcal{L}$ for every (i,t) observation, using (2.19). We evaluate the predictive performance in Table 2.9.

Table 2.9: Evaluation of Predicted Choice Probabilities

Rank $(\leq)$	**1**	**2**	**3**	**4**	**5**	**10**	**100**
Proportion	0.2173	0.3045	0.3496	0.3778	0.4054	0.4623	0.6323

In Table 2.9, we calculate for each observation (i,t), what is the rank of the true chosen set L_i^t among the consideration set $\mathcal{L}$, based on the predicted choice probabilities. There are 21.73%

observations making the correct prediction, which assign the highest probability to the actually chosen set. There are 40.54% observations ranking the actually chosen bundle at top 5, and 46.23% at top 10. Since we are effectively conducting a multiclass classification with 3998 classes, we regard this result as a satisfactory one.

2.6.5 Dimension Reduction Using Random Projection

From the previous two steps, we obtain the estimated expected current round utility, and the estimated choice probability, which are both of dimension $d = |\mathcal{L}|$ for each (i, t) observation. In this subsection, we apply the Random Projection (RP) method in [13] to reduce the dimension (on the number of choices) in this multinomial discrete choice problem. RP is a popular machine learning method for dimension reduction. This dimension reduction is achieved by conducting a linear projection on the high-dimensional vector to a low dimensional space, and such linear transformation is defined via a random matrix. By the Johnson-Lindenstrauss lemma, the Euclidean distance between original vectors on the high-dimensional space will be preserved to the low-dimensional space after RP, with large probability. On the other hand, in the last step of our estimation, the optimization objective function only involves Euclidean distance between data points. Therefore, RP could be used to do dimension reduction for our purpose, and it greatly saves the computational burden in the last step.

Particularly, let R be a random projection matrix with size $k \times d$, where $d = |\mathcal{L}|$, and $k < d$ is the projected down dimension. Let the deterministic part of the expected current round payoff be:

$$\begin{aligned}
\nu_i^t(L;\theta) &= \mathbb{E}[\overline{u}_i^t(L)|q, \mathcal{I}_i^t; \theta] \\
&= \sum_{j\in L}(\alpha_0 + \alpha_1 \operatorname{elig}_i + \alpha_2 \operatorname{pop}(j) + \alpha_3 \operatorname{elig}_i \operatorname{pop}(j))\widehat{q}_i^t(j) \\
&+ \frac{1}{2}\beta \sum_{j'\in L} \sum_{j'\neq j, j'\in L} \tau(j, j')\widehat{q}_i^t(j)\widehat{q}_i^t(j') - \sum_{j\in L} p_i^t(j)\widehat{q}_i^t(j) \\
&\equiv X_i^t(L)'\theta
\end{aligned} \tag{2.23}$$

where $\widehat{q}_i^t(j)$ are estimated from Step 1, and $\theta = (\alpha', \beta, -1)'$. For a $d \times 1$ vector $\mathbf{z}$, define $\widetilde{\mathbf{z}} = R\mathbf{z}$.

Then the projected down data is $(\widetilde{\mathbf{X}}_i^t, \widetilde{\widehat{\boldsymbol{\pi}}_i^t})$.

Following [13], we use [46]'s sparse random projection matrix as R, where a large number of elements are zero with high probability. Sparse random projection matrices are an attractive alternative to dense Gaussian random projection matrix that guarantees similar embedding quality while being much more memory efficient and allowing faster computation of the projected data. The sparse random projection matrix in [46] is:

$$R_{ij} = \sqrt{\frac{s}{k}} \begin{cases} 1, & \text{with probability} \quad 1/2s \\ 0, & \text{with probability} \quad 1 - 1/s \\ -1, & \text{with probability} \quad 1/2s \end{cases}, \tag{2.24}$$

where k is the dimension of the projected down space, and $s = \sqrt{d}$. We see that Li's random projection exercises a random subsampling or bootstrapping of the original data (some are sampled directly, some are sampled with a negative sign). In fact, it is a random drawing from both the dimension of observation and feature, so it is somewhat similar to a random forest algorithm.

The computational enhancement brought by RP is significant. For the optimization problem in the next step, if we use the original data $(\boldsymbol{X}_i^t, \widehat{\boldsymbol{\pi}}_i^t)$, it takes twice the time to compute the objective function compared to using the RP with $k = 300$. More importantly, directly using the original data takes more than 500 iterations for the optimization algorithm to converge, which is very expensive in terms of making inference. On the other hand, if we use the projected down data $(\widetilde{\mathbf{X}}_i^t, \widetilde{\widehat{\boldsymbol{\pi}}_i^t})$, it only takes around 20 iterations for the algorithm to find the optimum point.

2.6.6 Estimating the Structural Parameters Using Cyclic Monotonicity

In this subsection, we discuss the final step of our structural estimation, which is the estimation of our structural parameters of interest $\theta = (\alpha', \beta)'$ in (2.10). We use the method proposed in [12], which is a semiparametric estimation approach for multinomial discrete choice in panel data. It does not need to assume a specific distribution for the errors in the random utility model, and moreover, allows flexible dependence structure of errors among choices and across time. We

could also include individual fixed effects in bidders' utility function. This estimation is based on the *cyclic monotonicity* property of the choice probability function in the multinomial discrete choice model, and we use a series of inequalities from the cyclic monotonicity to construct the optimization objective function. Since these inequalities only concern Euclidean distance between the data vectors, RP could be applied to reduce the dimension.

2.6.6.1 Cyclic Monotonicity

Let bidders' expected current round payoff be

$$\mathbb{E}[u_i^t(L)|\boldsymbol{q}_i^t, \mathcal{I}_i^t] = \nu_i^t(L;\theta) + \epsilon_i^t(L), \tag{2.25}$$

where $\epsilon_i^t(L)$ is the error term, which may include the unobserved variables, random shocks, measurement errors, and estimation errors from previous steps.

Bidders choose the set L_i^t which maximizes their current-round expected payoff $\mathbb{E}[u_i^t(L)|\boldsymbol{q}_i^t, \mathcal{I}_i^t]$. Thus, the choice probability function is

$$\pi(L|\boldsymbol{\nu}_i^t) = \mathbb{P}[\nu_i^t(L) + \epsilon_i^t(L) \geq \max_{L' \neq L, L' \in \mathcal{L}} \nu_i^t(L') + \epsilon_i^t(L')]. \tag{2.26}$$

Let $\boldsymbol{\pi}(\boldsymbol{\nu}) = (\pi(L|\boldsymbol{\nu}))_{L\in\mathcal{L}}$, and $\boldsymbol{\nu} = (\nu(L))_{L\in\mathcal{L}} \in \mathcal{U} \subseteq \mathbb{R}^d, \boldsymbol{\epsilon} = (\epsilon(L))_{L\in\mathcal{L}}$. Suppose $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$ are independent, then $\boldsymbol{\pi}$ is *cyclic monotone* in $\mathcal{U}$.

Definition 2. *Consider a function $\boldsymbol{f} : \mathcal{F} \to \mathbb{R}^d$ with $\mathcal{F} \subseteq \mathbb{R}^d$. Take a length M-cycle in $\mathcal{F}$, denoted as $(\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_M, \boldsymbol{u}_1)$. Then $\boldsymbol{f}$ is called* cyclic monotone *with respect to the cycle $(\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_M, \boldsymbol{u}_1)$ if*

$$\sum_{m=1}^{M} (\boldsymbol{u}_m - \boldsymbol{u}_{m+1}) \cdot \boldsymbol{f}(\boldsymbol{u}_m) \geq 0. \tag{2.27}$$

If $\boldsymbol{f}$ is cyclic monotone with respect to all possible cycles of all lengths on $\mathcal{F}$, then we say $\boldsymbol{f}$ is cyclic monotone on $\mathcal{F}$.

The reason that choice probability function $\boldsymbol{\pi}$ is *cyclic monotone* is as follows. The independence between $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$ implies that the social surplus function [47] $S(\boldsymbol{\nu}) = \mathbb{E}[\max_{L\in\mathcal{L}}(\nu(L) +$

$\epsilon(L))|\boldsymbol{\nu}]$ is convex in $\boldsymbol{\nu}$. In addition, the choice probability function lies in the sub-gradient of the social surplus function: $\boldsymbol{\pi}(\boldsymbol{\nu}) \in \partial S(\boldsymbol{\nu})$. And it is known from the convex analysis [48] that, the sub-gradient of a convex function satisfies *cyclic monotonicity*.

Remark 1. *For a univariate convex and differentiable function, its gradient (derivative) is monotonically non-decreasing. Therefore, the* cyclic monotonicity *could be viewed as an appropriate extension of this feature to multivariate convex function.*

Remark 2. *Take $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathcal{U}$, we have*

$$(\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2) \cdot \boldsymbol{\pi}(\boldsymbol{\nu}_1) + (\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1) \cdot \boldsymbol{\pi}(\boldsymbol{\nu}_2) \geq 0. \tag{2.28}$$

Rearranging the terms, we get

$$\boldsymbol{\nu}_1 \boldsymbol{\pi}(\boldsymbol{\nu}_1) + \boldsymbol{\nu}_2 \boldsymbol{\pi}(\boldsymbol{\nu}_2) \geq \boldsymbol{\nu}_1 \boldsymbol{\pi}(\boldsymbol{\nu}_2) + \boldsymbol{\nu}_2 \boldsymbol{\pi}(\boldsymbol{\nu}_1). \tag{2.29}$$

The left hand side is the sum of expected utility for the two deterministic utilities $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2$, using the true choice probabilities. The right hand side is the sum of expected utility using the wrong choice probabilities, because they exchange the choice probabilities with each other. Therefore, the cyclic monotonicity for the length 2-cycle could be interpreted as that, using the true choice probabilities will always generate higher expected utility among the choices than exchanging the choice probabilities. This shares the similar idea with the pairwise stability property in the two-sided matching model [49].

In a general M-cycle, we have

$$\sum_{m=1}^{M} \boldsymbol{\nu}_m \boldsymbol{\pi}(\boldsymbol{\nu}_m) \geq \sum_{m=1}^{M} \boldsymbol{\nu}_m \boldsymbol{\pi}(\boldsymbol{\nu}_{m+1}), \tag{2.30}$$

which could be interpreted similarly.

2.6.6.2 Panel Data

Now consider the panel data environment as in our case. We only assume the errors to be identically distributed for different rounds. Without loss of generality, let $T \geq t_2 > t_1 \geq 1$. Let A_i be the bidder fixed effect in bidder i's value function, that is not taken into the expectation with respect to the winning probabilities.

Assumption 3. $\boldsymbol{\epsilon}_i^{t_1} \sim \boldsymbol{\epsilon}_i^{t_2} | (\boldsymbol{\nu}_i^{t_1}, \boldsymbol{\nu}_i^{t_2}, A_i), \forall i \in N, \forall t_1, t_2.$

Remark 3. *(i) First, we allow for time dependent shocks, for example, some outside information coming in for the licenses during each round t, possibly correlated across rounds, while we assume to have the same marginal distribution.*

(ii) In addition, we do not impose any distributional assumption on the joint distribution of $(\epsilon_i^t(L))_{L \in \mathcal{L}}$, meaning that the marginal distribution of the unobservable term in the expected current round utility can be arbitrary, and more importantly, they can be arbitrarily dependent of each other. Such flexibility is crucial to us, because different sets L and L' may have overlap of the same licenses, and thus their corresponding unobservables may neither come from the same marginal distribution, nor be independent of each other.

(iii) Furthermore, we allow for bidder fixed effect A_i.

(iv) Finally, the errors are permitted to be correlated with the covariates X_i^t , as well as the bidder fixed effect A_i. So we do not need to worry about endogeneity here.

Conclusively, the semiparametric feature of this estimation method is necessary in our setting. Given Assumption 3, by the cyclic monotonicity of $\boldsymbol{\pi}$, we have

$$(\boldsymbol{\nu}_i^{t_1} - \boldsymbol{\nu}_i^{t_2}) \cdot [\boldsymbol{\pi}(\boldsymbol{\nu}_i^{t_1}) - \boldsymbol{\pi}(\boldsymbol{\nu}_i^{t_2})] \geq 0. \tag{2.31}$$

Remark 4. *(i) Note that the fixed effect has been differenced out in $(\boldsymbol{\nu}_i^{t_1} - \boldsymbol{\nu}_i^{t_2})$, and in $[\boldsymbol{\pi}(\boldsymbol{\nu}_i^{t_1}) - \boldsymbol{\pi}(\boldsymbol{\nu}_i^{t_2})]$, the fixed effect is integrated out by the Law of Iterative Expectation.*

(ii) In [12], the choice probability function appears in the cyclic monotonicity inequality is $\boldsymbol{\pi}_i^{t_1}(\boldsymbol{\nu}_i^{t_1}, \boldsymbol{\nu}_i^{t_2})$ and $\boldsymbol{\pi}_i^{t_2}(\boldsymbol{\nu}_i^{t_1}, \boldsymbol{\nu}_i^{t_2})$, which uses both information from round t_1 and t_2 to form the choice probability at the two rounds. Nevertheless, since the later round information includes the early round information, then $\boldsymbol{\pi}_i^{t_2}(\boldsymbol{\nu}_i^{t_1}, \boldsymbol{\nu}_i^{t_2}) = \boldsymbol{\pi}_i^{t_2}(\boldsymbol{\nu}_i^{t_2})$. And since the decision on the early round does not depend on later round information, then $\boldsymbol{\pi}_i^{t_1}(\boldsymbol{\nu}_i^{t_1}, \boldsymbol{\nu}_i^{t_2}) = \boldsymbol{\pi}_i^{t_1}(\boldsymbol{\nu}_i^{t_1})$.

[12] show that our parameter of interest θ in $\boldsymbol{\nu}(\theta)$ is point identified from (2.31). One essential assumption of identification is the large support condition, i.e. the support of $(\boldsymbol{\epsilon}_i^t | A_i, \boldsymbol{\nu}_i^t)$ is $\mathbb{R}^d$ with positive probability. It implies that bidders should have positive probabilities of choosing any bundle L in $\mathcal{L}$. In our estimation of choice probabilities (Section 2.6.4), the logit-type estimated choice probability (2.19) ensures that all choice probabilities are non-zero.

2.6.6.3 The Estimation Method and Results

We use the projected-down data $(\widetilde{\boldsymbol{X}}_i^t, \widetilde{\widehat{\boldsymbol{\pi}}}_i^t)$ from RP in the last subsection in our final estimation. Recall that for the original data, $\boldsymbol{X}_i^t$ is the covariates for the expected payoff constructed from the estimated winning probabilities, and $\widehat{\boldsymbol{\pi}}_i^t$ is the estimated choice probabilities. We expect that the expected payoff (with estimated winning probabilities) is consistent with the estimated choice probabilities.

For each bidder i, we pick 10 rounds that she has bid on during round 21-120.[25] Consider all the 2-cycle's within the 10 rounds. Let the set of selected 2-cycle's of bidder i be H_i. Our objective function for optimization is

$$\begin{aligned} Q(\theta) =& \sum_{i \in N} \sum_{(t_1,t_2) \in H_i} \left[\left(\widetilde{\boldsymbol{\nu}}_i^{t_1}(\theta) - \widetilde{\boldsymbol{\nu}}_i^{t_2}(\theta) \right) \cdot \left(\widetilde{\widehat{\boldsymbol{\pi}}}_i^{t_1} - \widetilde{\widehat{\boldsymbol{\pi}}}_i^{t_2} \right) \right]_- \\ =& \sum_{i \in N} \sum_{(t_1,t_2) \in H_i} \left\{ \left[(\widetilde{\boldsymbol{X}}_i^{t_1} - \widetilde{\boldsymbol{X}}_i^{t_2})' \theta \right] \cdot \left(\widetilde{\widehat{\boldsymbol{\pi}}}_i^{t_1} - \widetilde{\widehat{\boldsymbol{\pi}}}_i^{t_2} \right) \right\}_-, \end{aligned} \tag{2.32}$$

where $[z]_- = |\min(0, z)|$. Our final estimator is $\widehat{\theta} = \arg\min_{\theta \in \Theta} Q(\theta)$. [12] show that $\widehat{\theta}$ is

[25] If a bidder i bids less than 10 rounds during round 21-120, just use all the rounds she bids on.

consistent.

Table 2.10: Estimation Result with 100 Random Projections, $k = 300$

Variable	**Mean**	**Median**	$(5\%, 95\%)$	$(25\%, 75\%)$
intercept: α_0	1.73	1.64	(-14.91,20.59)	(-7.41,8.39)
eligibility: α_1	-0.10	-0.12	(-0.87,0.67)	(-0.41, 0.16)
population: α_2	245.5	247.5	(160.22,331.71)	(207.5,279.2)
elig$\times$pop: α_3	-1.08	-1.12	(-2.79,0.95)	(-1.98,-0.29)
complementarity: β	79.96	74.51	(-75.06,251.13)	(16.66,130.08)
obs	1882			

Table 2.10 shows the estimators for the parameters in bidders' value function. I find a large and significant effect of the complementarity on the private value. The complementarity of the nationwide bundle is worth 8 billion dollars for an average bidder, which equals 59.54% of the sum of final prices of all licenses (13.43 billion dollars). This is close to the finding in [11]. For an bidder with average eligibility winning the nationwide bundle, the complementarity contributes 24.46% of the private value. This number is highly consistent with the result reported in [8]. Note that we use the same dataset as [8]. Even though we take different approaches for structural analysis, we end up having similar results on the complementarity effects as in [11] and [8], which demonstrates the validity of our structural modeling and estimation methods. For the bidders' stand-alone values, I document large and significant effect of the license-characteristic. For an average bidder, a license with 1 more million population, will be valued 98.71 million dollars higher. In comparison, on average 1 more million population of a license is associated with 70.32 million dollars higher in the final price as shown in a reduced-form regression in Table 2.11. On the other hand, we find very small and insignificant effect of the bidder-characteristics, as well as the interaction between bidder-characteristics and license-characteristics. Therefore, the variation of the stand-alone values is mostly generated by the heterogeneity of licenses, and the bidder heterogeneity is not essential in bidders' private value.

The final results are consistent with the results in the estimation of choice probabilities shown in Table 2.8, which are the estimated parameters for the “current-round utility for decision making".

Table 2.11: Regression of Final Price on Population for Licenses

Final Price	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8.7861e6	7.6242e6	-11.52	0.0000
Population	70.3229	0.5459	128.82	0.0000

2.7 Bidder Heterogeneity

In this section, I explore the bidder heterogeneity in bidders' stand-alone values and complementarity values. Because the effect of bidder-characteristic on value is insignificant, it is natural to attribute the heterogeneity in bidders' values to the complementarity, instead of the stand-alone value.

2.7.1 Evidence of Bidder Heterogeneity

In this subsection I look at the complementarity values of the ultimate winning bundles of each winner in the auction, calculated from the estimation results in Section 2.6.

There are 59 winners at the end of the auction who win at least two license, therefore the winning bundles exhibit complementarity. In Table 2.12 I provide three measures to represent the complementarity contribution of the winning bundle to a winner, where complem/price is the fraction of complementarity value in the total final price of the winning bundle, complem/value is the fraction of complementarity value in the value of the winning bundle, and complem/pop is the complementarity for 1 unit (1% of the US population) of population.

In Figure 2.5 I plot the three measures of complementarity contribution with respect to bidders' eligibility. We see that for all the winners, the complementarity contribution is increasing with eligibility. More importantly, for large and small bidders, the trends are quite different.

Therefore, there exists large bidder heterogeneity in the complementarity among the winners.

Table 2.12: Heterogeneity of Complementarity Value across Winners

	complem/price	complem/value	complem/pop
Min	0.0144	0.0022	0.0059
25% Qu	2.2674	0.6215	1.5618
Median	6.6004	1.3547	3.3590
Mean	8.0280	2.9934	6.7637
75% Qu	12.3799	3.2689	8.0595
Max	27.4404	22.3904	37.8290
Sd	7.1024	4.2905	8.5681
Nation	18.75	11.85	24.80
Obs	59		

Figure 2.5: Heterogeneity of Complementarity across Bidders

(a) complem/price (b) complem/value (c) complem/pop

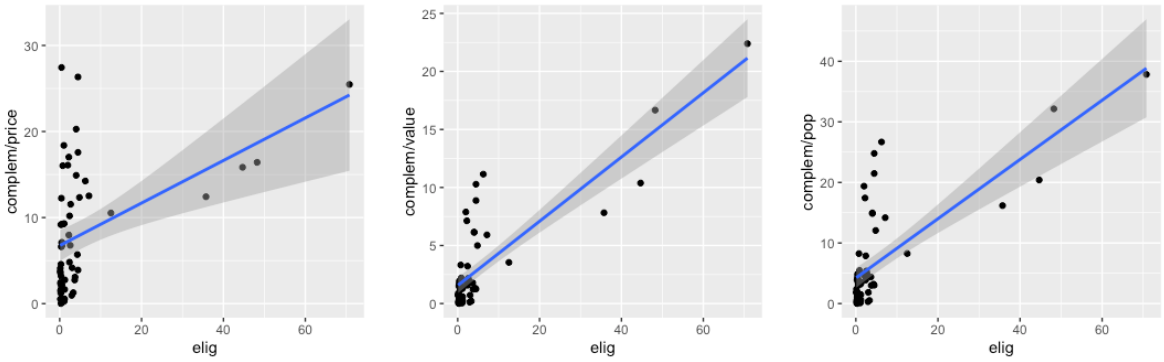


Table 2.13 summarizes their initial eligibility, profit margin, as well as the number of licenses and the total population in their winning bundles. We find a large variation among the winners and their winning results.

The variation among bidders in complementarity as well as the final winning results, is going to attribute to their private values. Since the stand-alone values are dominated by license characteristics instead of bidder characteristics, the bidder heterogeneity is supposed to reveal in the complementarity. I will estimate the bidder-heterogeneous complementarity effects in the next subsection.

Table 2.13: Winners and Their Bundles

	elig	# license	totalpop	margin
Min	0.01	1	0.01	-20.55
25% Qu	0.22	1	0.06	13.24
Median	0.58	3	0.22	37.44
Mean	3.8	5.6	1.15	88.83
75% Qu	2.51	6	0.59	91.33
Max	70.78	56	37.71	1350.08
Sd	10.77	8.09	4.34	175.19
Total	334.61	493	100	7817.09
Obs	88			

2.7.2 Estimation of Bidder Heterogeneity

To explore the bidder heterogeneity, I separately estimate the model for large, medium, and small bidders. I distinguish the types of bidders in terms of their initial eligibility: large bidders refers to those with eligibility greater than 10, small bidders are those less than 1, and medium bidders are the rest. The unit of eligibility is the total population (measured in the percentage point of the total US population) that a bidder could bid for, at the beginning of the auction. In our full sample of rounds 21-120, there are 12 large bidders, 54 medium bidders, and 133 small bidders.

Table 2.14, 2.15, and 2.16 shows the estimation results.

Table 2.14: Large Bidders

Variable	**Mean**	**Median**	$(5\%, 95\%)$	$(25\%, 75\%)$
intercept: α_0	3.13	7.07	(-82.75, 74.20)	(-28.03,36.05)
eligibility: α_1	-0.17	-0.16	(-1.56, 1.43)	(-0.75,0.39)
population: α_2	219.10	209.32	(-83.76, 572.71)	(68.42,345.42)
elig$\times$pop: α_3	-1.08	-0.99	(-5.79,3.30)	(-3.25,0.81)
complementarity: β	175.55	168.78	(-179.55, 508.82)	(51.31,328.08)

We find strong heterogeneity in bidders' complementarity values. For an average large bidders,

Table 2.15: Medium Bidders

Variable	**Mean**	**Median**	$(5\%, 95\%)$	$(25\%, 75\%)$
intercept: α_0	-5.954	-6.076	(-34.29, 21.45)	(-14.17,4.67)
eligibility: α_1	1.16	0.98	(-5.73, 9.03)	(-2.07,3.76)
population: α_2	269.7	264.6	(168.83, 377.50)	(219.4,318.3)
elig$\times$pop: α_3	-6.75	-6.55	(-30.33,17.19)	(-15.05,3.86)
complementarity: β	42.21	41.73	(-93.98, 236.82)	(-21.45,94.26)

Table 2.16: Small Bidders

Variable	**Mean**	**Median**	$(5\%, 95\%)$	$(25\%, 75\%)$
intercept: α_0	4.93	1.19	(-26.68, 40.83)	(-10.17,20.69)
eligibility: α_1	-2.12	0.65	(-68.39, 51.80)	(-22.91,21.59)
population: α_2	264.05	259.07	(133.21, 387.63)	(220.09,316.05)
elig$\times$pop: α_3	-50.23	-52.44	(-179.78,89.89)	(-108.64,12.63)
complementarity: β	85.65	75.04	(-68.33, 286.02)	(11.08,146.05)

the nationwide is worth 17.56 billion dollars, while for the medium and small bidders, this number becomes 4.22 and 8.57. Nevertheless, the estimates for stand-alone values are less heterogeneous, and quite robust to the full sample estimation. For all the large, medium, and small bidders, the estimated parameters on population are highly consistent with the full sample estimate. And the initial eligibility of bidders still plays insignificant role in their stand-alone values.

Hence, it is the heterogeneity on complementarity values that explains the variation on bidding behaviors and final results across bidders. Large bidders become large winners, because they value the complementarity among licenses higher. Medium and small bidders are less competitive in the auction because for the same bundle, they generate less complementarities.

2.8 Conclusion

In this paper, I develop a structural approach to analyze the US spectrum auction and recover the bidders' private values including stand-alone values and complementarity values from the empirical data, which will serve as the fundamental elements to quantify the performance of different

mechanism designs for the spectrum allocation. I find strong evidence of complementarity among licenses. The complementarity of a national-wide license is worth 8 billion dollars for an average bidder, which is 59.54% of the sum of final prices of all licenses. For the bidders' stand-alone values, I document large and significant effect of the license-characteristic. For an average bidder, a license with 1 more million population will be valued 98.71 million dollars higher. On the other hand, small and insignificant effect of the bidder-characteristics is recorded. Therefore, the variation in stand-alone values are mostly generated from the license rather than the bidder characteristics. In addition, I find bidder heterogeneity in the complementarity. Particularly, large bidders value the complementarity effects more than small and medium bidders.

The estimation method I propose for this structural model may be of independent interest. I provide the estimation of high-dimensional bundle choice problem with individual-level data, with rich applications in other contexts, for example, the demand estimation in industrial organization.

3. OPTIMAL MODEL AVERAGING OF MIXED-DATA KERNEL-WEIGHTED SPLINE REGRESSIONS

3.1 Introduction

When conducting applied econometric and statistical analysis, the presence of model uncertainty is ignored at the practitioner's peril [50]. The two dominant approaches for tackling this source of uncertainty when conducting regression analysis involve either *model selection* or *model averaging* [51]. Model selection deals with model uncertainty by selecting one model from among a set of candidate models using a model selection criterion such as [52] *Akaike Information Criterion* (AIC) or the Schwarz-Bayes information criterion [53], by way of illustration. That is, model selection takes a set of candidate models and applies weight 1 to one model and 0 to all others. An alternative is to apply model averaging, which deals with model uncertainty by instead *averaging* over the candidate models using a model averaging criterion that applies a vector of weights to the set of candidate models typically resulting in non-zero weights being applied to each of the candidate models. The goal in model averaging is to control misspecification bias while reducing estimation variance.

There is a longstanding literature on Bayesian model averaging; see [54] for a comprehensive review. There is also a rapidly-growing literature on frequentist model averaging, including [55], [56], [57], and [58], [59], [60] and [61], among others.

Practitioners who adopt model averaging often construct a weighted average defined over a set of *parametric* candidate models, and interest typically lies in one or more parameters common to all models. Our interest here is somewhat different though, and we adopt a nonparametric perspective. Ideally we want to average over a sufficiently rich set of candidate models so that we can consistently estimate a large class of DGPs. We also want the approach to be useful for practitioners and to admit a range of predictor types including continuous and categorical, both unordered and ordered. For these reasons we consider averaging over a recently proposed regression

spline technique that admits both categorical and continuous predictors [62]. The use of regression splines is appealing from a variety of perspectives. From a practical perspective, regression splines present global approximations that are computationally attractive since they require nothing more than solving a simple weighted least squares problem. From the theoretical perspective, the use of B-spline basis functions offers the maximally differentiable spline basis while its approximation capabilities have been widely studied and are well-established. Furthermore, for those accustomed to least-squares estimation methodology and the use of polynomials for modelling nonlinearities in a regression setting, they present a familiar yet powerful alternative to the use of parametric candidate models.

Our theoretical results, which are based on the Mallows criterion, apply both to nested and non-nested regression models, and we also allow for heterogeneous errors. Related to this is the work of [63] who examines the asymptotic risk of nested least-squares averaging estimators based on minimizing a generalized Mallows criterion in a linear model with heteroskedasticity, and the work of [64] who adopt the Mallows criterion to choose the weight vector in the model averaging estimator for linear regression models with heteroskedastic errors.

The rest of this paper proceeds as follows. Section 3.2 presents the approach of [62] and then derives an optimal weighting scheme for averaging across this set of candidate models. Section 3.3 considers the finite-sample behaviour of the proposed approach relative to popular model selection criteria along with an illustrative application, while Section 3.5 presents some concluding remarks. Detailed proofs appear in Appendix B. An R package that implements the proposed approach is available for the practitioner [65].

3.2 Model Averaging of Kernel-Weighted Spline Regression

We consider a nonparametric regression model containing both continuous and categorical predictors, which we write as

$$Y_i = \mu_i + \epsilon_i = g(\bar{X}_i, \bar{Z}_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{3.1}$$

where $\bar{X}_i = (\bar{X}_{i1}, \ldots, \bar{X}_{i\bar{q}})'$ is a $\bar{q}$-dimensional vector of continuous predictors, and $\bar{Z}_i = (\bar{Z}_{i1}, \ldots, \bar{Z}_{i\bar{r}})'$ is an $\bar{r}$-dimensional vector of categorical predictors. However, we only observe a sub-vector of $(\bar{X}_i, \bar{Z}_i)$. Specifically, we assume that we observe $(X_i, Z_i), i = 1, 2, \ldots, n$, where $X_i = (X_{i1}, \ldots, X_{iq})'$ is a q-dimensional sub-vector of $\bar{X}_i$, and $Z_i = (Z_{i1}, \ldots, Z_{ir})'$ is an r-dimensional sub-vector of $\bar{Z}_i$. Thus, we require that (X_i, Z_i) be a *proper* sub-vector of $(\bar{X}_i, \bar{Z}_i)$, i.e., we need either $q < \bar{q},\ r \leq \bar{r}$, or $q \leq \bar{q},\ r < \bar{r}$; the requirement that the candidate models are misspecified is prevalent in the model averaging literature.

Letting z_l denote the l-th component of z, $l = 1, ..., r$, we assume that z_l takes c_l different values in $D_l = \{0, 1, \ldots, c_l - 1\}$, $l = 1, \ldots, r$, where c_l is a finite constant. Assume for $l = 1, 2, \ldots, q$, each X_{il} is distributed on a compact interval $[a_l, b_l]$, and without loss of generality, we take all intervals $[a_l, b_l] = [0, 1]$. Let the support of Z be $\mathcal{M}_Z = \Pi_{l=1}^{r} D_l$, and the support of X be $\mathcal{M}_X = [0, 1]^q$. Let $(Y_i, \bar{X}_i', \bar{Z}_i')_{i=1}^n$ be independent and identically distributed as $(Y_1, \bar{X}_1', \bar{Z}_1')$. To allow for possibly heteroscedastic random errors, we assume $E(\epsilon_i|\bar{X}_i, \bar{Z}_i) = 0$ and $E(\epsilon_i^2|\bar{X}_i, \bar{Z}_i) = \sigma^2(\bar{X}_i, \bar{Z}_i) \equiv \sigma_i^2$, $i = 1, \ldots, n$. We denote the conditional expectation of the dependent variable by $\mu_i = E[Y_i|\bar{X}_i, \bar{Z}_i] = g(\bar{X}_i, \bar{Z}_i)$, where $g(\cdot, \cdot)$ is an unknown smooth function. Denote $Y = (Y_1, \ldots, Y_n)'$, $\mu = (\mu_1, \ldots, \mu_n)'$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$. Using matrix notation, Equation (3.1) can be written as $Y = \mu + \epsilon$.

Our goal is to approximate μ_i, which is particularly useful for prediction, and is also a typical object of interest in the literature on model averaging estimation (e.g., [56]; [66]). To this end, we use S_n candidate nonparametric regression models to approximate Equation (3.1). Since the dimensionality of the observed predictors is finite, i.e., $q < \infty, r < \infty$, the total number of candidate models (to be averaged) S_n is finite. For $s = 1, \ldots, S_n$, let the s-th candidate model be

$$Y_i = g_{(s)}(X_{i,(s)}, Z_{i,(s)}) + e_{i,(s)}, \quad i = 1, \ldots, n, \tag{3.2}$$

where $X_{i,(s)}$ is a q_s-dimensional sub-vector of X_i, $Z_{i,(s)}$ is a r_s-dimensional sub-vector of Z_i, $g_{(s)}(\cdot, \cdot)$ is an unknown smooth function, and $e_{i,(s)} = \mu_i - g_{(s)}(X_{i,(s)}, Z_{i,(s)}) + \epsilon_i$ represents the

approximation error in the s-th model. Also, we use $\mathcal{M}_{X,(s)}$ and $\mathcal{M}_{Z,(s)}$ to denote the supports of $X_{(s)}$ and $Z_{(s)}$, respectively.

To provide an optimal weighting scheme, we first need to estimate each candidate model. To handle the presence of categorical predictors, we follow [62] to estimate $g_{(s)}(\cdot,\cdot)$ by tensor-product polynomial splines weighted by categorical kernel functions. Let $L(Z_{i,(s)}, z_{(s)}, \lambda_{(s)})$ be a product categorical kernel function and let $\mathcal{B}_{(s)}(x_{(s)})$ be the tensor-product polynomial splines, both of which will be defined below. Then the nonparametric function $g_{(s)}(x_{(s)}, z_{(s)})$ can be approximated by $\mathcal{B}_{(s)}(x_{(s)})'\beta_{(s)}(z_{(s)})$, where $\mathcal{B}_{(s)}(x_{(s)})$ is of dimension $K_{(s)} \equiv K_{n,(s)}$, with $K_{n,(s)} \to \infty$ as $n \to \infty$. We estimate $\beta_{(s)}(z_{(s)})$ by minimizing the following weighted least squares criterion:

$$\widehat{\beta}_{(s)}(z_{(s)}) = \text{argmin}_{\beta\in\mathbb{R}^{K_{(s)}}} \sum_{i=1}^{n} [Y_i - \mathcal{B}_{(s)}(X_{i,(s)})'\beta]^2 L(Z_{i,(s)}, z_{(s)}, \lambda_{(s)}). \tag{3.3}$$

Thus $g_{(s)}(x_{(s)}, z_{(s)})$ can be estimated by $\widehat{g}_{(s)}(x_{(s)}, z_{(s)}) = \mathcal{B}_{(s)}(x_{(s)})'\widehat{\beta}_{(s)}(z_{(s)})$.

With the following specifications of $L(Z_{i,(s)}, z_{(s)}, \lambda_{(s)})$ and $\mathcal{B}_{(s)}(x_{(s)})$, one can write $\widehat{\beta}_{(s)}(z_{(s)})$ defined in Equation (3.3) as a linear function of Y. First, for the univariate kernel function $l(Z_{il,(s)}, z_{l,(s)}, \lambda_{l,(s)})$, let $l(Z_{il,(s)}, z_{l,(s)}, \lambda_{l,(s)})$ equal $\lambda_{l,(s)}$ if $Z_{il,(s)} \neq z_{l,(s)}$, and equal 1 otherwise. Then the product kernel function $L(Z_{i,(s)}, z_{(s)}, \lambda_{(s)}) = \Pi_{l=1}^{r_s} l(Z_{il,(s)}, z_{l,(s)}, \lambda_{l,(s)}) = \Pi_{l=1}^{r_s} \lambda_{l,(s)}^{1(Z_{il,(s)} \neq z_{l,(s)})}$. We have the following expression for $L(Z_{(s)}, z_{(s)}, \lambda_{(s)})$:

$$\begin{aligned}
L(Z_{i,(s)}, z_{(s)}, \lambda_{(s)}) &= \Pi_{l=1}^{r_s} l(Z_{il,(s)}, z_{l,(s)}, \lambda_{l,(s)}) \\
&= \mathbf{1}(Z_{i,(s)} = z_{(s)}) + \sum_{l=1}^{r_s} \lambda_{l,(s)} \mathbf{1}(Z_{il,(s)} \neq z_{l,(s)}) \prod_{j\neq l}^{r_s} \mathbf{1}(Z_{ij,(s)} = z_{j,(s)}) + O(||\lambda_{(s)}||^2) \\
&= \mathbf{1}(Z_{i,(s)} = z_{(s)}) + \sum_{l=1}^{r_s} \lambda_{l,(s)} \mathbb{I}_{(l,(s))}(Z_{il,(s)}, z_{l,(s)}) + O(||\lambda_{(s)}||^2),
\end{aligned} \tag{3.4}$$

where $\lambda_{(s)} = (\lambda_{1,(s)}, \ldots, \lambda_{r_s,(s)})$ is the vector of bandwidths for each of the categorical predictors, $||\lambda_{(s)}||^2 = \sum_{l=1}^{r_s} \lambda_{l,(s)}^2$, $\mathbf{1}(\cdot)$ is the indicator function, and $\mathbb{I}_{(l,(s))}(Z_{il,(s)}, z_{l,(s)}) = \mathbf{1}(Z_{il,(s)} \neq z_{l,(s)}) \prod_{j\neq l}^{r_s} \mathbf{1}(Z_{ij,(s)} = z_{j,(s)})$.

Next, we specify the tensor-product polynomial splines $\mathcal{B}_{(s)}(x_{(s)})$. Let $G_{l,(s)} = G_{l,(s)}^{(m_{l,(s)}-2)}$ be

the space of polynomial splines of order $m_{l,(s)}$ and pre-select an integer $N_{l,(s)} = N_{n,l,(s)}$. Divide $[0,1]$ into $(N_{l,(s)}+1)$ subintervals $I_{j_l,l,(s)} = [t_{j_l,l,(s)}, t_{j_l+1,l,(s)}]$, $j_l = 0,\ldots,N_{l,(s)}-1$, $I_{N_l,l,(s)} = [t_{N_{l,(s)},l,(s)}, 1]$, where $\{t_{j_l,l,(s)}\}_{j_l=1}^{N_{l,(s)}}$ is a sequence of equally spaced points, called interior knots, given as

$$t_{-(m_{l,(s)}-1),l,(s)} = \cdots = t_{0,l,(s)} = 0 < t_{1,l,(s)} < \cdots < t_{N_{l,(s)},l,(s)} < 1 = t_{N_{l,(s)}+1,l,(s)} = \cdots = t_{N_{l,(s)}+m_{l,(s)},l,(s)}$$

in which $t_{j_l,l,(s)} = j_l h_{l,(s)}$, $j_l = 0,1,\ldots,N_{l,(s)}+1$, $h_{l,(s)} = 1/(N_{l,(s)}+1)$ is the distance between neighboring knots. Let $K_{n,l,(s)} = N_{l,(s)} + m_{l,(s)}$, where $N_{l,(s)}$ is the number of interior knots and $m_{l,(s)}$ is the spline order, and let $B_{l,(s)}(x_{l,(s)}) = \{B_{j_l,l,(s)}(x_{l,(s)}) : 1 - m_{l,(s)} \le j_l \le N_{l,(s)}\}'$ be a basis system of the space $G_{l,(s)}$. We define the space of tensor-product polynomial splines $\mathcal{G}_{(s)} = \otimes_{l=1}^{q_s} G_{l,(s)}$. It is clear that $\mathcal{G}_{(s)}$ is a linear space of dimension $K_{(s)} \equiv K_{n,(s)} = \Pi_{l=1}^{q_s} K_{n,l,(s)}$. Then

$$\mathcal{B}_{(s)}(x_{(s)}) = \left[\{\mathcal{B}_{j_1,\ldots,j_{p_s}}(x_{(s)})\}_{j_1=1-m_{1,(s)},\ldots,j_{p_s}=1-m_{p_s}}^{N_{1,(s)},\ldots,N_{r_s,(s)}}\right] = B_{1,(s)}(x_{1,(s)}) \otimes \cdots \otimes B_{p_s,(s)}(x_{q_s,(s)})$$

is a basis system of the space $\mathcal{G}_{(s)}$, where $x_{(s)} = (x_{l,(s)})_{l=1}^{q_s}$.

Let $\mathcal{L}_{z_{(s)}} = \mathrm{diag}\{L(Z_{1,(s)}, z_{(s)}, \lambda_{(s)}), \ldots, L(Z_{n,(s)}, z_{(s)}, \lambda_{(s)})\}$ be a diagonal matrix with $L(Z_{i,(s)}, z_{(s)}, \lambda_{(s)})$ being the i-th diagonal entries for $1 \le i \le n$, and let $\mathbf{B}_{(s)} = [\{\mathcal{B}_{(s)}(X_{1,(s)}), \ldots, \mathcal{B}_{(s)}(X_{n,(s)})\}']_{n\times K_{(s)}}$. Then $\widehat{\beta}_{(s)}(z_{(s)})$ defined in Equation (3.3) can be written as a linear function of Y, i.e.,

$$\widehat{\beta}_{(s)}(z_{(s)}) = (\mathbf{B}'_{(s)}\mathcal{L}_{z_{(s)}}\mathbf{B}_{(s)})^{-1}\mathbf{B}'_{(s)}\mathcal{L}_{z_{(s)}}Y.$$

Furthermore, we can estimate $\mu_{i,(s)}$ by

$$\widehat{\mu}_{i,(s)} = \mathcal{B}_{(s)}(X_{i,(s)})'\widehat{\beta}_{(s)}(Z_{i,(s)}) = \mathcal{B}_{(s)}(X_{i,(s)})'(\mathbf{B}'_{(s)}\mathcal{L}_{Z_{i,(s)}}\mathbf{B}_{(s)})^{-1}\mathcal{B}'_{(s)}\mathcal{L}_{Z_{i,(s)}}Y.$$

We can rewrite this expression in matrix form as $\widehat{\mu}_{(s)} = P_{(s)}Y$, where $\widehat{\mu}_{(s)} = (\widehat{\mu}_{1,(s)}, \ldots, \widehat{\mu}_{n,(s)})'$

and $P_{(s)}$ is a square matrix of dimension n with the i-th row vector being

$$\mathcal{B}_{(s)}(X_{i,(s)})'(\mathbf{B}'_{(s)}\mathcal{L}_{Z_{i,(s)}}\mathbf{B}_{(s)})^{-1}\mathcal{B}'_{(s)}\mathcal{L}_{Z_{i,(s)}}. \tag{3.5}$$

Let the weight vector $w = (w_1, \ldots, w_{S_n})'$ belong to the set $\mathcal{W} = \{w \in [0,1]^{S_n} : \sum_{s=1}^{S_n} w_s = 1\}$, and let $P(w) = \sum_{s=1}^{S_n} w_s P_{(s)}$. The model averaging estimator of μ is given by

$$\widehat{\mu}(w) = \sum_{s=1}^{S_n} w_s \widehat{\mu}_{(s)} = P(w)Y.$$

3.2.1 Weight Choice Criterion and Asymptotic Optimality

Up to now, the weight vector in $\widehat{\mu}(w)$ is unspecified. Motivated by the Mallows criterion for model averaging estimators [56], we propose a similar method for choosing the weight vector w. Let $\Omega = E(\epsilon\epsilon') = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Define the predictive squared loss by

$$L_n(w) = n^{-1}\|\widehat{\mu}(w) - \mu\|^2,$$

and the conditional expected loss by

$$R_n(w) = E[L_n(w)|\bar{X}, \bar{Z}] = n^{-1}\|P(w)\mu - \mu\|^2 + n^{-1}\,\mathrm{tr}[\Omega P(w)'P(w)], \tag{3.6}$$

where $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_n)'$ and $\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_n)'$.

Let the Mallows-type criterion function be

$$C_n(w) = n^{-1}\|P(w)Y - Y\|^2 + 2n^{-1}\,\mathrm{tr}[P(w)\Omega]. \tag{3.7}$$

It is easy to show that $R_n(w) = E[C_n(w)|\bar{X}, \bar{Z}] - n^{-1}\,\mathrm{tr}(\Omega)$, which suggests that for the optimal choice of w in the sense of minimizing $R_n(w)$, we can choose w to minimize $C_n(w)$ by noting the fact that $n^{-1}\,\mathrm{tr}(\Omega)$ does not depend on w.

Supposing for the moment that Ω is known, the optimal choice of the weight vector is given by

$$\widetilde{w} = \text{argmin}_{w \in \mathcal{W}} \; C_n(w), \tag{3.8}$$

where $C_n(w)$ is defined in Equation (3.7). Then the optimal model averaging estimator of μ is $\widehat{\mu}(\widetilde{w}) = P(\widetilde{w})Y$.

In order to provide regularity conditions for the optimal choice of the weight vector, we need to introduce some notation. Let $\xi_n = \inf_{w \in \mathcal{W}} nR_n(w)$, and let w_s^0 be an $S_n \times 1$ vector in which the sth element is one and others are zeros. Recall that q_s is the dimension of the continuous covariate $X_{i,(s)}$. Let $g_{(s)}(x_{(s)}, z_{(s)})$ be m_s times differentiable with respect to $x_{(s)}$, for every fixed $z_{(s)} \in \mathcal{M}_{Z,(s)}$ (see [67]). Let $\alpha_s = m_s/q_s$. Denote $\delta_{\max}(A)$ as the largest singular value of the matrix A. Next, we give the assumptions required for establishing the asymptotic optimality of $\widetilde{w}$ defined in Equation (3.8).

Assumption 4.

1. *For every* $s = 1, 2, \ldots, S_n$, $m_s \geq q_s$.
2. *For some integer* $N \geq 1$, $n^{N/(1+2\underline{\alpha})} \xi_n^{-2N} \sum_{s=1}^{S_n} [nR_n(w_s^0)]^N \to 0$ *(as* $n \to \infty$*) almost surely, where* $\underline{\alpha} = \min_{1 \leq s \leq S_n} \alpha_s$.

Assumption 4-1 states the required smoothness of the true conditional expectation function $g_{(s)}(x_{(s)}, z_{(s)})$ with respect to the continuous component $x_{(s)}$. Assumption 4-2 requires $\xi_n \to \infty$, implying that there is no finite approximating model whose bias is zero. It also constrains the divergence rate of $\sum_{s=1}^{S_n} [nR_n(w_s^0)]^N \to \infty$. We emphasize that the observed predictors should be a *proper* sub-vector of the true predictors. Therefore, we necessarily exclude the possibility of having the true model in our candidate models. This type of assumption is prevalent in the model averaging literature.

Assumption 5. *For every* $s = 1, 2, \ldots, S_n$,

1. $E[\mathcal{B}_{(s)}(X_{i,(s)})\mathcal{B}_{(s)}(X_{i,(s)})'] = I_{K_{(s)}}$.

2. *There exists a sequence of constants* $\zeta_0(K_{(s)})$ *satisfying* $\sup_{x_{(s)}\in\mathcal{M}_{X,(s)}} \|\mathcal{B}_{(s)}(x_{(s)})\| \leq \zeta_0(K_{(s)})$ *such that* $\zeta_0(K_{(s)})^2 K_{(s)}/n \to 0$ *as* $n \to \infty$.

Assumption 5 is commonplace in the spline regression literature (e.g., [67]).

Assumption 6.

1. *For some fixed integer* $1 \leq N < \infty$, $\max_{1\leq i\leq n} E(\epsilon_i^{4N}|\bar{X}_i, \bar{Z}_i) < \infty$ *almost surely.*

2. $\min_{1\leq i\leq n} \sigma_i^2 \geq \bar{\sigma}^2 > 0$ *almost surely.*

Assumption 6-1 is typical in the literature on model averaging estimation (see, e.g., [56], [58], [57], and [68] by way of illustration). It imposes a finite moment bound and is satisfied by Gaussian noise. Assumption 6-2 requires that the conditional variance be bounded below by a positive constant, which is commonly used in the Mallows type model averaging literature [58, 69] as well as models with heteroskedasticity [70].

Assumption 7. *For every* $s = 1, 2, \ldots, S_n$,

1. $K_{(s)} = O(n^{1/(1+2\alpha_s)})$.

2. $\sum_{l=1}^{r_s} \lambda_{l,(s)} = O((n^{-1}K_{(s)})^{1/2})$.

Assumption 7-1 is the optimal rate of $K_{(s)}$ as shown in [67]. Assumption 7-2 gives the optimal order of the categorical smoothing parameter $\lambda_{(s)}$ in the tensor-product spline regression; see [62].

Theorem 1. *Under Assumptions 4-7, letting* $\widetilde{w}$ *be defined as in Equation* (3.8) *we have*

$$\frac{L_n(\widetilde{w})}{\inf_{w\in\mathcal{W}} L_n(w)} \to 1$$

in probability as $n \to \infty$.

Theorem 1 shows that the practitioner may do as well as if they knew the true μ_i, i.e., the weight vector $\widetilde{w}$ is asymptotically optimal in the sense that the average loss with $\widetilde{w}$ is asymptotically equal to that with the infeasible optimal weight vector.

Theorem 1 considers the case where the error variance Ω is known. In practice, however, Ω is often unknown, and we now turn our attention to this important case. We will consider two different ways of estimating Ω. The first estimator of Ω depends on $w = (w_1, \ldots, w_{S_n})$, and the second estimator only depends on w^*, where w^* is the weight for the largest model, i.e., the model with the largest $q_s + r_s$. To render the Mallows-type criterion given in Equation (3.7) computationally feasible, we estimate the unknown Ω based on residuals from model averaging estimation by

$$\widehat{\Omega}(w) = \mathrm{diag}(\hat{\epsilon}_1^2(w), \ldots, \hat{\epsilon}_n^2(w)),$$

where $\hat{\epsilon}_i(w) = y_i - \hat{\mu}_i(w)$.

Replacing Ω with $\widehat{\Omega}(w)$ in $C_n(w)$, we obtain the feasible criterion

$$\widehat{C}_n(w) = n^{-1}\|P(w)Y - Y\|^2 + 2n^{-1}\,\mathrm{tr}[P(w)\widehat{\Omega}(w)].$$

Correspondingly, the new optimal weights are defined as

$$\widehat{w} = \mathrm{argmin}_{w\in\mathcal{W}}\ \widehat{C}_n(w). \tag{3.9}$$

We now show that the weight $\widehat{w}$ is still asymptotically optimal. Let $\rho_{ii}^{(s)}$ be the i^{th} diagonal element of $P_{(s)}$. Then

$$\rho_{ii}^{(s)} = \mathcal{B}(X_{i,(s)})' \left(\sum_{m=1}^{n} \mathcal{B}(X_{m,(s)})\mathcal{B}(X_{m,(s)})' L(Z_{m,(s)}, Z_{i,(s)}, \lambda_{(s)}) \right)^{-1} \mathcal{B}(X_{i,(s)}).$$

The additional assumption required for establishing the asymptotic optimality of $\widehat{w}$ is the following:

Assumption 8. *There exists a constant c such that* $|\rho_{ii}^{(s)}| \leq cn^{-1}|\operatorname{tr}(P_{(s)})|$, $\forall s = 1, \ldots, S_n$, $\forall i = 1, \ldots, n$.

This condition is commonly used to ensure the asymptotic optimality of cross-validation (e.g., [70] and [58]).

Theorem 2. *Under Assumptions 4-8, letting* $\widehat{w}$ *be defined as in Equation* (3.9)*, then we have*

$$\frac{L_n(\widehat{w})}{\inf_{w \in \mathcal{W}} L_n(w)} \to 1 \tag{3.10}$$

in probability as $n \to \infty$.

An alternative strategy for estimating Ω is based on the largest model that includes all predictors (e.g., see the estimation of the variance of homoscedastic error terms in [56]). We estimate the unknown Ω based on residuals from the largest model indexed by s^*, therefore

$$\widehat{\Omega}^* = \operatorname{diag}(\widehat{\epsilon}_{s^*,1}^2, \ldots, \widehat{\epsilon}_{s^*,n}^2),$$

where $(\widehat{\epsilon}_{s^*,1}^2, \ldots, \widehat{\epsilon}_{s^*,n}^2)' = Y - \widehat{\mu}_{(s^*)} = Y - P_{(s^*)}Y$, and $s^* = \operatorname{argmax}_{1 \leq s \leq S_n}(q_s + r_s)$. The new Mallows criterion function becomes

$$\widehat{C}_n^*(w) = n^{-1}\|P(w)Y - Y\|^2 + 2n^{-1}\operatorname{tr}[P(w)\widehat{\Omega}^*].$$

The new optimal weights are defined as

$$\widehat{w}^* = \operatorname{argmin}_{w \in \mathcal{W}} \ \widehat{C}_n^*(w). \tag{3.11}$$

Assumption 9. $\|\mu\|^2 = O(n)$.

Assumption 9 has been used in [57] and [59]. [57] demonstrate that replacing $\widehat{\Omega}^*$ with $\widehat{\Omega}_{(s)}$ estimated from another model s will not alter the result of Theorem 3.

Theorem 3. *Under Assumptions 4-9, letting $\widehat{w}^*$ be defined in Equation* (3.11)*, then*

$$\frac{L_n(\widehat{w}^*)}{\inf_{w\in\mathcal{W}} L_n(w)} \to 1 \tag{3.12}$$

in probability as $n \to \infty$.

Proofs theorems 1-3 can be found in Appendix B.1.

3.3 Monte Carlo Assessment of Finite-Sample Performance

We consider two Monte Carlo simulation experiments designed to assess the finite-sample performance of the proposed methods relative to a representative set of competing methods that include a set of popular model selection techniques along with that based on the largest model (i.e., the model with the largest number of predictors).

3.3.1 Case (I)

Case (I) is a setting in which the candidate models are under-specified (the case covered by our theory). For Case (I), data is simulated from the following DGP where x_1, x_2, and x_4 are continuously distributed as $U[-1,1]$ while x_3 has discrete support generated as a binomially distributed random variate drawn from three independent Bernoulli trials with probability of success $\pi = 1/2$. The DGP is given by $y = x_1+x_2+x_1x_2+x_1^2+x_2^2+x_3+x_1x_3+x_2x_3+x_4+x_4x_1+x_4x_2+\epsilon$, where the variance of ϵ is set so that the signal/noise ratio is such that the expected R^2 for a correctly specified model would be $(0.95, 0.80, 0.50, 0.20)$ which corresponds to $\sigma = (0.25, 0.50, 1.00, 2.00)$ times the standard deviation of the systematic component of the DGP. Results are summarized in Table 3.1 (model average weights are summarized in Table B.1 in Appendix B.2).

For Case (I) we estimate the following six models: $(a)\ y_i = g_1(x_{1i})+\epsilon_i$, $(b)\ y_i = g_2(x_{2i})+\epsilon_i$, $(c)\ y_i = g_4(x_{1i}, x_{2i})+\epsilon_i$, $(d)\ y_i = g_5(x_{1i}, x_{3i})+\epsilon_i$, $(e)\ y_i = g_6(x_{2i}, x_{3i})+\epsilon_i$, $(f)\ y_i = g(x_{1i}, x_{2i}, x_{3i})+\epsilon_i$ (note that the model $y_i = g_3(x_{3i}) + \epsilon_i$ is not included since x_3 is discrete and we need at least one continuous predictor in the kernel-weighted spline specification). For each of these six models, we use cross-validation to select the degree of the tensor spline and smoothing parameter for the

discrete predictor, when present, then estimate the model by nonparametric series methods using kernel-weighted B-splines as outlined above. For the proposed approach we average these six estimators by assigning weights w_1, w_2, w_3, w_4, w_5 and w_6 to these estimators using the Mallows criterion outlined above. In particular, we choose weights to minimize the objective function in Equation (3.11) in which the unknown Ω is estimated based on residuals from the largest model.

We consider five estimators: (1) Mallows model averaging ('MMA'), (2) AIC model selection ('AIC'), (3) BIC model selection ('BIC'), (4) Mallows' C_p model selection, and (5) the largest model ('L'). To evaluate the estimators, we compute the risk (expected squared error). We do this by computing means across $1,000$ simulation draws.

The AIC and BIC criterion are given by $\log(\hat{\sigma}_s^2)+2n^{-1}\operatorname{trace}(P_{(s)})$ and $\log(\hat{\sigma}_s^2)+n^{-1}\operatorname{trace}(P_{(s)})\log(n)$, respectively. The C_p criterion is given by $\hat{\sigma}_s^2(n+2\operatorname{trace}(P_{(s)}))$. Here $s=1,2,...,6$, $P_{(s)}$ is defined via $\widehat{\mu}_{(s)}=P_{(s)}Y$, $\hat{\sigma}_s^2=n^{-1}\sum_{i=1}^n \hat{\epsilon}_{i,s}^2$, and the $\hat{\epsilon}_{i,s}$ are the residuals from the sth model.

Table 3.1: Relative MSE, Case (I); Numbers > 1 Indicate Inferior MSE Performance Relative to the Proposed Model Averaging Approach.

n	σ	MMA	AIC	BIC	C_p	L
50	0.25	1.00	1.20	1.37	1.20	1.25
	0.50	1.00	1.22	1.35	1.22	1.23
	1.00	1.00	1.27	1.31	1.34	1.22
	2.00	1.00	1.38	1.08	1.48	1.27
100	0.25	1.00	1.14	1.30	1.13	1.14
	0.50	1.00	1.17	1.44	1.17	1.18
	1.00	1.00	1.23	1.38	1.28	1.20
	2.00	1.00	1.34	1.22	1.46	1.28
200	0.25	1.00	1.04	1.11	1.04	1.04
	0.50	1.00	1.07	1.33	1.07	1.07
	1.00	1.00	1.17	1.53	1.17	1.16
	2.00	1.00	1.28	1.32	1.39	1.23
400	0.25	1.00	1.01	1.02	1.01	1.01
	0.50	1.00	1.03	1.09	1.03	1.03
	1.00	1.00	1.08	1.54	1.08	1.08
	2.00	1.00	1.22	1.46	1.25	1.21

3.3.2 Case (II)

Case (II) is a setting in which the candidate models contain the true model which also coincides with the largest model, a case not covered by our theory but one that may be of interest to the reader. The DGP is given by $y = x_1 + x_2 + x_1x_2 + x_1^2 + x_2^2 + x_3 + x_1x_3 + x_2x_3 + \epsilon$ which is identical to Case (I) except that x_4 no longer appears, and the residual variance is set per Case (I). We use the same set of models as per Case (I), and for each of the six models, we use the delete-one cross-validation method to select the spline degree and kernel bandwidth, when present, then we average over these estimates and select weights to minimize the Mallows objective function as per Case (I). Results are summarized in Table 3.2 (model average weights are summarized in Table B.2 in Appendix B.2).

Table 3.2: Relative MSE, Case (II); Numbers > 1 Indicate Inferior MSE Performance Relative to the Proposed Model Averaging Approach.

n	σ	MMA	AIC	BIC	C_p	L
50	0.25	1.00	1.29	1.41	1.28	1.32
	0.50	1.00	1.26	1.35	1.28	1.26
	1.00	1.00	1.30	1.34	1.38	1.24
	2.00	1.00	1.41	1.11	1.53	1.28
100	0.25	1.00	1.23	1.32	1.23	1.24
	0.50	1.00	1.24	1.56	1.24	1.24
	1.00	1.00	1.29	1.45	1.35	1.26
	2.00	1.00	1.36	1.26	1.51	1.30
200	0.25	1.00	1.10	1.15	1.10	1.11
	0.50	1.00	1.16	1.40	1.16	1.16
	1.00	1.00	1.24	1.81	1.25	1.23
	2.00	1.00	1.35	1.41	1.48	1.30
400	0.25	1.00	1.03	1.08	1.03	1.03
	0.50	1.00	1.08	1.16	1.08	1.08
	1.00	1.00	1.16	1.95	1.16	1.16
	2.00	1.00	1.28	1.61	1.32	1.26

3.3.3 Discussion

From Table 3.1 we observe that our proposed model averaging approach has the smallest estimation MSE for all cases considered. In general, the gain of our model averaging method is more substantial for smaller sample sizes with smaller signal/noise ratios.

Furthermore, Table 3.2 shows that, even when the true model is included in the set of candidate models (an unlikely event not covered by our theory), the use of model averaging can outperform model selection in small sample settings as Case (II) reveals.

3.4 Empirical Illustration

We consider panel data for two different groups of countries, the OECD and the "rest of the world" consisting of the lesser developed countries. We treat OECD status as categorical and use continuous predictors human capital and initial GDP [71]. We shuffle the data into two independent samples of size $n_1 = 600$ and $n_2 = 16$, fit the models on the n_1 training observations then evaluate the predicted square error (PSE) on the independent evaluation observations. We repeat this exercise 1,000 times then compute the average PSE for the models selected by the AIC, BIC, and C_p criteria, as well as the estimate from the largest model, relative to that for the Mallows procedure proposed herein, with numbers > 1 indicating that the proposed approach provides predictions that are more faithful to the underlying DGP than its model selection-based peers. The relative PSE for the AIC, BIC, and C_p criterion are 1.04, 1.06, and 1.04, respectively, while that for the largest model L is 1.04. We use the same set of candidate models as per Case (I) and Case (II), and the mean model average weights for the six models are $\bar{w}_1 = 0.001378478$, $\bar{w}_2 = 0.05537101$, $\bar{w}_3 = 0.2105344$, $\bar{w}_4 = 0.2822475$, $\bar{w}_5 = 0.07193286$, and $\bar{w}_6 = 0.3785357$.

3.5 Summary

We propose a novel model averaging approach where the candidate models are based on a recently proposed kernel-weighted spline regression method that admits both continuous and categorical predictors [62]. We provide theoretical underpinnings for optimal model average weights in this setting, assess finite-sample performance, and present an empirical illustration. The proposed

model average approach is capable of outperforming a range of popular model selection strategies, and may therefore be of interest to practitioners who wish to confront model uncertainty in applied settings. An R package that implements the proposed approach is available for practitioners [65].

4. MULTIVARIATE DENSITY FORECAST COMBINATION

4.1 Introduction

Density forecasts have attracted more and more attention among academic researchers as well as professional forecasters. Its increasing popularity is attributed to the capability of depicting the uncertainty around the point forecasts. Compared to univariate density forecasts, the multivariate density forecasts allow us to study the cross-variable interactions, such as time-varying conditional correlations between variables of interest. While the majority of the literature in multivariate density forecasts focuses on evaluation and testing ([72],[73]), less attention is paid to the attempts at improving the forecast performance. Forecast combination has been demonstrated as a valid approach to offer people a more accurate description of the true underlying uncertainty. The density forecast combination is pioneered by [74]. [75] establishes statistical asymptotic optimality for the optimal estimators of the combination weight. When it turns to the multivariate density forecast combination, little work is done so far. [76] consider the combination of multivariate density forecasts using the predictive likelihood as the weighting scheme. To the best of my knowledge, there is no existing result concerning the asymptotic property of the estimated combination weight in the context of multidimensional density forecasts.

This article aims to fill this blank space in the literature. I develop a density forecast combination scheme for the multivariate response. I prove the consistency of the estimated combination weight, and use simulations to demonstrate the validity of the econometric theory. Specifically, I decompose the multivariate distribution of the outcome variable into the product of the marginal distribution and conditional distributions, which are all univariate distributions. Then I follow [75] to construct the objective function with respect to the combination weight for each univariate distribution. Finally, in the spirit of [77], I define the objective function for multivariate density forecast combination to be a transformation of the univariate objective functions, for example, over the maximum function. By providing a forecast combination scheme, I anticipate that researchers and

policymakers may benefit from the combined multivariate density forecasts, especially from gaining information on contemporary correlation between variables, which is conveyed particularly from the multivariate structure.

The literature of multivariate density forecasts is still at its infancy. [72] discuss the evaluation and calibration of the multivariate density forecasts. They propose to analyze the histograms and correlograms of the marginal and conditional Probability Integral Transformation (PIT), to assess the appropriateness of a particular joint density forecast. From their discussion about the correlograms, we see that the multivariate density forecast indeed carries extra information about the contemporary correlations between the variables of interest. [77] provide formal tests for the multivariate distributions in GARCH models, where the null is joint normal or joint student-t. [73] propose a modified approach to evaluate the multidimensional density forecast, which has better power under the contemporary links between variables of interests. For the density forecast combination, our closest predecessor is [75] who proposes the weighting scheme of univariate density forecast combination leveraging the Probability Integral Transformation (PIT) and Kullback-Leibler Information Criterion (KLIC), and shows the statistical properties of the optimally chosen weights.

The rest of this essay is organized as follows. Section 4.2 sets up the underlying econometric structure, the estimation scheme of density forecasts, and the definition of a multivariate density forecast combination. Section 4.3 and 4.4 illustrates the PIT and KLIC based combination objective function, respectively. Section 4.5 defines the optimal choice of the combination weight and the asymptotic optimality. Section 4.6 presents the Monte-Carlo simulation. Section 4.7 provides concluding remarks and discussions.

4.2 Setup

Throughout I use the uppercase letters to denote the random variables (vectors), and the lowercase counterparts for their realizations. We observe $Z_t = (Y_t', X_t')', t = 1, ..., T + h$, where $Y_t = (Y_{t1}, ..., Y_{tk})'$ is a k-dimensional vector of variables of interest, and $X_t = (X_{t1}, ..., X_{td})'$ is a d-dimensional vector of predictors. I consider the h-step ahead density forecast. Let $\phi^*_{t+h}(y|\mathcal{I}_t)$

be the true conditional joint pdf of Y_{t+h}, given $\mathcal{I}_t$ which is the information set at time t, where $y = (y_1, ..., y_k)'$.

Consider a rolling window estimation scheme with window size R, so that the density forecast for time $t + h$ is based on the truncated information set $\mathcal{I}_{t-R+1}^t$, which contains information from time $t - R + 1$ to t. For a fixed forecast origin f, the time index t runs from $t = f - G - h + 1$ to $t = f - h$, where G is the total number of rolling windows. The estimation procedure is repeated for all forecast origins f from $f = G + h + R - 1$ to $f = T$. Therefore, we obtain $P = T - G - h - R$ out-of-sample density forecasts with the corresponding realizations, in order to assess the performance of the forecast combinations. For ease of notations, below we consider a fixed forecast origin f, and let the time index run from $t = 1, ..., G$.

We have M models to forecast the h-step ahead predictive joint pdf of Y_{t+h}, each of them being $\phi_{t+h}^m(y|\mathcal{I}_{t-R+1}^t), m = 1, ..., M$. The corresponding CDF is denoted as $\Phi_{t+h}^m(y|\mathcal{I}_{t-R+1}^t)$. We combine the M predictive densities using the convex combination, to have the combined predictive pdf, denoted by

$$\phi_{t+h}^C(y|\mathcal{I}_{t-R+1}^t) = \sum_{m=1}^{M} w_m \phi_{t+h}^m(y|\mathcal{I}_{t-R+1}^t), \tag{4.1}$$

where $w \in \mathcal{W} = \{w \in [0,1]^M : \sum_{m=1}^M w_m = 1\}$. The combined predictive CDF is then given by

$$\begin{aligned} \Phi_{t+h}^C(y|\mathcal{I}_{t-R+1}^t) &= \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_k} \sum_{m=1}^{M} w_m \phi_{t+h}^m(y|\mathcal{I}_{t-R+1}^t) \\ &= \sum_{m=1}^{M} w_m \Phi_{t+h}^m(y|\mathcal{I}_{t-R+1}^t). \end{aligned} \tag{4.2}$$

4.3 Probability Integral Transformation and Rosenblatt's Transformation

For the sake of brevity, in this section we do not specify the information set $\mathcal{I}_{t-R+1}^t$ in the predictive density and the corresponding CDF. We use [78]'s transformation to extend the univariate PIT to a multivariate environment. To convey the idea in a clearer way, in this section, we discuss

the bivariate case, i.e. $k = 2$. For a generic bivariate random vector (S_1, S_2) with pdf $\phi(s_1, s_2)$ and CDF $\Phi(s_1, s_2)$, denote the marginal pdf and CDF of the first argument as $\phi_1(s_1)$ and $\Phi_1(s_1)$. Denote the conditional pdf of S_2 given $S_1 = s_1$ to be $\phi_{2|1}(s_2|s_1)$, and the corresponding conditional CDF to be $\Phi_{2|1}(s_2|s_1)$. Then the marginal pdf and CDF of the combined predictive density are

$$\begin{aligned}
\phi^C_{t+h,1}(s_1) &= \int_{-\infty}^{\infty} \phi^C_{t+h}(s_1, v)dv = \int_{-\infty}^{\infty} \sum_{m=1}^{M} w_m \phi^m_{t+h}(s_1, v)dv \\
&= \sum_{m=1}^{M} w_m \int_{-\infty}^{\infty} \phi^m_{t+h}(s_1, v)dv = \sum_{m=1}^{M} w_m \phi^m_{t+h,1}(s_1). \qquad (4.3) \\
\Phi^C_{t+h,1}(s_1) &= \int_{-\infty}^{s_1} \phi^C_{t+h,1}(v)dv = \sum_{m=1}^{M} w_m \Phi^m_{t+h,1}(s_1). \qquad (4.4)
\end{aligned}$$

And the conditional pdf and CDF could be expressed as

$$\phi^C_{t+h,2|1}(s_2|s_1) = \frac{\phi^C_{t+h}(s_1, s_2)}{\phi^C_{t+h,1}(s_1)} = \frac{\sum_{m=1}^{M} w_m \phi^m_{t+h}(s_1, s_2)dv}{\sum_{m=1}^{M} w_m \phi^m_{t+h,1}(s_1)}. \qquad (4.5)$$

$$\begin{aligned}
\Phi^C_{t+h,2|1}(s_2|s_1) &= \int_{-\infty}^{s_2} \phi^C_{t+h,2|1}(v|s_1)dv = \int_{-\infty}^{s_2} \frac{\phi^C_{t+h}(s_1, v)}{\phi^C_{t+h,1}(s_1)}dv \\
&= \frac{\int_{-\infty}^{s_2} \phi^C_{t+h}(s_1, v)dv}{\phi^C_{t+h,1}(s_1)} \\
&= \frac{\int_{-\infty}^{s_2} \sum_{m=1}^{M} w_m \phi^m_{t+h}(s_1, v)dv}{\sum_{m=1}^{M} w_m \phi^m_{t+h,1}(s_1)} \\
&= \frac{\sum_{m=1}^{M} w_m \int_{-\infty}^{s_2} \phi^m_{t+h}(s_1, v)dv}{\sum_{m=1}^{M} w_m \phi^m_{t+h,1}(s_1)}. \qquad (4.6)
\end{aligned}$$

Define the Probability Integral Transformation (PIT) of the marginal and conditional distribution of the predictive density:

$$U_{t,1} = \Phi^C_{t+h,1}(Y_{t+h,1}), \;\; U_{t,2} = \Phi^C_{t+h,2|1}(Y_{t+h,2}|Y_{t+h,1}). \qquad (4.7)$$

If $\phi^C_{t+h}(y)$ coincides with the true conditional density $\phi^*_{t+h}(y|\mathcal{I}_t)$, then $U_{t,1}, U_{t,2}$ are independent $U[0,1]$ random variables (see [77]). For a fixed h, Let $\widehat{U}_{t,1} = \Phi^C_{t+h,1}(y_{t+h,1}), \widehat{U}_{t,2} = \Phi^C_{t+h,2|1}(y_{t+h,2}|y_{t+h,1})$, that is, to plug in the realized values of y_{t+h}. Define

$$V_{t,i}(r,w) = \mathbb{1}(\widehat{U}_{ti} \leq r) - r, \;\; i = 1,2 \tag{4.8}$$

for a given quantile $r \in [0,1]$, where $\mathbb{1}(\cdot)$ is the indicator function. Let $J_{0,i} = Pr(U_{ti} \leq r) - r$, which is supposed to be 0 under the correct specification of $\phi^C_{t+h}(\cdot)$. Define the sample counterpart of $J_{0,i}$ to be

$$J_{G,i}(r,w) = G^{-1}\sum_{t=1}^{G} V_{ti}(r,w), \;\; i = 1,2, \tag{4.9}$$

which measures the distance between the empirical CDF of the PIT to the CDF of the uniform distribution, at a given quantile r. The desired optimal combined density forecast should make this discrepancy as small as possible. Next we introduce three commonly used statistics to represent this distance uniformly over r. Let $\rho \subset [0,1]$ denote a finite union of neither empty nor singleton, closed intervals, which is selected by the user.

In the spirit of Kolmogorov-Smirnov statistic, we construct three objective functions that allow us to select the optimal weight. Similar statistics also appear in [77], where they are used as test statistics for the test of the multivariate density function in GARCH models. The proposed Kolmogorov-Smirnov type objective functions consist of three variants:

$$\begin{aligned} K^{(1)}_G(w) &= \max\{\Psi_{G,1}(w), \Psi_{G,2}(w)\}, \\ K^{(2)}_G(w) &= \Psi_{G,1}(w) + \Psi_{G,2}(w), \\ K^{(3)}_G(w) &= \sup_{r\in\rho}(|J_{G,1}(r,w) + J_{G,2}(r,w)|), \end{aligned} \tag{4.10}$$

where $\Psi_{G,i}(w) = \sup_{r_i\in\rho}|J_{G,i}(r_i,w)|$. Similarly, the Cramer-von Mises type objective functions

are

$$
\begin{aligned}
C_G^{(1)}(w) &= \max\{\Theta_{G,1}(w), \Theta_{G,2}(w)\}, \\
C_G^{(2)}(w) &= \Theta_{G,1}(w) + \Theta_{G,2}(w), \\
C_G^{(3)}(w) &= \int_\rho (J_{G,1}(r,w) + J_{G,2}(r,w))^2 dr.
\end{aligned} \tag{4.11}
$$

where $\Theta_{G,i}(w) = \int_\rho J_{G,i}^2(r,w)dr$. Finally, the Anderson-Darling type objective functions are

$$
\begin{aligned}
A_G^{(1)}(w) &= \max\{\Lambda_{G,1}(w), \Lambda_{G,2}(w)\}, \\
A_G^{(2)}(w) &= \Lambda_{G,1}(w) + \Lambda_{G,2}(w), \\
A_G^{(3)}(w) &= \int_\rho \left[\frac{(J_{G,1}(r,w) + J_{G,2}(r,w))^2}{r(1-r)} \right] dr.
\end{aligned} \tag{4.12}
$$

where $\Lambda_{G,i}(w) = \int_\rho \frac{J_{G,i}^2(r,w)}{r(1-r)} dr$.

4.4 Kullback-Leibler Information Criterion

Another way to measure the discrepancy between the combined predictive density and the true one is to leverage the Kullback-Leibler Information Criterion (KLIC). Let ϱ_1 denote a finite union of closed, non-empty, non-singleton intervals on the support of true CDF $\Phi_{t+h,1}^*$, and ϱ_2 for $\Phi_{t+h,2|1}^*$ similarly. The KLIC between the marginal CDF is

$$
\begin{aligned}
\mathrm{KLIC}(\Phi_{t+h,1}^*(\cdot), \Phi_{t+h,1}^C(\cdot)) = E_{\phi^*}[\log \phi_{t+h,1}^*(y_{t+h,1}) \mathbb{1}(y_{t+h,1} \in \varrho_1)] \\
- E_{\phi^*}[\log \phi_{t+h,1}^C(y_{t+h,1}) \mathbb{1}(y_{t+h,1} \in \varrho_1)],
\end{aligned} \tag{4.13}
$$

and the KILC between the conditional CDF is

$$
\begin{aligned}
\mathrm{KLIC}(\Phi^*(\cdot|y_{t+h,1}), \Phi_{t+h,2|1}^C(\cdot|y_{t+h,1})) = E_{\phi^*}[\log \phi^*(y_{t+h,2}|y_{t+h,1}) \mathbb{1}(y_{t+h,2} \in \varrho_2)] \\
- E_{\phi^*}[\log \phi_{t+h,1}^C(y_{t+h,2}|y_{t+h,1}) \mathbb{1}(y_{t+h,2} \in \varrho_2)].
\end{aligned} \tag{4.14}
$$

Following [74], we construct the sample counterpart of the KLIC (leaving out the term that is irrelevant to w)

$$\begin{aligned}\mathrm{KLIC}_{G,1}(w) &= \frac{1}{G}\sum_{t=1}^{G} -\log \phi^{C}_{t+h,1}(y_{t+h,1})\mathbb{1}(y_{t+h,1} \in \varrho_1),\\ \mathrm{KLIC}_{G,2}(w) &= \frac{1}{G}\sum_{t=1}^{G} -\log \phi^{C}_{t+h,2|1}(y_{t+h,2}|y_{t+h,1})\mathbb{1}(y_{t+h,2} \in \varrho_2).\end{aligned} \tag{4.15}$$

Now we are able to define the KLIC-based objective function

$$\begin{aligned}H^{(1)}_{G}(w) &= \max\{\mathrm{KLIC}_{G,1}(w), \mathrm{KLIC}_{G,2}(w)\},\\ H^{(2)}_{G}(w) &= \mathrm{KLIC}_{G,1}(w) + \mathrm{KLIC}_{G,2}(w).\end{aligned} \tag{4.16}$$

On the other hand, it is natural to directly use the multivariate density forecast to construct the KLIC. Let ϱ_3 be a finite union of closed, non-empty, non-singleton intervals on the support of true CDF Φ^{*}_{t+h}. Define

$$H^{(3)}_{G}(w) = \mathrm{KLIC}_{G,3}(w) = \frac{1}{G}\sum_{t=1}^{G} -\log \phi^{C}_{t+h}(y_{t+h,1}, y_{t+h,2})\mathbb{1}(y_{t+h} \in \varrho_3) \tag{4.17}$$

4.5 Optimal Weight Estimation

In this section we define our estimator for the optimal weight in the multivariate density forecast combination, and establish its asymptotic optimality. The proofs are in Appendix C.1.

We select the combination weights by minimizing the distance between the combined density and the true density.

$$\widehat{w} = \underset{w\in\mathcal{W}}{\operatorname{argmin}}\, F_G(w), \tag{4.18}$$

where $F_G(w)$ is either $K^{(l)}_{G}(w), C^{(l)}_{G}(w), A^{(l)}_{G}(w)$, or $H^{(l)}_{G}(w)$ with $l = 1, 2, 3$.

For each type of the PIT-based objective functions $K_G^{(l)}(w).C_G^{(l)}(w)$ and $A_G^{(l)}(w)$ with $l = 1, 2, 3$, I define their population counterparts to be $K_0^{(l)}(w), C_0^{(l)}(w)$ and $A_0^{(l)}(w)$, respectively, where I replace $J_{G,i}(r, w)$ in the objective functions with $J_{0,i}(r, w)$.

I make parallel assumptions in [75]. Specifically, the continuity condition (Assumption 3 of [75]) is modified with respect to the marginal and conditional CDF's, and the identification condition (Assumption 5 of [75]) is in with respect to $K_0^{(l)}(w), C_0^{(l)}(w)$ and $A_0^{(l)}(w)$.

Assumption 10. *1. (Dependence) $\{Z_t\}\}$ is ϕ-mixing of size $-k/(2k-1), k \geq 1$ or α-mixing of size $-k/(k-1), k > 1$.*

2. (Contunuity) The combined marginal and conditional CDF are both continuous, i.e.

$$\Pr\left(\Phi^C_{t+h,1}(y_{t+h,1}) = r\right) = 0, \;\; \Pr\left(\Phi^C_{t+h,2|1}(y_{t+h,2}|y_{t+h,1}) = r\right) = 0, \tag{4.19}$$

for all $(w, r) \in \mathcal{W} \times \rho$ and for all t.

3. (Estimation Scheme) $R < \infty$ as $G, T \to \infty$, $1 \leq h < \infty$ and fixed. The number of models M is finite.

4. (Identification) There exists a unique $w^* \in \mathcal{W}$ such that w^* minimizes $K_0^{(l)}(w), C_0^{(l)}(w)$, or $A_0^{(l)}(w)$, $l = 1, 2, 3$, depending on the user-chosen objective function.

5. (Anderson-Darling Assumption) There exists $0 < \delta < 0.5$ such that for $i = 1, 2$,

$$\sup_{w\in\mathcal{W}} \left| \int_0^\delta \frac{J^2_{G,i}(r, w) - J^2_{0,i}(r, w)}{r(1-r)} dr \right| \xrightarrow{a.s.} 0, \;\; \sup_{w\in\mathcal{W}} \left| \int_{1-\delta}^1 \frac{J^2_{G,i}(r, w) - J^2_{0,i}(r, w)}{r(1-r)} dr \right| \xrightarrow{a.s.} 0. \tag{4.20}$$

In addition, since the PIT-based estimation scheme involves conditional distributions, we require the conditional pdf to be well defined.

Assumption 11. $\Pr\left[\phi^C_{t+h,1}(y_{t+h,1}) > 0\right] = 1$ *for all* $w \in \mathcal{W}$.

We have the following result of the PIT-based estimators of the optimal weight.

Theorem 4. *Under Assumptions 10 and 11, we have* $\widehat{w} \xrightarrow{a.s.} w^*$*, where* w^* *is the unique minimizer of* $K_0^{(l)}, C_0^{(l)}$*, or* $A_0^{(l)}, l = 1, 2, 3$*.*

For the KLIC-based objective function, define

$$\begin{aligned}
\mathrm{KLIC}_{0,1}(w) &= \frac{1}{G}\sum_{t=1}^{G} -E_{\phi^*}[\log \phi^C_{t+h,1}(y_{t+h,1})\mathbb{1}(y_{t+h,1} \in \rho_1)], \\
\mathrm{KLIC}_{0,2}(w) &= \frac{1}{G}\sum_{t=1}^{G} -E_{\phi^*}[\log \phi^C_{t+h,2|1}(y_{t+h,2}|y_{t+h,1})\mathbb{1}(y_{t+h,2} \in \rho_2)], \\
\mathrm{KLIC}_{0,3}(w) &= \frac{1}{G}\sum_{t=1}^{G} -E_{\phi^*}[\log \phi^C_{t+h}(y_{t+h,1}, y_{t+h,2})\mathbb{1}(y_{t+h} \in \varrho_3)].
\end{aligned} \tag{4.21}$$

And for $l = 1, 2, 3$, let $H_0^{(l)}(w)$ be the population counterparts of $H_G^{(l)}$ with $\mathrm{KLIC}_{G,i}(w)$ replaced by $\mathrm{KLIC}_{0,i}(w)$.

I make parallel assumptions as in [75]. In particular, I require that $\phi^*_{t+h,1}$ and $\phi^*_{t+h,2|1}$ satisfying Assumption 7 in [75]. In addition, $\Phi^C_{t+h,1}$ and $\Phi^C_{t+h,2|1}$ should satisfy Assumption 8-11 in [75]. Finally, the identification condition (Assumption 12 in [75]) is preserved.

Assumption 12. *1. (Existence)* $E_{\phi^*}[\log \phi^*_{t+h,1}(y_{t+h,1})\mathbb{1}(y_{t+h,1} \in \varrho_1)]$, $E_{\phi^*}[\log \phi^*_{t+h,2|1}(y_{t+h,2}|y_{t+h,1})\mathbb{1}(y_{t+h,2} \in \varrho_2)]$*, and* $E_{\phi^*}[\log \phi^*_{t+h}(y_{t+h})\mathbb{1}(y_{t+h} \in \varrho_3)]$ *exist for all* t*.*

2. *(Continuity)* $\log \phi^C_{t+h,1}(y_{t+h,1})$*,* $\log \phi^C_{t+h,2|1}(y_{t+h,2}|y_{t+h,1})$*, and* $\log \phi^*_{t+h}(y_{t+h})$ *are continuous over* ϱ_1*,* ϱ_2*, and* ϱ_3*, respectively.*

3. *(Dominance) For all* $w \in \mathcal{W}$ *and all* t*, we have* $|\log \phi^C_{t+h,1}(y_{t+h,1})| \leq b_1(y_{t+h,1})$*,* $|\log \phi^C_{t+h,2|1}(y_{t+h,2}|y_{t+h,1})| \leq b_2(y_{t+h,2})$*, and* $|\log \phi^C_{t+h}(y_{t+h})| \leq b_3(y_{t+h})$*, over* ϱ_1, ϱ_2*, and* ϱ_3*, respectively, where* $b_j, j = 1, 2, 3$ *are integrable with respective to its argument for all* t*.*

4. *(Moment Condition)* $\mathbb{E}|\log \phi^C_{t+h,1}(y_{t+h,1})|^{k+\tau} < \Delta < \infty$ *for some* $\tau > 0$ *for all* t *and for all* $w \in \mathcal{W}$*. So are* $\phi^C_{t+h,2|1}(y_{t+h,2}|y_{t+h,1})$ *and* $\phi^C_{t+h}(y_{t+h})$*.*

5. *(Identification) There exists a unique $w^* \in \mathcal{W}$ minimizing $H_0^{(j)}(w), j = 1, 2, 3$, depending on the user-chosen objective function.*

Then we have the following result for the KLIC-based estimator of the optimal weight.

Theorem 5. *Under Assumptions 11 and 12, we have $\widehat{w} \xrightarrow{a.s.} w^*$, where w^* is the unique minimizer of $H_0^{(j)}, j = 1, 2, 3$.*

4.6 Monte-Carlo Simulation

I conduct simulation experiments to verify the validity of our estimators and the asymptotic theory. I consider bivariate density forecast ($Y = (Y_1, Y_2)$) and one step ahead prediction ($h = 1$). All simulations are repeated 2000 times. Sample sizes are $G = \{80, 500, 2000\}$. Without loss of generality, we use the true parameters in the data generating process to estimate the weights. For the true data generating process, I use VAR(1) models with bivariate normal innovations. I combine the true candidate models, so that according to the consistency result, the estimated weight $\widehat{w}$ should converge to the true w^*. The (true) candidate models are given by

$$Z_{t+1} = c^{(j)} + A^{(j)} Z_t + \epsilon_{t+1}^{(j)}, \quad j = 1, 2, 3, \tag{4.22}$$

where $Z_t = [Z_{t1}, Z_{t2}]'$, and $\epsilon_t^{(j)}$ is iid bivariate normally distributed, i.e. $\epsilon_t^{(j)} \sim \mathcal{N}(0, \Sigma^{(j)})$, and $j \in \{1, 2, 3\}$ stands for model M1,M2 and M3, respectively. The true model is a weighted mixture of M1, M2 and M3, with weight $w = (w_1, w_2, w_3)$. Then the true data generating process is

$$Z_{t+1} = c + AZ_t + \epsilon_{t+1}, \tag{4.23}$$

where $c = w_1 c^{(1)} + w_2 c^{(2)} + w_3 c^{(3)}$, $A = w_1 A^{(1)} + w_2 A^{(2)} + w_3 A^{(3)}$, and $\epsilon_t \sim \mathcal{N}(0, \Sigma)$, with $\Sigma = w_1^2 \Sigma^{(1)} + w_2^2 \Sigma^{(2)} + w_3^2 \Sigma^{(3)}$.

M3 is an irrelevant model included in the estimation to see how our estimation procedure can eliminate the redundant model. The true weight is set to be $w^* = (0.4, 0.6, 0)$. In addition, I let the position parameters ($c^{(j)}$ and $A^{(j)}$) of the candidate models are the same, so that the mixture

density is a unimodal one. The parameter specification is as followed:

$$c^{(1)} = c^{(2)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \; A^{(1)} = A^{(2)} = \begin{bmatrix} 0.5 & 0.3 \\ 0.6 & 0.2 \end{bmatrix}, \; \Sigma^{(1)} = \begin{bmatrix} 1 & 2 \\ 2 & 9 \end{bmatrix}, \; \Sigma^{(2)} = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix}.$$

The results are shown in Figure 4.1-4.4 and Table 4.1-4.4. We see that all the proposed estimators of the optimal weight $\widehat{w}$ show convergence to the true weight w^*, as the sample size grows. Within each type of the PIT-based estimators (KS, CvM, and AD), the second variation (KS2, CvM2, and AD2) outperforms the other two variations. This advantage appears almost uniformly over different sample sizes, different statistics (bias, variance, MSE), and different components of w (w_1, w_2 and w_3). The first variation (KS1, CvM1, and AD1) and the third variation (KS3, CvM3, and AD3) make little differences. In a very tiny scale, the first variation is better than the third one in CvM and AD type estimators, and the third one beats the first one in KS type estimators. We therefore compare the second variation across the PIT-based estimators. We see that AD has the best performance, and KS is the worst. This finding corresponds to [75]. For the KLIC type estimator, the first (KLIC1) and the third variations (KLIC3) dominate the second one (KLIC2). Computationally, the KLIC type estimation is much faster than the PIT-based estimations, and the convergence rate (MSE) seems also better than its PIT-based competitors. Moreover, the KLIC estimator performs extremely well in eliminating the irrelevant model M3, even in small sample size.

Figure 4.1: KS Type Estimator of w. True $w^* = (0.4, 0.6, 0)$

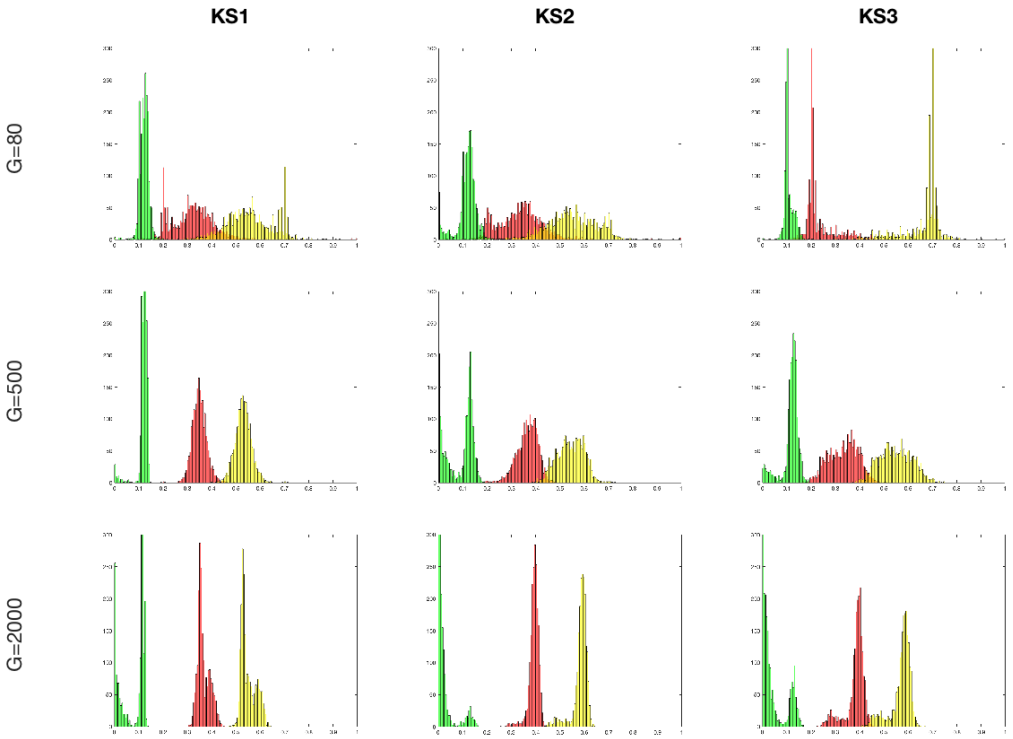


Table 4.1: KS Type Estimators of w

Sample size	Statistic	KS1			KS2			KS3		
		w_1	w_2	w_3	w_1	w_2	w_3	w_1	w_2	w_3
	Bias	-0.0776	-0.0440	0.1217	-0.0625	-0.0529	0.1154	-0.1734	0.0686	0.1048
G=80	Variance	0.0081	0.0104	0.0004	0.0092	0.0121	0.0017	0.0042	0.0062	0.0002
	MSE	0.0141	0.0124	0.0152	0.0131	0.0149	0.0150	0.0342	0.0109	0.0112
	Bias	-0.0502	-0.0688	0.1190	-0.0341	-0.0551	0.0892	-0.0659	-0.0506	0.1165
G=500	Variance	0.0009	0.0012	0.0005	0.0019	0.0030	0.0030	0.0036	0.0045	0.0012
	MSE	0.0034	0.0059	0.0147	0.0031	0.0060	0.0110	0.0080	0.0071	0.0148
	Bias	-0.0329	-0.0488	0.0817	-0.0065	-0.0153	0.0218	-0.0166	-0.0293	0.0459
G=2000	Variance	0.0006	0.0010	0.0024	0.0004	0.0009	0.0013	0.0015	0.0018	0.0025
	MSE	0.0017	0.0033	0.0090	0.0005	0.0011	0.0018	0.0018	0.0026	0.0046

Figure 4.2: CvM Type Estimator of w. True $w^* = (0.4, 0.6, 0)$

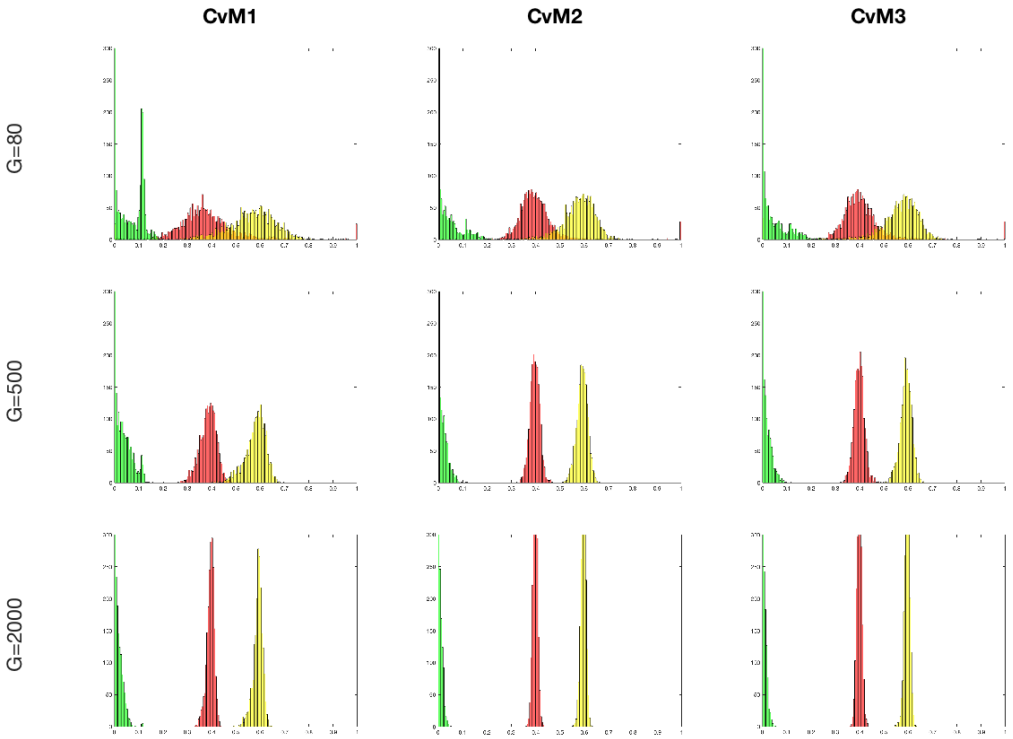


Table 4.2: CvM Type Estimators of w

Sample size	Statistic	CvM1			CvM2			CvM3		
		w_1	w_2	w_3	w_1	w_2	w_3	w_1	w_2	w_3
	Bias	-0.0177	-0.0429	0.0606	0.0055	-0.0314	0.0259	0.0114	-0.0391	0.0277
G=80	Variance	0.0152	0.0144	0.0029	0.0097	0.0100	0.0021	0.0104	0.0110	0.0024
	MSE	0.0155	0.0162	0.0065	0.0097	0.0110	0.0028	0.0105	0.0125	0.0031
	Bias	-0.0138	-0.0186	0.0324	-0.0034	-0.0076	0.0110	-0.0023	-0.0088	0.0111
G=500	Variance	0.0012	0.0018	0.0011	0.0004	0.0005	0.0003	0.0004	0.0006	0.0003
	MSE	0.0014	0.0021	0.0022	0.0004	0.0006	0.0004	0.0005	0.0006	0.0004
	Bias	-0.0045	-0.0093	0.0138	-0.0015	-0.0040	0.0054	-0.0011	-0.0044	0.0055
G=2000	Variance	0.0002	0.0004	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	MSE	0.0003	0.0005	0.0005	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

Figure 4.3: AD Type Estimator of w. True $w^* = (0.4, 0.6, 0)$

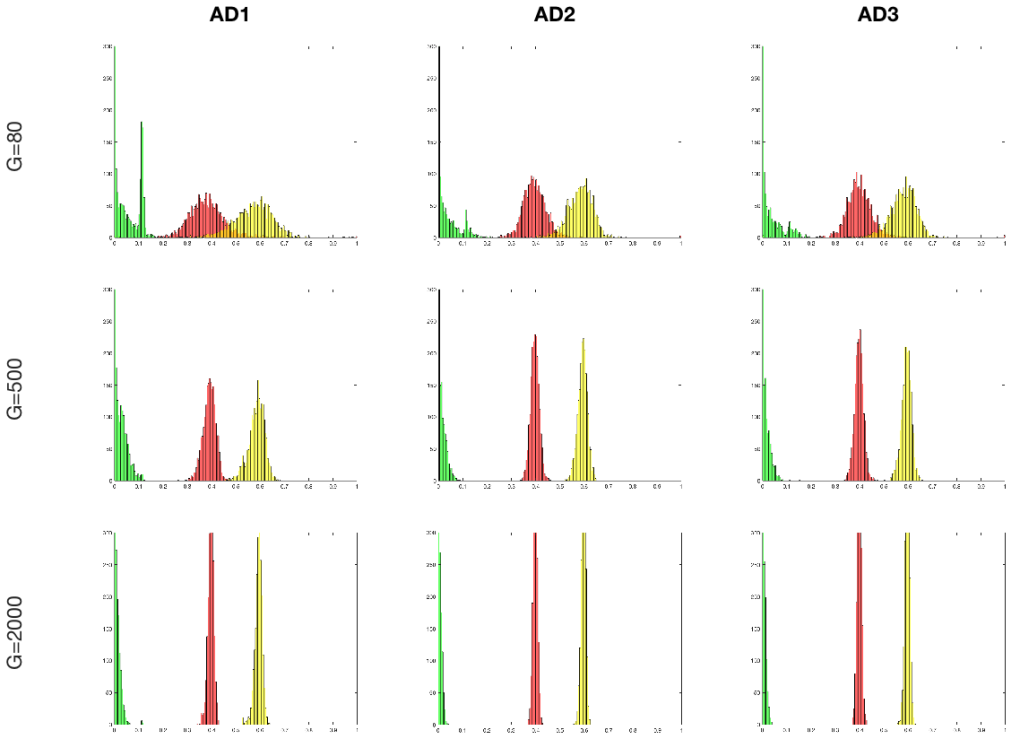


Table 4.3: AD Type Estimators of w

Sample size	Statistic	AD1			AD2			AD3		
		w_1	w_2	w_3	w_1	w_2	w_3	w_1	w_2	w_3
	Bias	-0.0176	-0.0316	0.0492	-0.0035	-0.0176	0.0211	0.0006	-0.0217	0.0210
G=80	Variance	0.0067	0.0073	0.0025	0.0027	0.0032	0.0015	0.0034	0.0037	0.0016
	MSE	0.0070	0.0083	0.0049	0.0027	0.0035	0.0020	0.0034	0.0042	0.0020
	Bias	-0.0093	-0.0142	0.0234	-0.0029	-0.0062	0.0091	-0.0023	-0.0070	0.0093
G=500	Variance	0.0007	0.0011	0.0007	0.0003	0.0004	0.0002	0.0003	0.0004	0.0002
	MSE	0.0008	0.0013	0.0012	0.0003	0.0004	0.0003	0.0003	0.0005	0.0003
	Bias	-0.0030	-0.0068	0.0098	-0.0013	-0.0033	0.0046	-0.0011	-0.0037	0.0047
G=2000	Variance	0.0001	0.0002	0.0002	0.0001	0.0001	0.0000	0.0001	0.0001	0.0000
	MSE	0.0002	0.0003	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

Figure 4.4: KLIC Type Estimator of w. True $w^* = (0.4, 0.6, 0)$

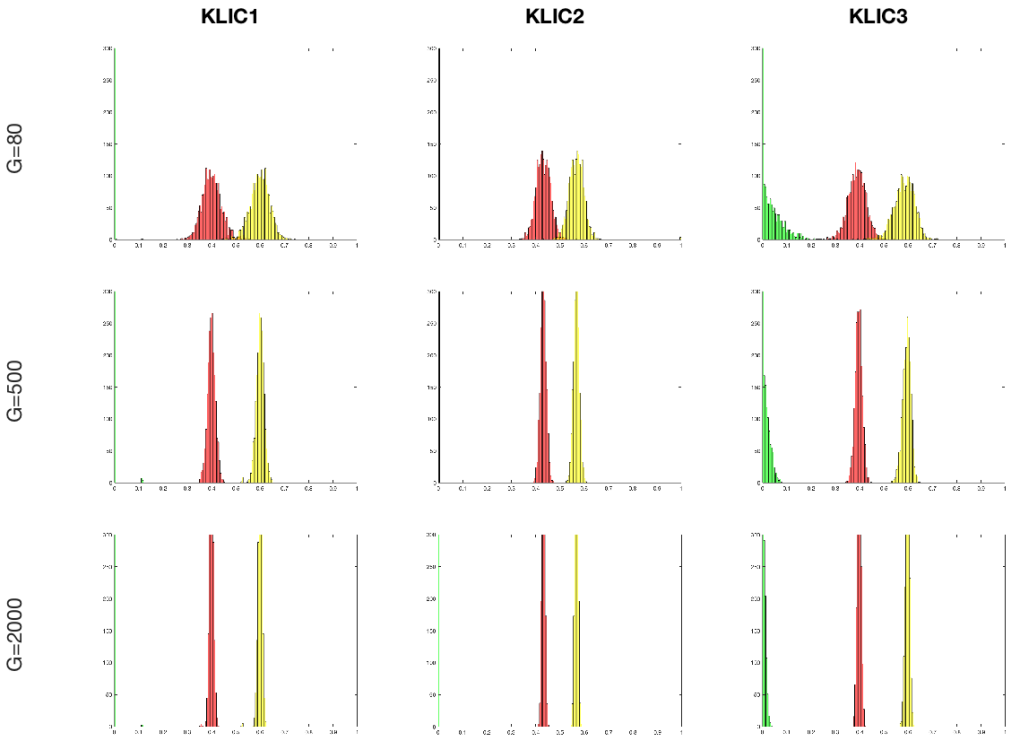


Table 4.4: KLIC Type Estimators of w

Sample size	Statistic	KLIC1			KLIC2			KLIC3		
		w_1	w_2	w_3	w_1	w_2	w_3	w_1	w_2	w_3
	Bias	0.0004	-0.0005	0.0001	0.0310	-0.0310	0.0000	-0.0107	-0.0167	0.0274
G=80	Variance	0.0016	0.0016	0.0000	0.0012	0.0012	0.0000	0.0014	0.0019	0.0017
	MSE	0.0016	0.0016	0.0000	0.0022	0.0022	0.0000	0.1099	0.0321	0.2054
	Bias	-0.0006	0.0000	0.0006	0.0317	-0.0317	0.0000	-0.3526	-0.1051	0.4577
G=500	Variance	0.0003	0.0003	0.0001	0.0001	0.0001	0.0000	0.0026	0.0038	0.0078
	MSE	0.0003	0.0003	0.0001	0.0011	0.0011	0.0000	0.1269	0.0148	0.2173
	Bias	-0.0002	-0.0002	0.0003	0.0322	-0.0322	0.0000	-0.3631	-0.1124	0.4755
G=2000	Variance	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0008	0.0010	0.0023
	MSE	0.0001	0.0001	0.0000	0.0011	0.0011	0.0000	0.1327	0.0136	0.2284

4.7 Conclusion and Discussions

This paper develops a forecast combination method for multidimensional variables, which renders the information of interdependency between multiple quantities of interest. I provide a class of weighting schemes over different forecasting models, prove the asymptotic optimality of the estimated weights, and use simulation to demonstrate the validity of theoretical results.

The current study opens the revenue for further research in the multivariate density forecast combination. Firstly, instead of using the Rosenblatt's transformation to deal with the multivariate density, one could work on the Kendell distribution function $K_W(r) = Pr(H(W) \leq r)$ where $H(w)$ is the joint CDF of $W = (W_1, W_2)'$. This could be viewed as the multivariate probability integral transformation (MPIT) of W. However, in general, K_W is not necessarily uniformly distributed in $[0, 1]$. Actually, by [79], $K_W(r) \geq r$ and $K_W(0^-) = 0$, and K_W is $U[0, 1]$ if and only if W_1, W_2 are independent. In lack of an explicit way to derive the Kendall distribution, one could use simulation to approximate it. Then we can choose the combination weight by minimizing the distance between the empirical MPIT and the true MPIT as the operation with PIT. We may also obtain the Kendall distribution via copulas. As a result of the Sklar's Theorem, the Kendall distribution function $K_W(\cdot)$ depends only on the copula of W. We can estimate (or directly take) the copula function of each forecast model and then derive the corresponding Kendall distribution. Secondly, since it is easier to obtain univariate density forecasts for each variable of interest separately, one might be interested to know what if we just employ the marginal densities into our combination scheme. Intuitively, this leaves out the important message of interactions among variables. The discrepancy between this "naive" combination and our proposed method would exhibit the significance of interdependency. Thirdly, note that there are $k!$ ways to factor the joint density function where k is the dimension of Y_t. Thus, we could conduct different ways of decomposition in our method as a robustness check. Finally, exploring more sophisticated simulation designs would help to verify the efficacy of our approach, and I am also interested in working with a real dataset to combine multivariate density forecasts from different professional forecasters.

5. CONCLUSION

In this dissertation, I develop econometric methods for studying three economic issues: the structural analysis of spectrum auctions, the estimation of conditional expectations, and the multivariate density forecasts.

In the first essay, I provide a structural approach to analyze US spectrum auctions and recover the bidders' values including stand-alone values and complementarity values. I find that the complementarity of a national-wide license is worth 8 billion dollars for an average bidder, which is 59.54% of the sum of final prices of all licenses, indicating strong evidence of complementarity. For the bidders' stand-alone values, I document a significant effect of the license-characteristic, while the effect of the bidder-characteristic is insignificant. The exploration of bidder heterogeneity reveals that large bidders evaluate complementarity higher than small and medium bidders. Methodologically, within the estimation scheme, I establish a framework for estimating the high-dimensional bundle choice problem with individual-level data. I leverage the random projection technique in machine learning to achieve dimension reduction. There are rich applications of this method in other contexts, for example, the demand estimation in industrial organization and marketing.

In the second essay, we study model averaging in the mixed-data environment with nonparametric regression spline models. We provide conditions and proofs of the asymptotic optimality properties of our estimators. The proposed model averaging approach performs better than several appealing model selection methods, and may, therefore, be of interest to applied researchers faced with model uncertainty in various settings.

In the third essay, I propose a multivariate density forecast combination method. My strategy is to exploit the Rosenblatt's transformation and decompose the multivariate density into univariate densities. Then I construct the forecast combination scheme based on the univariate density forecasts [75] and the test statistics in [77]. Simulation results support my theoretical results on the asymptotic optimality of the estimated weights.

REFERENCES

[1] M. Bichler and J. K. Goeree, *Handbook of Spectrum Auction Design*. Cambridge University Press, 2017.

[2] P. Milgrom, "Putting auction theory to work: The simultaneous ascending auction," *Journal of Political Economy*, vol. 108, no. 2, pp. 245–272, 2000.

[3] L. M. Ausubel, P. Cramton, and P. Milgrom, "The clock-proxy auction: A practical combinatorial auction design," *Handbook of Spectrum Auction Design*, pp. 120–140, 2006.

[4] J. Levin and A. Skrzypacz, "Properties of the combinatorial clock auction," *American Economic Review*, vol. 106, no. 9, pp. 2528–51, 2016.

[5] J. Bulow, J. Levin, and P. Milgrom, "Winning play in spectrum auctions," tech. rep., National Bureau of Economic Research, 2009.

[6] P. Milgrom, "Auction market design: Recent innovations," *Annual Review of Economics*, vol. 11, pp. 383–405, 2019.

[7] P. Milgrom, *Putting Auction Theory to Work*. Cambridge University Press, 2004.

[8] J. T. Fox and P. Bajari, "Measuring the efficiency of an fcc spectrum auction," *American Economic Journal: Microeconomics*, vol. 5, no. 1, pp. 100–146, 2013.

[9] J. K. Goeree and Y. Lien, "An equilibrium analysis of the simultaneous ascending auction," *Journal of Economic Theory*, vol. 153, pp. 506–533, 2014.

[10] P. Cramton, "Simultaneous ascending auctions," *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

[11] M. Xiao and Z. Yuan, "License complementarity and package bidding: The US spectrum auctions," *Available at SSRN 3266253*, 2018.

[12] X. Shi, M. Shum, and W. Song, "Estimating semi-parametric panel multinomial choice models using cyclic monotonicity," *Econometrica*, vol. 86, no. 2, pp. 737–761, 2018.

[13] K. X. Chiong and M. Shum, "Random projection estimation of discrete-choice models with large choice sets," *Management Science*, vol. 65, no. 1, pp. 256–271, 2018.

[14] K. Bel, D. Fok, and R. Paap, "Parameter estimation in multivariate logit models with many binary choices," *Econometric Reviews*, vol. 37, no. 5, pp. 534–550, 2018.

[15] K. Hyndman and C. F. Parmeter, "Efficiency or competition? a structural econometric analysis of canada's aws auction and the set-aside provision," *Production and Operations Management*, vol. 24, no. 5, pp. 821–839, 2015.

[16] P. A. Haile and E. Tamer, "Inference with an incomplete model of english auctions," *Journal of Political Economy*, vol. 111, no. 1, pp. 1–51, 2003.

[17] D. McFadden, "Modeling the choice of residential location," *Transportation Research Record*, no. 673, 1978.

[18] J. T. Fox, "Semiparametric estimation of multinomial discrete-choice models using a subset of choices," *The RAND Journal of Economics*, vol. 38, no. 4, pp. 1002–1019, 2007.

[19] D. Nibbering, "A high-dimensional multinomial choice model," tech. rep., Monash University, Department of Econometrics and Business Statistics, 2019.

[20] F. J. Ruiz, S. Athey, and D. M. Blei, "Shopper: A probabilistic model of consumer choice with substitutes and complements," *arXiv preprint arXiv:1711.03560*, 2017.

[21] M. Bruins, J. A. Duffy, M. P. Keane, and A. A. Smith Jr, "Generalized indirect inference for discrete choice models," *Journal of Econometrics*, vol. 205, no. 1, pp. 177–203, 2018.

[22] A. Iaria and A. Wang, "Identification and estimation of demand for bundles," *Available at SSRN 3458543*, 2019.

[23] P. Bajari, D. Nekipelov, S. P. Ryan, and M. Yang, "Demand estimation with machine learning and model combination," tech. rep., National Bureau of Economic Research, 2015.

[24] R. Donnelly, F. R. Ruiz, D. Blei, and S. Athey, "Counterfactual inference for consumer choice across many product categories," *arXiv preprint arXiv:1906.02635*, 2019.

[25] B. J. Gillen, S. Montero, H. R. Moon, and M. Shum, "Blp-2lasso for aggregate discrete choice models with rich covariates," *The Econometrics Journal*, vol. 22, no. 3, pp. 262–281, 2019.

[26] J. H. Kagel, Y. Lien, and P. Milgrom, "Ascending prices and package bidding: A theoretical and experimental analysis," *American Economic Journal: Microeconomics*, vol. 2, no. 3, pp. 160–85, 2010.

[27] C. Brunner, J. K. Goeree, C. A. Holt, and J. O. Ledyard, "An experimental test of flexible combinatorial spectrum auction formats," *American Economic Journal: Microeconomics*, vol. 2, no. 1, pp. 39–57, 2010.

[28] L. M. Ausubel, "An efficient dynamic auction for heterogeneous commodities," *American Economic Review*, vol. 96, no. 3, pp. 602–629, 2006.

[29] S. Brusco and G. Lopomo, "Collusion via signalling in simultaneous ascending bid auctions with heterogeneous objects, with and without complementarities," *The Review of Economic Studies*, vol. 69, no. 2, pp. 407–436, 2002.

[30] R. Engelbrecht-Wiggans and C. M. Kahn, "Low-revenue equilibria in simultaneous ascending-bid auctions," *Management Science*, vol. 51, no. 3, pp. 508–518, 2005.

[31] J. W. Hatfield and P. R. Milgrom, "Matching with contracts," *American Economic Review*, vol. 95, no. 4, pp. 913–935, 2005.

[32] X. Meng and H. Gunay, "Exposure problem in multi-unit auctions," *International Journal of Industrial Organization*, vol. 52, pp. 165–187, 2017.

[33] L. Lamy, "Ascending auctions: some impossibility results and their resolutions with final price discounts," tech. rep., HAL, 2009.

[34] V. Krishna, R. W. Rosenthal, *et al.*, "Simultaneous auctions with synergies," *Games and Economic Behavior*, vol. 17, no. 1, pp. 1–31, 1996.

[35] W. Shin, "Simultaneous auctions for complementary goods," *arXiv preprint arXiv:1312.2641*, 2013.

[36] M. Gentry, T. Komarova, P. Schiraldi, and W. Shin, "On monotone strategy equilibria in simultaneous auctions for complementary goods," *Journal of Mathematical Economics*, vol. 85, pp. 109–128, 2019.

[37] N. Sun and Z. Yang, "An efficient and incentive compatible dynamic auction for multiple complements," *Journal of Political Economy*, vol. 122, no. 2, pp. 422–466, 2014.

[38] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. Springer, 2013.

[39] B. E. Hansen, "Least squares model averaging," *Econometrica*, vol. 75, no. 4, pp. 1175–1189, 2007.

[40] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. Springer Series in Statistics New York, 2001.

[41] G. J. Russell and A. Petersen, "Analysis of cross category dependence in market basket selection," *Journal of Retailing*, vol. 76, no. 3, pp. 367–392, 2000.

[42] B. Dai, S. Ding, G. Wahba, *et al.*, "Multivariate bernoulli distribution," *Bernoulli*, vol. 19, no. 4, pp. 1465–1483, 2013.

[43] C. Varin, N. Reid, and D. Firth, "An overview of composite likelihood methods," *Statistica Sinica*, pp. 5–42, 2011.

[44] C. F. Manski and S. R. Lerman, "The estimation of choice probabilities from choice based samples," *Econometrica*, pp. 1977–1988, 1977.

[45] G. King and L. Zeng, "Logistic regression in rare events data," *Political Analysis*, vol. 9, no. 2, pp. 137–163, 2001.

[46] P. Li, T. J. Hastie, and K. W. Church, "Very sparse random projections," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 287–296, ACM, 2006.

[47] C. F. Manski and D. McFadden, *Structural Analysis of Discrete Data with Econometric Applications*. MIT press Cambridge, MA, 1981.

[48] R. T. Rockafellar, *Convex Analysis*, vol. 28. Princeton University Press, 1970.

[49] A. E. Roth and M. A. O. Sotomayor, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Econometric Society Monographs, Cambridge University Press, 1990.

[50] E. E. Leamer, "Let's take the con out of econometrics," *The American Economic Review*, vol. 73, no. 1, pp. 31–43, 1983.

[51] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*. Cambridge, 2008.

[52] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions in Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[53] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[54] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: A tutorial," *Statistical Science*, vol. 14, p. 382417, 1999.

[55] S. T. Buckland, K. P. Burnhamn, and N. H. Augustin, "Model selection: An integral part of inference," *Biometrics*, vol. 53, p. 603618, 1997.

[56] B. E. Hansen, "Least squares model averaging," *Econometrica*, vol. 75, pp. 1175–1189, 2007.

[57] A. T. Wan, X. Zhang, and G. Zou, "Least squares model averaging by mallows criterion," *Journal of Econometrics*, vol. 156, no. 2, pp. 277–283, 2010.

[58] B. E. Hansen and J. S. Racine, "Jackknife model averaging," *Journal of Econometrics*, vol. 167, no. 1, pp. 38–46, 2012.

[59] X. Zhang and W. Wang, "Optimal model averaging estimation for partially linear models," *Statistica Sinica*, vol. 29, no. 2, pp. 693–718, 2019.

[60] X. Zhang, G. Zou, and R. Carroll, "Model averaging based on Kullback-Leibler distance," *Statistic Sinica*, vol. 25, pp. 1583–1598, 2015.

[61] X. Zhang, D. Yu, G. Zou, and H. Liang, "Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models," *Journal of the American Statistical Association*, vol. 111, no. 516, pp. 1775–1790, 2016.

[62] S. Ma, J. S. Racine, and L. Yang, "Spline regression in the presence of categorical predictors," *Journal of Applied Econometrics*, vol. 30, pp. 703–717, 2015.

[63] B. E. Hansen, "Model averaging, asymptotic risk, and regressor groups," *Quantitative Economics*, vol. 5, no. 3, pp. 495–530, 2014.

[64] Q. Liu, R. Okui, and A. Yoshimura, "Generalized least squares model averaging," *Econometric Reviews*, pp. 1–61, 2016.

[65] J. S. Racine, *ma: Model Averaging*, 2017. R package version 1.0-8.

[66] X. Lu and L. Su, "Jackknife model averaging for quantile regressions," *Journal of Econometrics*, vol. 188, no. 1, pp. 40–58, 2015.

[67] W. K. Newey, "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, vol. 79, no. 1, pp. 147–168, 1997.

[68] T. Ando and K.-C. Li, "A model-averaging approach for high-dimensional regression," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 254–265, 2014.

[69] S. Zhao, X. Zhang, and Y. Gao, "Model averaging with averaging covariance matrix," *Economics Letters*, vol. 145, pp. 214–217, 2016.

[70] D. W. Andrews, "Asymptotic optimality of generalized c l, cross-validation, and generalized cross-validation in regression with heteroskedastic errors," *Journal of Econometrics*, vol. 47, no. 2, pp. 359–377, 1991.

[71] E. Maasoumi, J. S. Racine, and T. Stengos, "Growth and convergence: A profile of distribution dynamics and mobility," *Journal of Econometrics*, vol. 136, pp. 483–508, 2007.

[72] F. X. Diebold, J. Hahn, and A. S. Tay, "Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange," *Review of Economics and Statistics*, vol. 81, no. 4, pp. 661–673, 1999.

[73] S. I. Ko and S. Y. Park, "Multivariate density forecast evaluation: A modified approach," *International Journal of Forecasting*, vol. 29, no. 3, pp. 431–441, 2013.

[74] S. G. Hall and J. Mitchell, "Combining density forecasts," *International Journal of Forecasting*, vol. 23, no. 1, pp. 1–13, 2007.

[75] G. Ganics, "Optimal density forecast combinations," Working Papers 1751, Banco de España, 2017.

[76] H. Gerard and K. Nimark, "Combining multivariate density forecasts using predictive criteria," *Working Papers (Universitat Pompeu Fabra. Departamento de Economía y Empresa)*, no. 1117, p. 1, 2008.

[77] J. Bai and Z. Chen, "Testing multivariate distributions in garch models," *Journal of Econometrics*, vol. 143, no. 1, pp. 19–36, 2008.

[78] M. Rosenblatt, "Remarks on a multivariate transformation," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 470–472, 1952.

[79] R. B. Nelsen, J. J. Quesada-Molina, J. A. Rodriguez-Lallena, and M. Ubeda-Flores, "Kendall distribution functions," *Statistics and Probability Letters*, vol. 65, no. 3, pp. 263–268, 2003.

[80] P. Whittle, "Bounds for the moments of linear and quadratic forms in independent variables," *Theory of Probability & Its Applications*, vol. 5, no. 3, pp. 302–305, 1960.

[81] L. Hogben, *Handbook of Linear Algebra*. CRC Press, 2006.

APPENDIX A

APPENDIX FOR THE FIRST ESSAY

A.1 Equilibrium of the Multiple-Object Clock Auction

In this section, I derive the Bayesian Nash equilibrium (BNE) for the multiple-object clock auction where objects are heterogeneous and complementary. I first consider the two-object case, which is used in Section 2.3 for equilibrium analysis. Then I extent the results to multiple-object case. Related proofs are provided in Appendix A.2.

A.1.1 Two Objects

Lemma 1 (Drop-Out Level Ordering)**.** *At any time, for a global bidder i, her strategy is the drop-out levels (s_i^A, s_i^B) such that $v_i^B \leq s_i^B \leq s_i^A \leq v_i^A + \theta_i$.*

Next, we study several simple subgames, in which the (weakly) dominant strategies for the global bidders are obvious. Similar results are also found in [32].

Lemma 2 (Trivial Subgames)**.** *For bidder i,*

(i) If she has already dropped out on object B at a price $s_i^B < v_i^A$, then she would drop out on A up to her stand-alone value of A, i.e. $s_{i,0}^A = v_i^A$. In other words, she plays the weakly dominant strategy in the English auction.

(ii) If she has already dropped out on object B at a price $s_i^B > v_i^A$, then she would drop out on object A at the same time, i.e. $s_{i,0}^A = s_i^B$.

(iii) If all other bidders have dropped out on object B, that is, bidder i obtains object B (at a price $s_i^B < v_i^A + \theta_i$). Then her drop-out level on object A is $s_{i,1}^A = v_i^A + \theta_i$.

Thus, the only situation left to be considered is where there are more than one bidders staying on both objects. In this case, let $p^t(A) = p^t(B) = p$. Suppose there are n^B other global bidders

staying on object B, and n^A other global bidders staying on object A. For a bidder i, let $s^B_{-i} = s^B_{-i}(n^B) \equiv \max_{j\neq i} s^B_j$ be the maximum of other bidders' drop-out level on B, which is a function of all other bidders' values. Let $G^B_{-i}(\cdot|p, n^B)$ be the CDF of s^B_{-i}, given $s^B_{-i} \geq p$ and the number of other remaining bidder of B, n^B. A bidder's drop-out level on object A depends on whether she wins object B or not, given by Lemma 2. Her belief about other bidders' drop-out level on A also depends on her winning or losing B. Let $s^A_{-i,1}$ and $s^A_{-i,0}$ be drop-out level on A when she wins B and loses B, respectively. Then we have

$$s^A_{-i,1} = s^A_{-i,1}(n^A) \equiv \max_{j\neq i} s^A_{j,0}, \quad s^A_{-i,0} = s^A_{-i,0}(n^A) \equiv \max\left\{\max_{j\neq i,k} s^A_{j,0}, s^A_{k,1}\right\}, \tag{A.1}$$

where k denotes the winner on B, suppose i did not win it. Let $G^A_{-i,1}(\cdot|p, n^A)$ and $G^A_{-i,0}(\cdot|p, n^A)$ be the CDF (given they are no less than p) of $s^A_{-i,1}$ and $s^A_{-i,0}$, respectively.

It will be clear later on that G^B_{-i} does not affect the equilibrium strategy of the global bidders. Let $F^C(s) \equiv Pr(v^A_i + \theta_i < s)$ be the distribution function of $v^A_i + \theta_i$. By the drop-out level analysis in Lemma 2, with simple calculations, we have

$$G^A_{-i,1}(s|p, n^A) = \left(\frac{F^A(s) - F^A(p)}{1 - F^A(p)}\right)^{n^A} \tag{A.2}$$

$$G^A_{-i,0}(s|p, n^A) = \frac{[F^A(s) - F^A(p)]^{n^A-1} \cdot [F^C(s) - F^C(p)]}{[1 - F^A(p)]^{n^A-1} \cdot [1 - F^C(p)]}. \tag{A.3}$$

We have $G^A_{-i,1} - G^A_{-i,0} \geq 0$. One can verify it from the above expression. In addition, it could be directly seen because, if i loses B, then one of his opponent wins it, and this bidder's updated value for object A would be adding θ, so

$$\begin{aligned}\mathbb{P}(s^A_{-i,0} < s) &= \mathbb{P}(\max_{j\neq i,k} v^A_j < s) \cdot \mathbb{P}(v^A_k + \theta < s) \leq \mathbb{P}(\max_{j\neq i,k} v^A_j < s) \cdot \mathbb{P}(v^A_k < s) \\ &= \mathbb{P}(\max_{j\neq i} v^A_j < s) = \mathbb{P}(s^A_{-i,1} < s).\end{aligned} \tag{A.4}$$

Let $\lambda_i \equiv \min\{v^A_i + \theta, \overline{v}^A\}$. If bidder i's drop-out level on object B is s^B_i, then her expected utility when the current price is $p = p^A = p^B$ is given by

$$\Pi_i(s_i^B; p, \mathbf{v}_i, \mathbf{n}, F) = \int_p^{s_i^B} \Bigg[v_i^B - s_{-i}^B + \mathbb{1}(s_{-i}^B < \overline{v}^A) \int_{s_{-i}^B}^{\lambda_i} (v_i^A + \theta_i - s_{-i}^A) d\, G_{-i,1}^A(s_{-i}^A | s_{-i}^B, n^A) \\ + \mathbb{1}(s_{-i}^B > \overline{v}^A)(v_i^A + \theta_i - s_{-i}^B) \Bigg] d\, G_{-i}^B(s_{-i}^B | p, n^B) \\ + \int_{s_i^B}^{\max\{v_i^A, s_i^B\}} \Bigg[\int_{s_{-i}^B}^{v_i^A} (v_i^A - s_{-i}^A) d\, G_{-i,0}^A(s_{-i}^A | s_{-i}^B, n^A) \Bigg] d\, G_{-i}^B(s_{-i}^B | p, n^B), \tag{A.5}$$

where $\mathbf{v}_i = (v_i^A, v_i^B, \theta_i), \mathbf{n} = (n^A, n^B), F = (F^A, F^B, F^\theta)$, and $G_{-i,1}^A$ and $G_{-i,0}^A$ are defined in (A.2) and (A.3), respectively. The first term on the right hand side is the expected utility in the case of winning B, i.e. $s_i^B > s_{-i}^B$, while the second term is when losing B, i.e. $s_i^B < s_{-i}^B$. Then the first order condition results in:

$$s_i^B = v_i^B + \mathbb{1}(s_i^B < \overline{v}^A) \int_{s_i^B}^{\lambda_i} (v_i^A + \theta_i - s_{-i}^A) d\, G_{-i,1}^A(s_{-i}^A | s_i^B, n^A) \\ + \mathbb{1}(s_i^B > \overline{v}^A)(v_i^A + \theta_i - s_i^B) - \mathbb{1}(s_i^B < v_i^A) \int_{s_i^B}^{v_i^A} (v_i^A - s_{-i}^A) d\, G_{-i,0}^A(s_{-i}^A | s_i^B, n^A), \tag{A.6}$$

The equilibrium drop-out level s_i^B solves the above equation. We see that the first order condition does not involve G_{-i}^B. If $s_i^B > \overline{v}^A$, then the (A.6) simplifies to

$$s_i^B = v_i^B + v_i^A + \theta_i - s_i^B. \tag{A.7}$$

Therefore, in this case, $s_i^B = s_i^A = (v_i^A + v_i^B + \theta)/2$. The condition is $(v_i^A + v_i^B + \theta)/2 > \overline{v}^A$. In the following we discuss a bidder with $(v_i^A + v_i^B + \theta)/2 < \overline{v}^A$, so that we have $s_i^B < \overline{v}^A$ and thus $s_i^B \in [v_i^B, \lambda_i]$. The first order condition becomes:

$$s_i^B = v_i^B + \int_{s_i^B}^{\lambda_i} (v_i^A + \theta_i - s_{-i}^A) d\, G_{-i,1}^A(s_{-i}^A | s_i^B, n^A) - \mathbb{1}(s_i^B < v_i^A) \int_{s_i^B}^{v_i^A} (v_i^A - s_{-i}^A) d\, G_{-i,0}^A(s_{-i}^A | s_i^B, n^A), \tag{A.8}$$

Let

$$J(s, n^A) = v_i^B - s + \int_{s}^{\lambda_i} (v_i^A + \theta_i - s_{-i}^A) d\, G_{-i,1}^A(s_{-i}^A | s, n^A) - \mathbb{1}(s < v_i^A) \int_{s}^{v_i^A} (v_i^A - s_{-i}^A) d\, G_{-i,0}^A(s_{-i}^A | s, n^A). \tag{A.9}$$

Hence, the equilibrium strategy s_i^B solves the equation $J(s_i^B, n^A) = 0$. In Appendix A.2 we prove that $J(s, n^A)$ is non-increasing in $s \in [v_i^B, \lambda_i]$. One observes that $J(v_i^B, n^A) > 0$ since $\lambda_i > v_i^A$ and $G_{-i,1}^A > G_{-i,0}^A$. In addition, $J(\lambda_i, n^A) = v_i^B - \lambda_i < 0$. Therefore, there is a unique root for $J(s, n^A) = 0$ in $s_i^B \in [v_i^B, \lambda_i]$. Further, if $J(v_i^A, n^A) > 0$, then the unique root is $s_i^B \in [v_i^A, v_i^A + \theta_i]$; if $J(v_i^A, n^A) < 0$, then the unique root is $s_i^B \in [v_i^B, v_i^A]$. One immediately sees that if $v_i^A + \theta_i < \overline{v}^A$, then

$$J(v_i^A, n^A) = J_1(v_i^A, n^A) = v_i^B - v_i^A + \int_{v_i^A}^{v_i^A + \theta_i} G_{-i,1}^A(x | v_i^A, n^A) dx, \tag{A.10}$$

and if $v_i^A + \theta_i > \overline{v}^A$, then

$$J(v_i^A, n^A) = J_2(v_i^A, n^A) = v_i^B + \theta_i - \overline{v}^A + \int_{v_i^A}^{\overline{v}^A} G_{-i,1}^A(x | v_i^A, n^A) dx. \tag{A.11}$$

We summarize the above discussions in the following theorem.

Theorem 6. *Let Assumption 1 holds. Suppose at a time bidder i is staying on both objects, and both price clocks have not stopped. Then there exists a unique symmetric Bayesian Nash equilibrium, at which bidder i will first drop out on object B, with the drop-out level s_i^B that solves equation* (A.6). *Specifically, if bidder i's average value $(v_i^A + v_i^B + \theta)/2$, exceeds $\overline{v}^A$, then bidder i will drop out on A and B simultaneously at her average value, i.e. $s_i^A = s_i^B = (v_i^A + v_i^B + \theta)/2$. Otherwise,*

there are four situations:

(i) If $v_i^A + \theta_i < \overline{v}^A$*, and* $J_1(v_i^A, n^A) > 0$*, then* $s_i^B \in [v_i^A, v_i^A + \theta_i]$*, and* $s_i^A = s_i^B$ *after she drops out on* B*.*

(ii) If $v_i^A + \theta_i < \overline{v}^A$*, and* $J_1(v_i^A, n^A) < 0$*, then* $s_i^B \in [v_i^B, v_i^A]$*, and* $s_i^A = v_i^A$ *after she drops out on* B*.*

(iii) If $v_i^A + \theta_i > \overline{v}^A$*, and* $J_2(v_i^A, n^A) > 0$*, then* $s_i^B \in [v_i^A, v_i^A + \theta_i]$*, and* $s_i^A = s_i^B$ *after she drops out on* B*.*

(iv) If $v_i^A + \theta_i > \overline{v}^A$*, and* $J_2(v_i^A, n^A) < 0$*, then* $s_i^B \in [v_i^B, v_i^A]$*, and* $s_i^A = v_i^A$ *after she drops out on* B*.*

The full characterization of the Bayesian Nash Equilibrium in this two-object clock auction is given in the following theorem.

Theorem 7 (Bayesian Nash Equilibrium)**.** *Consider a clock auction with two heterogeneous objects whose value structure distribution satisfies Assumption 1, where there are no local bidders and multiple global bidders. Then the Bayesian Nash equilibrium is the strategy profile* $(s_i^A, s_i^B)_{i=1}^n$ *at every time, such that*

(i) If bidder i *has dropped out or won an object (which must be object* B *according to Lemma 1), then bidder* i*'s drop out level on* A*,* s_i^A *follow Lemma 2.*

(ii) If bidder i *is staying on both objects and both price clocks have not stopped, then bidder* i*'s strategy* (s_i^A, s_i^B) *follows Theorem 6.*

A.1.2 Multiple Objects

For the case where there are $m > 2$ objects ($j = 1, \cdots, m$), we first analyze a two-bidder situation ($i = 1, 2$), to show how the bidders update the information of the winner identity during the auction. We derive the equilibrium strategy in a recursive manner. Finally we turn to the case where there are multiple objects and multiple (global) bidders.

The knowledge of the identity of winners on previously closed objects will enormously complicate bidders' beliefs, because winners will update their values on the remaining objects, due to the complementarity coming from the newly obtained object. As the number of objects increase, such complexity will quickly explode so that bidders may not be able to form a correct belief. In such a extremely complicated decision making, it is not unreasonable to assume that players only make use of partial information in hand. We assume that bidders' strategy only depend on the number of active opponents in the remaining objects. In other words, a bidder does not take into account which bidder has won which objects before. Such restricted information structure removes the necessarity of updating opponents' value functions, making bidders' strategy more tractable.

Parallel to Assumption 1, in the multiple-object case, we also need to impose the assumption on the value structure: bidders' stand-alone values are ordered in the same way. Let $M_j \equiv \{1, \cdots, j\}$, $j = 1, \cdots, m$.

Assumption 13 (Strict Value Ordering)**.** *The value structure distribution F satisfies:*

$$\mathbb{P}[\underline{v}^j \geq \overline{v}^{j+1} + \theta_i^{j+1}] = 1, \;\; j = 1, \cdots, m-1, \tag{A.12}$$

where $\underline{v}^j$ and $\overline{v}^j$ are the lower bound and upper bound of the distribution of the stand-alone value of object j, and $\theta_i^j = \theta_i(M_j) - \theta_i(M_{j-1})$ is the complementarity that object j brings to objects M_{j-1}.

Let $\Pi_i^j(v_i, p, w, F)$ denote the expected utility of bidder i at equilibrium, when there are j objects remaining in the auction. The current price is p, and the current allocation result is $w = (w_1, w_2)$, with w_i being the set of objects bidder i has already won. F is the full distribution function on all the m objects. $\mathbf{v}_i = \mathbf{v}_i(M_j)|w_i$ is the *conditional* value structure of bidder i on object set $M_j = \{1, \cdots, j\}$, given she obtains w_i. Then,

$$\Pi_i^j(\mathbf{v}_i, p, w, F) = \max_{s_i^j} \Bigg\{ \int_{s_i^j}^{\overline{v}^1} \Pi_i^{j-1}(\mathbf{v}_i, s_{-i}^j, w'', F'') d\, G_{-i}^j(s_{-i}^j | p, w, F) + \int_p^{s_i^j} \left[v_i^j - s_{-i}^j + \Pi_i^{j-1}(v_i', s_{-i}^j, w', F') \right] d\, G_{-i}^j(s_{-i}^j | p, w, F) \Bigg\}, \tag{A.13}$$

where w' is updated from w by adding object j to w_{-i}, and w'' is by adding j to w_i; F' and F'' are bidder i's updated beliefs on the value structure of objects $M_{j-1} = \{1, \cdots, j-1\}$ of bidder $-i$, given w' and w'', respectively; $v_i^j = v_i^j | w_i$ is the marginal stand-alone value of bidder i on object j, given she wins w_i; $v_i' = \mathbf{v}_i(M_{j-1}) | w_i'$ is the marginal value structure of bidder i on object set M_{j-1}, given she wins w_i'. First order condition yields:

$$s_i^j = v_i^j + \Pi_i^{j-1}(\mathbf{v}_i', s_i^j, w', F') - \Pi_i^{j-1}(\mathbf{v}_i, s_i^j, w'', F''). \tag{A.14}$$

The complementarity effect happens in $v_i' \geq v_i$, (to be rigorously proved) leading to: $\Pi_i^{j-1}(\mathbf{v}_i', s_i^j, w', F') \geq \Pi_i^{j-1}(\mathbf{v}_i, s_i^j, w'', F'')$. However, we expect that exposure problem exists as well, i.e. $s_i^j < v_i^j + \theta_i^j$, where $\theta_i^j = \theta_i(M_j) - \theta_i(M_{j-1})$ is the complementarity that object j brings to objects M_{j-1}, which is the difference between v_i' and v_i. The intuition is that the bidder is at risk of winning only a subset of the remaining objects.

Theorem 8 (Multiple Objects with Two Bidders)**.** *Let Assumption 13 holds. Consider multiple objects with two bidders. The Bayesian Nash equilibrium strategy s_i^j solves equation* (A.14)*, in which Π_i^j follows a recursive form* (A.13)*. In addition, the belief system $F(w)$ is consistent with the equilibrium outcome. For $j = 1, 2$, the equilibrium strategy follows from Theorem 6.*

Suppose there are n bidders in the auction. By Assumption 13 on the value structure, each bidder will drop out on object $m, m-1, \cdots, 1$ in turn, i.e. the equilibrium drop-out level satisfies $s_i^m \leq s_i^{m-1} \leq \cdots \leq s_i^1$. This is similar to Lemma 1 in A.1.1. What's more, when the last bidder is dropping out at object j, there will be n bidders active on object set M_{j-1}. Therefore, bidders

do not update their beliefs on other bidders' value based on number of remaining bidders. Let $\Pi_i^j(\mathbf{v}_i, p, n, F(w_i))$ denote the expected utility of bidder i at equilibrium, when object set M_j is alive in the auction. With some abuse of notation, let $\mathbf{v}_i$ be the updated value structure, given bidder i's currently winning objects w_i. Let $F(w_i)$ be the distribution on all other bidders' value structure on set M_j when objects $J/M_j = \{j+1, \cdots, m\}$ are closed, given w_i. Then, the expected payoff is

$$\begin{aligned}\Pi_i^j(v_i, p, n, F(w_i)) = \max_{s_i^j} \Bigg\{ & \int_{s_i^j}^{\overline{v}^1} \Pi_i^{j-1}(v_i, s_{-i}^j, n, F(w_i)) d\, G_{-i}^j(s_{-i}^j | p, n, F(w_i)) \\ & + \int_p^{s_i^j} \left[v_i^j - s_{-i}^j + \Pi_i^{j-1}(v_i', s_{-i}^j, n, F(w_i')) \right] d\, G_{-i}^j(s_{-i}^j | p, n, F(w_i)) \Bigg\},\end{aligned} \tag{A.15}$$

And the first order condition is

$$s_i^j = v_i^j + \Pi_i^{j-1}(v_i', s_i^j, n, F(w_i')) - \Pi_i^{j-1}(v_i, s_i^j, n, F(w_i)). \tag{A.16}$$

Theorem 9 (Multiple Objects with Multiple Bidders)**.** *Let Assumption 13 holds. Consider multiple objects with multiple bidders. The Bayesian Nash equilibrium strategy s_i^j solves equation* (A.16)*, in which Π_i^j follows a recursive form* (A.15)*. In addition, the belief system $F(w)$ is consistent with the equilibrium outcome. For $j = 1, 2$, the equilibrium strategy follows from Theorem 6.*

A.2 Proof of Theorem 6

We consider the four cases in Theorem 6. From the previous discussion, it suffices to show that $J_1(s, n^A)$ and $J_2(s, n^A)$ are non-increasing in $s \in [v_i^B, \lambda_i]$. Using the technique of integration by part, we derive the first order condition for each case.

Case I. $v_i^A + \theta_i < \overline{v}^A$**, and** $s_i^B < v_i^A$**.**

In this case the first order condition becomes:

$$
\begin{aligned}
s_i^B &= v_i^B + \int_{s_i^B}^{v_i^A+\theta_i} (v_i^A + \theta_i - s_{-i}^A) d\, G_{-i,1}^A(s_{-i}^A | s_i^B, n^A) - \int_{s_i^B}^{v_i^A} (v_i^A - s_{-i}^A) d\, G_{-i,0}^A(s_{-i}^A | s_i^B, n^A) \\
&= v_i^B + \int_{s_i^B}^{v_i^A+\theta_i} G_{-i,1}^A(x | s_i^B, n^A) dx - \int_{s_i^B}^{v_i^A} G_{-i,0}^A(x | s_i^B, n^A) dx \\
&= v_i^B + \int_{v_i^A}^{v_i^A+\theta_i} G_{-i,1}^A(x | s_i^B, n^A) dx + \int_{s_i^B}^{v_i^A} \left[G_{-i,1}^A(x | s_i^B, n^A) - G_{-i,0}^A(x | s_i^B, n^A) \right] dx. \quad \text{(A.17)}
\end{aligned}
$$

Case II. $v_i^A + \theta_i < \overline{v}^A$**, and** $s_i^B > v_i^A$**.**

In this case the first order condition becomes:

$$
\begin{aligned}
s_i^B &= v_i^B + \int_{s_i^B}^{v_i^A+\theta_i} (v_i^A + \theta_i - s_{-i}^A) d\, G_{-i,1}^A(s_{-i}^A | s_i^B, n^A) \\
&= v_i^B + \int_{s_i^B}^{v_i^A+\theta_i} G_{-i,1}^A(x | s_i^B, n^A) dx. \quad \text{(A.18)}
\end{aligned}
$$

Case III. $v_i^A + \theta_i > \overline{v}^A$**, and** $s_i^B < v_i^A$**.**

In this case the first order condition becomes:

$$
\begin{aligned}
s_i^B &= v_i^B + \int_{s_i^B}^{\overline{v}^A} (v_i^A + \theta_i - s_{-i}^A) d\, G_{-i,1}^A(s_{-i}^A | s_i^B, n^A) - \int_{s_i^B}^{v_i^A} (v_i^A - s_{-i}^A) d\, G_{-i,0}^A(s_{-i}^A | s_i^B, n^A) \\
&= v_i^B + \theta_i + v_i^A - \overline{v}^A + \int_{s_i^B}^{\overline{v}^A} G_{-i,1}^A(x | s_i^B, n^A) dx - \int_{s_i^B}^{v_i^A} G_{-i,0}^A(x | s_i^B, n^A) dx. \quad \text{(A.19)} \\
&= v_i^B + \theta_i + v_i^A - \overline{v}^A + \int_{v_i^A}^{\overline{v}_i^A} G_{-i,1}^A(x | s_i^B, n^A) dx + \int_{s_i^B}^{v_i^A} \left[G_{-i,1}^A(x | s_i^B, n^A) - G_{-i,0}^A(x | s_i^B, n^A) \right] dx.
\end{aligned}
$$

Case IV. $v_i^A + \theta_i > \overline{v}^A$**, and** $s_i^B > v_i^A$**.**

In this case the first order condition becomes:

$$
\begin{aligned}
s_i^B &= v_i^B + \int_{s_i^B}^{\overline{v}^A} (v_i^A + \theta_i - s_{-i}^A) d\, G_{-i,1}^A(s_{-i}^A | s_i^B, n^A) \\
&= v_i^B + \theta_i + v_i^A - \overline{v}^A + \int_{s_i^B}^{\overline{v}^A} G_{-i,1}^A(x | s_i^B, n^A) dx. \qquad \text{(A.20)}
\end{aligned}
$$

From the four cases, it is clear that in order to show $J_1(s, n^A)$ and $J_2(s, n^A)$ are non-increasing in $s \in [v_i^B, \lambda_i]$, it suffices to show that $G_{-i,1}^A(x|s, n^A)$ and $\Delta G_{-i}^A(x|s, n^A) \equiv G_{-i,1}^A(x|s, n^A) - G_{-i,0}^A(x|s, n^A)$ are both non-increasing in s, $\forall x \in [s, \overline{v}^A]$.

Without loss of generality, consider θ_i to be a constant. From equation (A.2), let $G_{-i,1}^A(x|s, n) = [K(s, x)]^n$, where $K(s, x) = (F^A(x) - F^A(s))/(1 - F^A(s))$. It is obvious that $K(s, x)$ is non-increasing in s for every $x \in [s, \overline{v}^A]$, and so is $G_{-i,1}^A(x|s, n)$. On the other hand, we can write $\Delta G_{-i}^A(x|s, n) = [K(s, x)]^{n-1} \cdot [K(s, x) - K(s - \theta; x - \theta)]$. Let $k(s, x) = \partial K(s, x)/\partial s$, which is non-positive from above. Then we have

$$
\begin{aligned}
\frac{\partial \Delta G_{-i}^A(x|s, n)}{\partial s} &= (n-1) k(s, x) \big[K(s, x)\big]^{n-2} \big[K(s, x) - K(s - \theta, x - \theta)\big] \\
&\quad + \big[K(s, x)\big]^{n-1} \cdot \big[k(s, x) - k(s - \theta, x - \theta)\big] \\
&= n\, k(s, x) \big[K(s, x)\big]^{n-1} - \big[K(s, x)\big]^{n-2} \cdot \\
&\quad \big[(n-1) k(s, x) K(s - \theta, x - \theta) + K(s, x) k(s - \theta, x - \theta)\big] \\
&\leq 0. \qquad \text{(A.21)}
\end{aligned}
$$

A.3 Proof of Proposition 1

In this section, we prove the comparative static properties of the Bayesian Nash equilibrium of the two-object clock auction, which is summarized in Proposition 1.

A.3.1 s_i^B Increasing in v_i^B

For the four cases, differentiate with respect to v_i^B on both sides of the first order conditions derived in Appendix A.2.

Case I. $v_i^A + \theta_i < \overline{v}^A$**, and** $s_i^B < v_i^A$**.**

$$\begin{aligned}\frac{\partial s_i^B}{\partial v_i^B} = 1 + \int_{v_i^A}^{v_i^A+\theta_i} \left[\frac{\partial G_{-i,1}^A(x|s_i^B, n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial v_i^B}\right]dx \\ + \int_{s_i^B}^{v_i^A} \left[\frac{\partial \Delta G_{-i}^A(x|s_i^B, n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial v_i^B}\right]dx - \Delta G_{-i}^A(x|s_i^B, n^A)\frac{\partial s_i^B}{\partial v_i^B},\end{aligned} \tag{A.22}$$

which leads to

$$\frac{\partial s_i^B}{\partial v_i^B}\left\{1 - \int_{v_i^A}^{v_i^A+\theta_i} \frac{\partial G_{-i,1}^A(x|s_i^B, n^A)}{\partial s_i^B}dx - \int_{s_i^B}^{v_i^A}\left[\frac{\partial \Delta G_{-i}^A(x|s_i^B, n^A)}{\partial s_i^B}\right]dx + \Delta G_{-i}^A(x|s_i^B, n^A)\right\} = 1. \tag{A.23}$$

Case II. $v_i^A + \theta_i < \overline{v}^A$**, and** $s_i^B > v_i^A$**.**

$$\frac{\partial s_i^B}{\partial v_i^B} = 1 + \int_{s_i^B}^{v_i^A+\theta_i} \left[\frac{\partial G_{-i,1}^A(x|s_i^B, n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial v_i^B}\right]dx - G_{-i,1}^A(x|s_i^B, n^A)\frac{\partial s_i^B}{\partial v_i^B}, \tag{A.24}$$

resulting in

$$\frac{\partial s_i^B}{\partial v_i^B}\left\{1 - \int_{s_i^B}^{v_i^A+\theta_i} \frac{\partial G_{-i,1}^A(x|s_i^B, n^A)}{\partial s_i^B} + G_{-i,1}^A(x|s_i^B, n^A)dx\right\} = 1. \tag{A.25}$$

Case III. $v_i^A + \theta_i > \overline{v}^A$**, and** $s_i^B < v_i^A$**.**

The derivative is the same as Case I.

Case IV. $v_i^A + \theta_i > \overline{v}^A$**, and** $s_i^B > v_i^A$**.**

The derivative is the same as Case II.

It could be easily verified that $\frac{\partial G^A_{-i,1}(x|s_i^B,n^A)}{\partial s_i^B} \leq 0$ and $\frac{\Delta\partial G^A_{-i}(x|s_i^B,n^A)}{\partial s_i^B} \leq 0$. In addition, $\Delta G^A_{-i}(x|s_i^B,n^A) \geq 0$. Therefore, Hence, for all the four cases, we have $\frac{\partial s_i^B}{\partial v_i^B} \geq 0$.

A.3.2 s_i^B Increasing in v_i^A

Again, we differentiate s_i^B with respect to v_i^A on both sides of the first order conditions derived in Appendix A.2.

Case I. $v_i^A + \theta_i < \overline{v}^A$**, and** $s_i^B < v_i^A$**.**

$$\frac{\partial s_i^B}{\partial v_i^A} = \int_{v_i^A}^{v_i^A+\theta_i} \left[\frac{\partial G^A_{-i,1}(x|s_i^B,n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial v_i^A}\right]dx + \int_{s_i^B}^{v_i^A} \left[\frac{\partial \Delta G^A_{-i}(x|s_i^B,n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial v_i^A}\right]dx$$
$$+\Delta G^A_{-i}(x|s_i^B,n^A) - \Delta G^A_{-i}(x|s_i^B,n^A)\frac{\partial s_i^B}{\partial v_i^A},$$

which gives us,

$$\frac{\partial s_i^B}{\partial v_i^A}\left\{1 - \int_{v_i^A}^{v_i^A+\theta_i}\left[\frac{\partial G^A_{-i,1}(x|s_i^B,n^A)}{\partial s_i^B}\right]dx - \int_{s_i^B}^{v_i^A}\left[\frac{\partial \Delta G^A_{-i}(x|s_i^B,n^A)}{\partial s_i^B}dx\right] + \Delta G^A_{-i}(x|s_i^B,n^A)\right\}$$
$$= \Delta G^A_{-i}(x|s_i^B,n^A). \tag{A.26}$$

Case II. $v_i^A + \theta_i < \overline{v}^A$**, and** $s_i^B > v_i^A$**.**

$$\frac{\partial s_i^B}{\partial v_i^A} = \int_{s_i^B}^{v_i^A+\theta_i}\left[\frac{\partial G^A_{-i,1}(x|s_i^B,n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial v_i^A}\right]dx + G^A_{-i,1}(x|s_i^B,n^A) - G^A_{-i,1}(x|s_i^B,n^A)\frac{\partial s_i^B}{\partial v_i^A}.$$

Thus,

$$\frac{\partial s_i^B}{\partial v_i^A}\left\{1 - \int_{s_i^B}^{v_i^A+\theta_i}\left[\frac{\partial G^A_{-i,1}(x|s_i^B,n^A)}{\partial s_i^B}\right]dx + G^A_{-i,1}(x|s_i^B,n^A)\right\} = G^A_{-i,1}(x|s_i^B,n^A). \tag{A.27}$$

Case III. $v_i^A + \theta_i > \overline{v}^A$**, and** $s_i^B < v_i^A$**.**

$$\frac{\partial s_i^B}{\partial v_i^A} = 1 - G^A_{-i,1}(x|s_i^B, n^A) + \int_{v_i^A}^{\overline{v}_i^A} \left[\frac{\partial G^A_{-i,1}(x|s_i^B, n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial v_i^A}\right]dx$$
$$+ \int_{s_i^B}^{v_i^A} \left[\frac{\partial \Delta G^A_{-i}(x|s_i^B, n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial v_i^A}\right]dx + \Delta G^A_{-i}(x|s_i^B, n^A) - \Delta G^A_{-i}(x|s_i^B, n^A)\frac{\partial s_i^B}{\partial v_i^A}, \quad \text{(A.28)}$$

leading to

$$\frac{\partial s_i^B}{\partial v_i^A}\left\{1 - \int_{v_i^A}^{\overline{v}_i^A} \left[\frac{\partial G^A_{-i,1}(x|s_i^B, n^A)}{\partial s_i^B}\right]dx - \int_{s_i^B}^{v_i^A} \left[\frac{\partial \Delta G^A_{-i}(x|s_i^B, n^A)}{\partial s_i^B}dx\right] + \Delta G^A_{-i}(x|s_i^B, n^A)\right\}$$
$$= 1 - G^A_{-i,1}(x|s_i^B, n^A) + \Delta G^A_{-i}(x|s_i^B, n^A). \quad \text{(A.29)}$$

Case IV. $v_i^A + \theta_i > \overline{v}^A$**, and** $s_i^B > v_i^A$**.**

$$\frac{\partial s_i^B}{\partial v_i^A} = 1 + \int_{s_i^B}^{\overline{v}^A} \frac{\partial G^A_{-i,1}(x|s_i^B, n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial v_i^A}dx - G^A_{-i,1}(x|s_i^B, n^A)\frac{\partial s_i^B}{\partial v_i^A},$$

yielding

$$\frac{\partial s_i^B}{\partial v_i^A}\left\{1 - \int_{s_i^B}^{\overline{v}^A} \frac{\partial G^A_{-i,1}(x|s_i^B, n^A)}{\partial s_i^B}dx + G^A_{-i,1}(x|s_i^B, n^A)\right\} = 1.$$

Overall, for all the four cases, by the fact that $\frac{\partial G^A_{-i,1}(x|s_i^B, n^A)}{\partial s_i^B} \leq 0$, $\frac{\Delta \partial G^A_{-i}(x|s_i^B, n^A)}{\partial s_i^B} \leq 0$, and $\Delta G^A_{-i}(x|s_i^B, n^A) \geq 0$, we have $\frac{\partial s_i^B}{\partial v_i^A} \geq 0$.

A.3.3 s_i^B Increasing in θ_i

Take derivative of s_i^B with respect to θ_i for the four cases.

Case I. $v_i^A + \theta_i < \overline{v}^A$**, and** $s_i^B < v_i^A$**.**

$$\frac{\partial s_i^B}{\partial \theta_i} = \int_{v_i^A}^{v_i^A+\theta_i} \left[\frac{\partial G_{-i,1}^A(x|s_i^B, n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial \theta_i}\right]dx + G_{-i,1}^A(x|s_i^B, n^A) + \int_{s_i^B}^{v_i^A} \left[\frac{\partial \Delta G_{-i}^A(x|s_i^B, n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial \theta_i}\right]dx$$
$$-\Delta G_{-i}^A(x|s_i^B, n^A)\frac{\partial s_i^B}{\partial \theta_i},$$

giving,

$$\frac{\partial s_i^B}{\partial v_i^A}\left\{1 - \int_{v_i^A}^{v_i^A+\theta_i} \left[\frac{\partial G_{-i,1}^A(x|s_i^B, n^A)}{\partial s_i^B}\right]dx - \int_{s_i^B}^{v_i^A} \left[\frac{\partial \Delta G_{-i}^A(x|s_i^B, n^A)}{\partial s_i^B}dx\right] + \Delta G_{-i}^A(x|s_i^B, n^A)\right\}$$
$$= G_{-i,1}^A(x|s_i^B, n^A). \tag{A.30}$$

Case II. $v_i^A + \theta_i < \overline{v}^A$**, and** $s_i^B > v_i^A$**.**

$$\frac{\partial s_i^B}{\partial \theta_i} = \int_{s_i^B}^{v_i^A+\theta_i} \left[\frac{\partial G_{-i,1}^A(x|s_i^B, n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial \theta_i}\right]dx + G_{-i,1}^A(x|s_i^B, n^A) - G_{-i,1}^A(x|s_i^B, n^A)\frac{\partial s_i^B}{\partial \theta_i}.$$

Thus, we have

$$\frac{\partial s_i^B}{\partial \theta_i}\left\{1 - \int_{s_i^B}^{v_i^A+\theta_i} \left[\frac{\partial G_{-i,1}^A(x|s_i^B, n^A)}{\partial s_i^B}\right]dx + G_{-i,1}^A(x|s_i^B, n^A)\right\} = G_{-i,1}^A(x|s_i^B, n^A). \tag{A.31}$$

Case III. $v_i^A + \theta_i > \overline{v}^A$**, and** $s_i^B < v_i^A$**.**

$$\frac{\partial s_i^B}{\partial \theta_i} = 1 + \int_{v_i^A}^{\overline{v}_i^A} \left[\frac{\partial G_{-i,1}^A(x|s_i^B, n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial \theta_i}\right]dx$$
$$+ \int_{s_i^B}^{v_i^A} \left[\frac{\partial \Delta G_{-i}^A(x|s_i^B, n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial \theta_i}\right]dx - \Delta G_{-i}^A(x|s_i^B, n^A)\frac{\partial s_i^B}{\partial \theta_i}, \tag{A.32}$$

leading to

$$\frac{\partial s_i^B}{\partial \theta_i}\left\{1-\int_{v_i^A}^{\overline{v}_i^A}\left[\frac{\partial G_{-i,1}^A(x|s_i^B,n^A)}{\partial s_i^B}\right]dx-\int_{s_i^B}^{v_i^A}\left[\frac{\partial \Delta G_{-i}^A(x|s_i^B,n^A)}{\partial s_i^B}dx\right]+\Delta G_{-i}^A(x|s_i^B,n^A)\right\}=1. \tag{A.33}$$

Case IV. $v_i^A+\theta_i>\overline{v}^A$**, and** $s_i^B>v_i^A$**.**

$$\frac{\partial s_i^B}{\partial \theta_i}=1+\int_{s_i^B}^{\overline{v}^A}\frac{\partial G_{-i,1}^A(x|s_i^B,n^A)}{\partial s_i^B}\frac{\partial s_i^B}{\partial \theta_i}dx-G_{-i,1}^A(x|s_i^B,n^A)\frac{\partial s_i^B}{\partial \theta_i},$$

which give rise to

$$\frac{\partial s_i^B}{\partial \theta_i}\left\{1-\int_{s_i^B}^{\overline{v}^A}\frac{\partial G_{-i,1}^A(x|s_i^B,n^A)}{\partial s_i^B}dx+G_{-i,1}^A(x|s_i^B,n^A)\right\}=1.$$

Hence, for all the four cases, since $\frac{\partial G_{-i,1}^A(x|s_i^B,n^A)}{\partial s_i^B}\leq 0$, $\frac{\Delta\partial G_{-i}^A(x|s_i^B,n^A)}{\partial s_i^B}\leq 0$, and $\Delta G_{-i}^A(x|s_i^B,n^A)\geq 0$, we have $\frac{\partial s_i^B}{\partial \theta_i}\geq 0$.

A.3.4 s_i^B Decreasing in n^A

Observing the four cases, it suffices to show that

$$\frac{\partial G_{-i,1}^A(x|s_i^B,n^A)}{\partial n^A}\leq 0,\quad \frac{\partial \Delta G_{-i}^A(x|s_i^B,n^A)}{\partial n^A}\leq 0, \tag{A.34}$$

which could be verified by their definition (A.2).

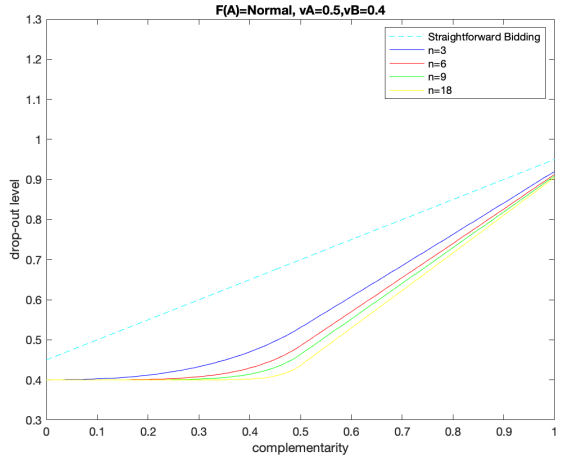
A.4 Numerical Solution for Equilibrium: F^A is Normal

Based on Theorem 7, I calculate numerically the BNE of the two-object clock auction when F^A is a truncated normal distribution in $[\underline{v}^A,\overline{v}^A]$. Using the same parameter settings as Section 2.3.3, I plot the corresponding graphs with F^A being normal. We can see that the pattern of s_i^B is robust to the specification of the common distribution for bidders' stand-alone value F^A. See

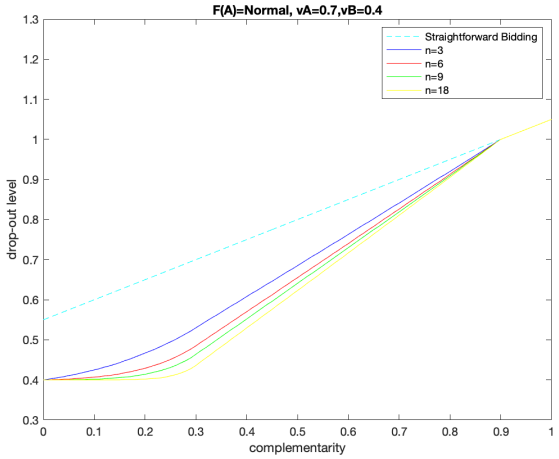
Figure A.1 and A.2.

Figure A.1: Equilibrium Bidding Strategy s_i^B and Complementarity θ_i.
Setting: F^A Normal. Property: s_i^B Non-decreasing in θ_i, and Non-increasing in n^A.

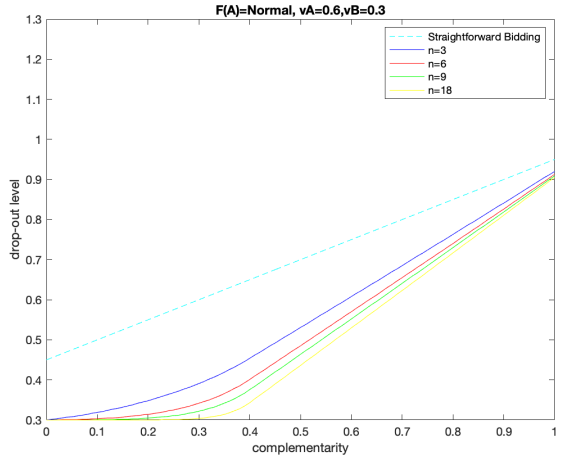
(a) $v_i^A = 0.5, v_i^B = 0.4$



(b) $v_i^A = 0.7, v_i^B = 0.4$



(c) $v^B = 0.3, v_i^A = 0.6$



(d) $v^A = 0.5, v_i^A = 0.6$

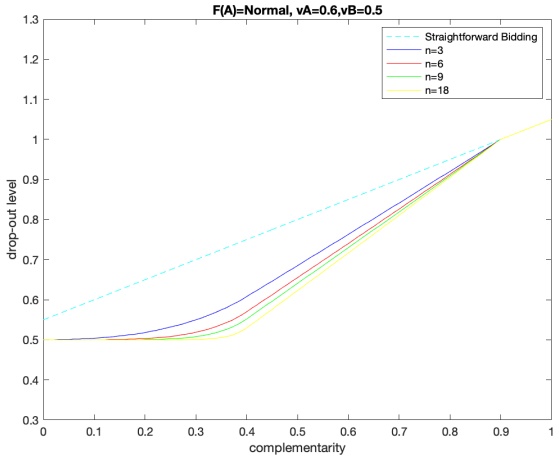
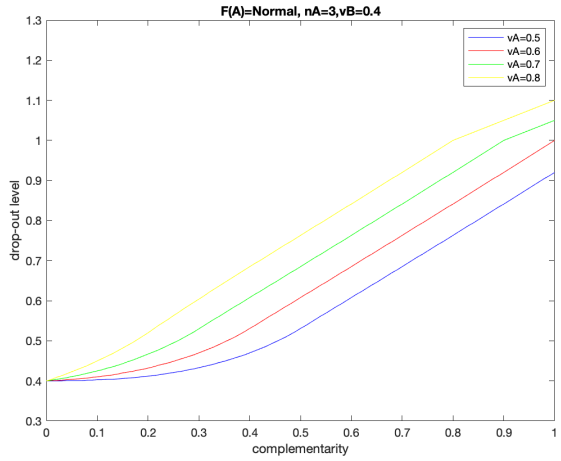
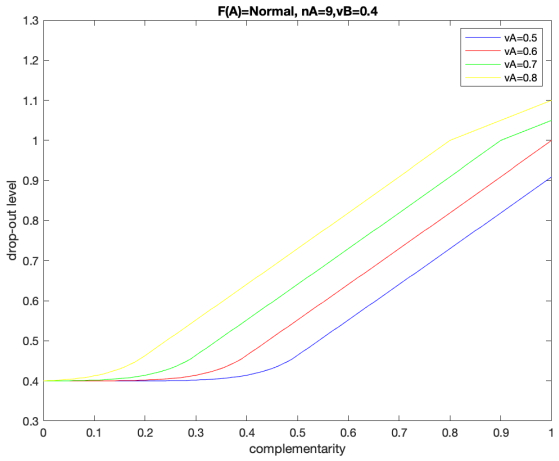


Figure A.2: Equilibrium Bidding Strategy s_i^B and Complementarity θ_i.
Setting: F^A Normal. Property: s_i^B Non-decreasing in v_i^A, v_i^B, and θ_i

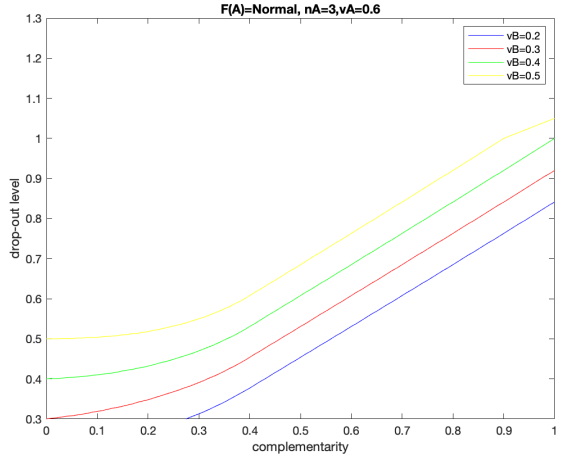
(a) $n^A = 3, v_i^B = 0.4$



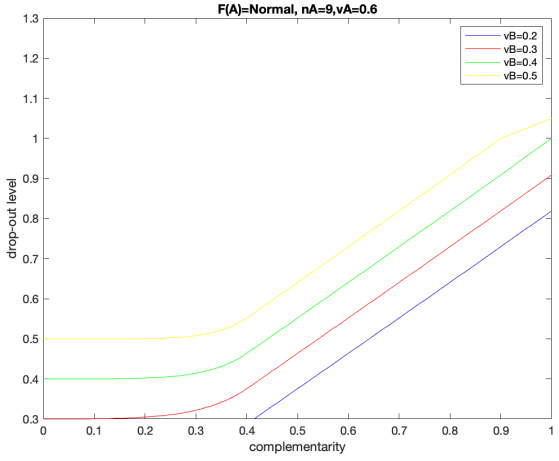
(b) $n^A = 9, v_i^B = 0.4$



(c) $n^A = 3, v_i^A = 0.6$



(d) $n^A = 9, v_i^A = 0.6$



APPENDIX B

APPENDIX FOR THE SECOND ESSAY

B.1 Proofs of Main Theorems and Propositions

Throughout the Appendix, for a $k \times k$ matrix A, let $\delta_{\max}(A)$ and $\delta_{\min}(A)$ be the largest and smallest singular value of A, respectively.

Lemma 3. *When $Z_{i,(s)}$ is a scalar, and $\lambda_{(s)} = 0$, we have*

$$\delta^2_{\max}(P_{(s)}) = O(K_{(s)}).$$

Proof of Lemma 3. We leave out the subscript (s) in this proof for notational ease. In this case $L(Z, z, \lambda) = \mathbf{1}(Z = z)$, and $\mathbf{1}(\cdot)$ is the indicator function. Let $\mathbf{1}_z = \mathrm{diag}\{\mathbf{1}(Z_1 = z), \ldots, \mathbf{1}(Z_n = z)\}$. Then $\mathcal{L}_z = \mathbf{1}_z$. We start by giving a simple identity: for every Z_i,

$$1 = \sum_{z \in \mathcal{M}_Z} \mathbf{1}(Z_i = z). \tag{B.1}$$

Then,

$$
\begin{aligned}
\delta^2_{\max}(P) \le \|P\|^2 = \operatorname{tr}(P'P) &= \sum_{i=1}^{n} P_i' P_i \\
&= \sum_{i=1}^{n} \left[\mathcal{B}_i' (\mathbf{B}' \mathcal{L}_{Z_i} \mathbf{B})^{-1} \mathbf{B}' \mathcal{L}_{Z_i} \cdot \mathcal{L}_{Z_i} \mathbf{B} (\mathbf{B}' \mathcal{L}_{Z_i} \mathbf{B})^{-1} \mathcal{B}_i \right] \\
&= \sum_{i=1}^{n} \left[\mathcal{B}_i' (\mathbf{B}' \mathbf{1}_{Z_i} \mathbf{B})^{-1} \mathbf{B}' \mathbf{1}_{Z_i} \cdot \mathbf{1}_{Z_i} \mathbf{B} (\mathbf{B}' \mathbf{1}_{Z_i} \mathbf{B})^{-1} \mathcal{B}_i \right] \\
&= \sum_{i=1}^{n} \left[\mathcal{B}_i' (\mathbf{B}' \mathbf{1}_{Z_i} \mathbf{B})^{-1} \mathcal{B}_i \right] \\
&= \sum_{i=1}^{n} \left[\mathcal{B}_i' \left(\sum_{m=1}^{n} \mathcal{B}_m \mathcal{B}_m' \mathbf{1}(Z_m = Z_i) \right)^{-1} \mathcal{B}_i \right] \\
&= \sum_{i=1}^{n} \sum_{z \in \mathcal{M}_Z} \mathbf{1}(Z_i = z) \left[\mathcal{B}_i' \left(\sum_{m=1}^{n} \mathcal{B}_m \mathcal{B}_m' \mathbf{1}(Z_m = Z_i) \right)^{-1} \mathcal{B}_i \right] \text{ by Equation (B.1),} \\
&= \sum_{z \in \mathcal{M}_Z} \sum_{i=1}^{n} \mathbf{1}(Z_i = z) \left[\mathcal{B}_i' \left(\sum_{m=1}^{n} \mathcal{B}_m \mathcal{B}_m' \mathbf{1}(Z_m = Z_i) \right)^{-1} \mathcal{B}_i \right] \\
&= \sum_{z \in \mathcal{M}_Z} \operatorname{tr} \left\{ \sum_{i=1}^{n} \mathbf{1}(Z_i = z) \left[\mathcal{B}_i' \left(\sum_{m=1}^{n} \mathcal{B}_m \mathcal{B}_m' \mathbf{1}(Z_m = z) \right)^{-1} \mathcal{B}_i \right] \right\} \\
&= \sum_{z \in \mathcal{M}_Z} \operatorname{tr} \left\{ \left(\sum_{i=1}^{n} \mathbf{1}(Z_i = z) \mathcal{B}_i \mathcal{B}_i' \right) \left(\sum_{m=1}^{n} \mathcal{B}_m \mathcal{B}_m' \mathbf{1}(Z_m = z) \right)^{-1} \right\} \\
&= \sum_{z \in \mathcal{M}_Z} \operatorname{tr}\{I_K\} = c \cdot K,
\end{aligned}
$$

where $c = \Pi_{l=1}^{r} c_l$. Therefore, $\delta^2_{\max}(P) = O(K)$. □

Lemma 4. *Under Assumption 5, for every* $s = 1, 2, \ldots, S_n$,

$$
\delta^2_{\max}(P_{(s)}) \le \operatorname{tr}(P_{(s)}' P_{(s)}) = O_p \left(K_{(s)} + \sum_{l=1}^{r_s} \lambda_{l,(s)} \cdot K_{(s)}^2 \right),
$$

where $P_{(s)}$ *is the projection matrix of the* s*-th model whose expression is given by Equation* (3.5).

Proof of Lemma 4. We follow the notation in Lemma 3. Recall from Equation (3.4) that $L(Z, z, \lambda) =$

$\mathbf{1}(Z = z) + \sum_{l=1}^{r} \lambda_l \mathbb{I}_{(l)}(Z, z) + O(||\lambda||^2)$, where $\mathbb{I}_{(l)}(Z, z) = \mathbf{1}(Z_l \neq z_l) \prod_{j \neq l}^{r} \mathbf{1}(Z_j = z_j)$. Let $\mathcal{H}_{z,l} = diag(\mathbb{I}_{(l)}(Z_1, z), \cdots, \mathbb{I}_{(l)}(Z_n, z))$. Then $\mathcal{L}_z = \mathcal{J}_z + \sum_{l=1}^{r} \lambda_l \mathcal{H}_{z,l} + O(||\lambda||^2)\, I_n$. Because $\mathbb{I}_{(l)}(Z, z)\mathbb{I}_{(j)}(Z, z) = 0$ for $l \neq j$, and $\mathbb{I}^2_{(l)}(Z, z) = \mathbb{I}_{(l)}(Z, z)$ for any l, we have

$$\begin{aligned} L^2(Z, z, \lambda) &= \left(\mathbf{1}(Z = z) + \sum_{l=1}^{r} \lambda_l \mathbb{I}_{(l)}(Z_l, z_l) + O(||\lambda||^2) \right)^2 \\ &= \mathbf{1}(Z = z) + \sum_{l=1}^{r} \lambda_l^2 \mathbb{I}^2_{(l)}(Z_l, z_l) + O(\|\lambda\|^3) \\ &= \mathbf{1}(Z = z) + \sum_{l=1}^{r} \lambda_l^2 \mathbb{I}_{(l)}(Z_l, z_l) + O(\|\lambda\|^3) \end{aligned}$$

Therefore, $\mathcal{L}_z^2 = \mathcal{J}_z + \sum_{l=1}^{r} \lambda_l^2 \mathcal{H}_{l,z} + O(\|\lambda\|^3)\, I_n$.

Letting $S_z = \frac{1}{n}\mathbf{B}'\mathcal{J}_z\mathbf{B}$, $S_i = S_{Z_i} = \frac{1}{n}\mathbf{B}'\mathcal{J}_{Z_i}\mathbf{B}$, $T_{l,i} = \frac{1}{n}\mathbf{B}'\mathcal{H}_{l,Z_i}\mathbf{B}$, and $U = \frac{1}{n}\mathbf{B}'\mathbf{B}$, then

$$\begin{aligned} \frac{1}{n}\mathbf{B}'\mathcal{L}_{Z_i}\mathbf{B} &= S_i + \sum_{l=1}^{r} \lambda_l T_{l,i} + O(||\lambda||^2)\, U, \\ \frac{1}{n}\mathbf{B}'\mathcal{L}^2_{Z_i}\mathbf{B} &= S_i + \sum_{l=1}^{r} \lambda_l^2 T_{l,i} + O(\|\lambda\|^3)\, U. \end{aligned}$$

Hence,

$$
\begin{aligned}
\delta^2_{\max}(P) \le \|P\|^2 = \operatorname{tr}(P'P) &= \sum_{i=1}^{n} P_i' P_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{B}_i'\left(\frac{1}{n}\mathbf{B}'\mathcal{L}_{Z_i}\mathbf{B}\right)^{-1}\frac{1}{n}\mathbf{B}'\mathcal{L}_{Z_i}\cdot\mathcal{L}_{Z_i}\mathbf{B}\left(\frac{1}{n}\mathbf{B}'\mathcal{L}_{Z_i}\mathbf{B}\right)^{-1}\mathcal{B}_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{B}_i'\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\left(S_i+\sum_{l=1}^{r}\lambda_l^2 T_{l,i}+O(||\lambda||^3)\,U\right)\right. \\
&\left.\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\mathcal{B}_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{B}_i'\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\cdot S_i\cdot\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\mathcal{B}_i\right] \\
&+\sum_{l=1}^{r}\lambda_l^2\frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{B}_i'\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}T_{l,i}\right. \\
&\left.\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\mathcal{B}_i\right] \\
&+\frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{B}_i'\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}U\right. \\
&\left.\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\mathcal{B}_i\right]\cdot O(\|\lambda\|^3) \\
&\equiv \Theta_n+\sum_{l=1}^{r}\lambda_l^2\Psi_n+\,O(\|\lambda\|^3)\,\Xi_n.
\end{aligned}
$$

We observe that $\mathbb{I}_{(l)}(Z,z)=\sum_{\tilde{z}\in\mathcal{M}_Z}\mathbf{1}(Z=\tilde{z})$, where $\tilde{z}$ satisfies $\tilde{z}_l\neq z_l$, and $\tilde{z}_j=z_j$ for any $j\neq l$. Thus,

$$
T_{l,i}=\frac{1}{n}\mathbf{B}'\mathcal{H}_{l,Z_i}\mathbf{B}=\frac{1}{n}\mathbf{B}'\left(\sum_{\substack{\tilde{z}\in\mathcal{M}_Z\\ \tilde{z}_l\neq Z_{il}}}\mathcal{J}_{\tilde{z}}\right)\mathbf{B}=\sum_{\substack{\tilde{z}\in\mathcal{M}_Z\\ \tilde{z}_l\neq Z_{il}}}S_{\tilde{z}}.
$$

And, obviously, $U=\sum_{z\in\mathcal{M}_Z}S_z$. Thus, by Lemma C.4, the singular values of S_i, $T_{l,i}$ and U

are all bounded below away from zero and bounded above from a finite constant, in probability. Therefore,

$$\delta_{\max}(S_i^{-1}) = \delta_{\min}^{-1}(S_i) = O_p(1), \;\; \delta_{\max}(T_{l,i}) = O_p(1), \;\; \delta_{\max}(U) = O_p(1) \tag{B.2}$$

Consequently, with $\sum_{l=1}^{r} \lambda_l = o(1)$, we have

$$\begin{aligned}
\delta_{\max}\left(\sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U\right) &\leq \sum_{l=1}^{r} \delta_{\max}(\lambda_l S_i^{-1} T_{l,i}) + O(||\lambda||^2)\, \delta_{\max}(S_i^{-1} U) \\
&\leq \sum_{l=1}^{r} \lambda_l\, \delta_{\max}(S_i^{-1})\, \delta_{\max}(T_{l,i}) + O(||\lambda||^2)\, \delta_{\max}(S_i^{-1})\, \delta_{\max}(U) \\
&= O_p\left(\sum_{l=1}^{r} \lambda_l\right) + O(||\lambda||^2)\, O_p(1) \\
&= O_p\left(\sum_{l=1}^{r} \lambda_l\right) = o_p(1) < 1.
\end{aligned} \tag{B.3}$$

In addition,

$$\begin{aligned}
\left(S_i + \sum_{l=1}^{r} \lambda_l T_{l,i} + O(||\lambda||^2)\, U\right)^{-1} &= \left(S_i \left(I_k + \sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U\right)\right)^{-1} \\
&= \left(I_k + \sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U\right)^{-1} S_i^{-1},
\end{aligned} \tag{B.4}$$

provided the inverse exists. Note that S_i is positive definite and therefore invertible, and the invertibility of $\left(I_k + \sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U\right)$ is guaranteed by Equation (B.3) and Lemma C.3.

Now we calculate the rate of Θ_n, Ψ_n and Ξ_n.

$$
\begin{aligned}
\Theta_n &= \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{B}_i'\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1} S_i\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\mathcal{B}_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{B}_i'\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1} S_i^{-1}S_i\right. \\
&\quad \left.\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1} S_i^{-1}\mathcal{B}_i\right], \text{ by Equation (B.4)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}\left\{S_i^{-1}\mathcal{B}_i\mathcal{B}_i'\left[\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1}\right]^2\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}\left[S_i^{-1}\mathcal{B}_i\mathcal{B}_i'\right]+\frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}\left\{S_i^{-1}\mathcal{B}_i\mathcal{B}_i'\left[\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-2}-I_k\right]\right\} \\
&= \Theta_{1,n}+\Theta_{2,n}.
\end{aligned}
$$

From Lemma 3 we have that

$$
\Theta_{1,n}=\sum_{i=1}^{n}\left[\mathcal{B}_i'(\mathbf{B}'\mathcal{J}_{Z_i}\mathbf{B})^{-1}\mathcal{B}_i\right]=O(K). \tag{B.5}
$$

Next we find $\Theta_{2,n}$. Using Lemma C.2, C.5, and C.3, we have

$$
\begin{aligned}
\Theta_{2,n} &\leq \frac{1}{n}\sum_{i=1}^{n} K \cdot \delta_{\max}\left(S_i^{-1}\mathcal{B}_i\mathcal{B}_i' \left[\left(I_k + \sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U \right)^{-2} - I_k \right] \right), \text{ by Lemma C.2} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} K \cdot \delta_{\max}\left(S_i^{-1}\mathcal{B}_i\mathcal{B}_i' \right) \cdot \delta_{\max}\left(\left(I_k + \sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U \right)^{-2} - I_k \right) \\
&= \frac{K}{n}\sum_{i=1}^{n} \delta_{\max}\left(S_i^{-1}\mathcal{B}_i\mathcal{B}_i' \right) \cdot O\left(\delta_{\max}\left(\sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U \right) \right), \text{ by Lemma C.3-3} \\
&= \frac{K}{n}\sum_{i=1}^{n} \delta_{\max}\left(S_i^{-1}\mathcal{B}_i\mathcal{B}_i' \right) \cdot O_p\left(\sum_{l=1}^{r} \lambda_l \right), \text{ by Equation (B.3)} \\
&\leq \frac{K}{n}\sum_{i=1}^{n} \delta_{\max}(S_i^{-1}) \cdot \delta_{\max}(\mathcal{B}_i\mathcal{B}_i') \cdot O_p\left(\sum_{l=1}^{r} \lambda_l \right) \\
&= \frac{K}{n}\sum_{i=1}^{n} \delta_{\max}(\mathcal{B}_i\mathcal{B}_i') \cdot O_p\left(\sum_{l=1}^{r} \lambda_l \right), \text{ by Equation (B.2)} \\
&= \frac{K}{n}\sum_{i=1}^{n} (\mathcal{B}_i'\mathcal{B}_i) \cdot O_p\left(\sum_{l=1}^{r} \lambda_l \right) \\
&= O_p(K^2) \cdot O_p\left(\sum_{l=1}^{r} \lambda_l \right), \text{ by Lemma C.5} \\
&= O_p\left(\left(\sum_{l=1}^{r} \lambda_l \right) K^2 \right),
\end{aligned}
$$

where the seventh line follows by

$$
\frac{1}{n}\sum_{i=1}^{n} \delta_{\max}(\mathcal{B}_i\mathcal{B}_i') = \frac{1}{n}\sum_{i=1}^{n} \|\mathcal{B}_i\mathcal{B}_i'\| = \frac{1}{n}\sum_{i=1}^{n} (\mathcal{B}_i'\mathcal{B}_i) = O_p(K). \tag{B.6}
$$

For Ψ_n, we have

$$
\begin{aligned}
\Psi_n &= \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{B}_i'\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}T_{l,i}\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U]\right)^{-1}\mathcal{B}_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{B}_i'\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1}S_i^{-1}T_{l,i}\times\right. \\
&\qquad \left.\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1}S_i^{-1}\mathcal{B}_i\right],\ \text{by Equation (B.4)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}\left[\mathcal{B}_i'\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1}S_i^{-1}T_{l,i}\times\right. \\
&\qquad \left.\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1}S_i^{-1}\mathcal{B}_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}\left[S_i^{-1}\mathcal{B}_i\mathcal{B}_i'\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1}S_i^{-1}T_{l,i}\times\right. \\
&\qquad \left.\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1}\right] \\
&\le \frac{1}{n}K\sum_{i=1}^{n}\delta_{\max}\left(S_i^{-1}\mathcal{B}_i\mathcal{B}_i'\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1}S_i^{-1}T_{l,i}\times\right. \\
&\qquad \left.\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1}\right),\ \text{by Lemma C.2} \\
&\le \frac{K}{n}\sum_{i=1}^{n}\delta_{\max}(S_i^{-1}\mathcal{B}_i\mathcal{B}_i')\,\delta_{\max}(S_i^{-1}T_{l,i})\,\delta_{\max}^2\left(\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1}\right) \\
&= \frac{K}{n}\sum_{i=1}^{n}\delta_{\max}\left(S_i^{-1}\mathcal{B}_i\mathcal{B}_i'\right)\,O_p(1),\ \text{by Lemma C.3-1 and Equation (B.2)} \\
&\le O_p(K^2),
\end{aligned}
$$

where the last inequality is similar to the calculation of $\Theta_{2,n}$.

Finally, for Ξ_n we can see the proof should be similar to Ψ_n, which follows by noticing that by

Lemma C.5,

$$
\begin{aligned}
\Xi_n &= \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{B}_i'\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\frac{1}{n}\mathbf{B}'\mathbf{B}\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\mathcal{B}_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}\left[\mathcal{B}_i'\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\frac{1}{n}\mathbf{B}'\mathbf{B}\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\mathcal{B}_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}\left[\frac{1}{n}\sum_{m=1}^{n}\mathcal{B}_i\mathcal{B}_i'\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\mathcal{B}_i\mathcal{B}_i'\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\right] \\
&\leq \delta_{\max}\left(\frac{1}{n}\sum_{m=1}^{n}\mathcal{B}_i\mathcal{B}_i'\right)\frac{K}{n}\sum_{i=1}^{n}\delta_{\max}^2\left(\left(S_i+\sum_{l=1}^{r}\lambda_l T_{l,i}+O(||\lambda||^2)\,U\right)^{-1}\right)\delta_{\max}(\mathcal{B}_i\mathcal{B}_i') \\
&\leq O_p(1)\,\frac{K}{n}\sum_{i=1}^{n}\delta_{\max}^2\left(\left(I_k+\sum_{l=1}^{r}\lambda_l S_i^{-1}T_{l,i}+O(||\lambda||^2)\,S_i^{-1}U\right)^{-1}S_i^{-1}\right)\delta_{\max}(\mathcal{B}_i\mathcal{B}_i'),\ \text{by Lemma C.4} \\
&= O_p(K)\,\frac{1}{n}\sum_{i=1}^{n}\delta_{\max}(\mathcal{B}_i\mathcal{B}_i')\,O_p(1),\ \text{by Equation (B.2) and Lemma C.3} \\
&\leq O_p(K)O_p(K)O_p(1),\ \text{by Equation (B.6)} \\
&= O_p(K^2).
\end{aligned}
$$

Since $\lambda_l = o(1)$, then combining the above results we have

$$
\begin{aligned}
\delta_{\max}^2(P) \leq \operatorname{tr}(P'P) &\leq \Theta_n + \sum_{l=1}^{r}\lambda_l^2\Psi_n + \Xi_n\cdot O(\|\lambda\|^3) \\
&\leq O_p(K) + O_p\left(\left(\sum_{l=1}^{r}\lambda_l\right)K^2\right) + O_p\left(\left(\sum_{l=1}^{r}\lambda_l^2\right)K^2\right) + O_p\left(\|\lambda\|^3K^2\right) \\
&= O_p\left(K+\left(\sum_{l=1}^{r}\lambda_l\right)K^2\right).
\end{aligned}
$$

This completes the proof of Lemma 4.

□

Proof of Theorem 1. The proof is similar to that of Theorem 1 of [57]. First it is straightforward

to see that

$$\delta_{\max}(\Omega) = O(1). \tag{B.7}$$

Following Assumption 4, Assumption 7, and Lemma 4, we obtain

$$\begin{aligned}\delta^2(P(w_s^o)) = \delta^2(P_{(s)}) &= O_p\left(K_{(s)} + \sum_{l=1}^{r_s} \lambda_{l,(s)} \cdot K_{(s)}^2\right) \\ &= O_p(K_{(s)}) = O_p\left(n^{1/(1+2\alpha_s)}\right) \\ &\leq O_p\left(n^{1/(1+2\underline{\alpha})}\right).\end{aligned}$$

Therefore, we have

$$\delta(P(w_s^o))^{2N} \xi_n^{-2N} \sum_{t=1}^{S_n} [nR_n(w_t^o)]^N \to 0, \forall s, \tag{B.8}$$

as $n \to \infty$.

Let $A(w) = I - P(w)$. Note that

$$C_n(w) = L_n(w) + n^{-1}\|\epsilon\|^2 + 2n^{-1}\langle \epsilon, A(w)\mu\rangle + 2n^{-1}\left\{\mathrm{tr}[P(w)\Omega] - \langle \epsilon, P(w)\epsilon\rangle\right\}$$

Theorem 1 is valid if the following is true: as $n \to \infty$,

$$\sup_{w\in\mathcal{W}} |\langle \epsilon, A(w)\mu\rangle| / [nR_n(w)] \xrightarrow{p} 0, \tag{B.9}$$

$$\sup_{w\in\mathcal{W}} |\,\mathrm{tr}[P(w)\Omega] - \langle \epsilon, P(w)\epsilon\rangle| / [nR_n(w)] \xrightarrow{p} 0, \tag{B.10}$$

$$\sup_{w\in\mathcal{W}} |L_n(w)/R_n(w) - 1| \xrightarrow{p} 0. \tag{B.11}$$

First, we consider Equation (B.9). For any $\delta > 0$, by the triangle inequality, Chebyshev's inequal-

ity, Theorem 2 of [80], Equation (B.7), and Equation (B.8), we obtain

$$
\begin{aligned}
&\Pr\left\{\sup_{w\in\mathcal{W}}|\langle\epsilon, A(w)\mu\rangle|/[nR_n(w)] > \tau\right\} \\
&\quad\le \Pr\left\{\sup_{w\in\mathcal{W}}\sum_{s=1}^{S_n} w_s|\epsilon'(I-P_{(s)})\mu| > \tau\xi_n\right\} \\
&\quad\le \Pr\left\{\max_{1\le s\le S_n}|\epsilon'(I-P_{(s)})\mu| > \tau\xi_n\right\} \\
&\quad= \Pr\left\{\{|\langle\epsilon, A(w_1^o)\mu\rangle| > \tau\xi_n\}\bigcup\{|\langle\epsilon, A(w_2^o)\mu\rangle| > \tau\xi_n\}\bigcup\ldots\bigcup\left\{|\langle\epsilon, A(w_{S_n}^o)\mu\rangle| > \tau\xi_n\right\}\right\} \\
&\quad\le \sum_{s=1}^{S_n}\Pr\left\{|\langle\epsilon, A(w_s^o)\mu\rangle| > \tau\xi_n\right\} \text{ by the triangle inequality,} \\
&\quad\le \sum_{s=1}^{S_n} E\left\{\frac{\langle\epsilon, A(w_s^o)\mu\rangle^{2N}}{\tau^{2N}\xi_n^{2N}}\right\} \text{ by Chebyshev's inequality,} \\
&\quad\le C_1\tau^{-2N}\xi_n^{-2N}\sum_{s=1}^{S_n}\|\Omega(2N)^{1/2}A(w_s^o)\mu\|^{2N} \text{ by (7) in Theorem 2 of [80],}
\end{aligned}
$$

where C_1 is a constant, $\Omega(2N) = \mathrm{diag}(\gamma_1^2(2N), \ldots, \gamma_n^2(2N))$, and $\gamma_j(2N) = E(\epsilon_j^{2N}|X_j, Z_j)^{1/2N}$. By Assumption 6-1, $\gamma_j(2N) < \infty$, thus $\delta_{\max}(\Omega(2N))^N = O(1)$. In addition, notice that $\mu' A\mu \le \delta_{\max}(A)\mu'\mu$ and $\delta(AA) = \delta_{\max}(A)^2$ for any symmetric positive semi-definite matrix A, along with $nR_n(w_s^o) \ge \|A(w_s^o)\mu\|^2$, which is implied by Equation (3.6), we have

$$
\begin{aligned}
&\Pr\left\{\sup_{w\in\mathcal{W}}|\langle\epsilon, A(w)\mu\rangle|/[nR_n(w)] > \tau\right\} \\
&\quad\le C_1\tau^{-2N}\xi_n^{-2N}\delta_{\max}(\Omega(2N))^N\sum_{s=1}^{S_n}\|A(w_s^o)\mu\|^{2N} \\
&\quad\le C_1'\tau^{-2N}\xi_n^{-2N}\sum_{s=1}^{S_n}[nR_n(w_s^o)]^N \to 0 \quad \text{as } n\to\infty \text{ by Equation (B.7) and Equation (B.8).}
\end{aligned}
$$

Similarly for Equation (B.10), we have

$$
\begin{aligned}
&\Pr\left\{\sup_{w\in\mathcal{W}}\left|\mathrm{tr}[P(w)\Omega]-\langle\epsilon,P(w)\epsilon\rangle\right|/[nR_n(w)]>\tau\right\}\\
&=\Pr\left\{\sup_{w\in\mathcal{W}}\left|\sum_{s=1}^{S_n}w_s[\mathrm{tr}(P_{(s)}\Omega)-\langle\epsilon,P_{(s)}\epsilon\rangle]\right|/[nR_n(w)]>\tau\right\}\\
&\le\Pr\left\{\max_{1\le s\le S_n}\left|\mathrm{tr}(P_{(s)}\Omega)-\langle\epsilon,P_{(s)}\epsilon\rangle\right|/[nR_n(w)]>\tau\right\}\\
&\le\sum_{s=1}^{S_n}\Pr\left\{\left|\mathrm{tr}[P(w_s^o)\Omega]-\langle\epsilon,P(w_s^o)\epsilon\rangle\right|>\tau\xi_n\right\}\\
&\le\sum_{s=1}^{S_n}E\left\{\frac{[\mathrm{tr}[P(w_s^o)\Omega]-\langle\epsilon,P(w_s^o)\epsilon\rangle]^{2N}}{\tau^{2N}\xi_n^{2N}}\right\}\\
&\le C_2\tau^{-2N}\xi_n^{-2N}\sum_{s=1}^{S_n}\left\{\mathrm{tr}\left[P(w_s^o)'\Omega(4N)P(w_s^o)\right]\right\}^N\ \text{by (8) in Theorem 2 of [80],}
\end{aligned}
$$

where C_2 is a constant, $\Omega(4N)=\mathrm{diag}(\gamma_1^2(4N),\ldots,\gamma_n^2(4N))$, and $\gamma_j(4N)=E(\epsilon_j^{4N}|X_j,Z_j)^{1/4N}$. By Equation (3.6) and Assumption 6-2, we have $nR_n(w_s^o)\ge\mathrm{tr}\left[\Omega P(w_s^o)'P(w_s^o)\right]\ge\bar{\sigma}^2\,\mathrm{tr}[P(w_s^o)'P(w_s^o)]$. By Assumption 6-1, $\gamma_j(4N)<\infty$, thus $\delta_{\max}(\Omega(4N))^N=O(1)$. Along with Lemma C.1-3, we have

$$
\begin{aligned}
&\Pr\left\{\sup_{w\in\mathcal{W}}\left|\mathrm{tr}[P(w)\Omega]-\langle\epsilon,P(w)\epsilon\rangle\right|/[nR_n(w)]>\tau\right\}\\
&\le C_2\tau^{-2N}\xi_n^{-2N}\delta_{\max}(\Omega(4N))^N\sum_{s=1}^{S_n}\mathrm{tr}[P(w_s^o)'P(w_s^o)]\\
&\le C_2'\tau^{-2N}\xi_n^{-2N}\sum_{s=1}^{S_n}[nR_n(w_s^o)]^N\to 0\quad\text{as } n\to\infty.
\end{aligned}
$$

Note that Equation (B.11) is equivalent to

$$
\sup_{w\in\mathcal{W}}\left|\frac{n^{-1}\|P(w)\epsilon\|^2-n^{-1}\,\mathrm{tr}[\Omega P(w)'P(w)]-2n^{-1}\langle A(w)\mu,P(w)\epsilon\rangle}{R_n(w)}\right|\xrightarrow{p}0.
$$

Thus, Equation (B.11) holds if, as $n \to \infty$, we have

$$\sup_{w \in \mathcal{W}} \left| \frac{\langle A(w)\mu, P(w)\epsilon \rangle}{nR_n(w)} \right| \xrightarrow{p} 0, \tag{B.12}$$

and

$$\sup_{w \in \mathcal{W}} \left| \frac{\|P(w)\epsilon\|^2 - \mathrm{tr}[\Omega P(w)'P(w)]}{nR_n(w)} \right| \xrightarrow{p} 0. \tag{B.13}$$

For Equation (B.12), we have

$$\begin{aligned}
&\Pr\left\{ \sup_{w \in \mathcal{W}} \left| \frac{\langle A(w)\mu, P(w)\epsilon \rangle}{nR_n(w)} \right| > \tau \right\} \\
&\leq \Pr\left\{ \sup_{w \in \mathcal{W}} \sum_{m=1}^{S_n} \sum_{s=1}^{S_n} w_t w_s \left| \epsilon' P_{(s)} (I - P_{(m)}) \mu \right| > \tau \xi_n \right\} \\
&\leq \Pr\left\{ \max_{1 \leq m \leq S_n} \max_{1 \leq s \leq S_n} \left| \epsilon' P_{(s)} (I - P_{(m)}) \mu \right| > \tau \xi_n \right\} \\
&\leq \sum_{t=1}^{S_n} \sum_{s=1}^{S_n} E\left[\frac{\langle P(w_m^o)\epsilon, A(w_s^o)\mu \rangle^{2N}}{\tau^{2N} \xi_n^{2N}} \right] \\
&\leq C_3 \tau^{-2N} \xi_n^{-2N} \sum_{m=1}^{S_n} \sum_{s=1}^{S_n} \left\| P(w_m^o) \Omega (2N)^{1/2} A(w_s^o) \mu \right\|^{2N} \text{ by (7) in Theorem 2 of [80]}, \\
&\leq C_3 \left(\max_m \delta_{\max} [P(w_m^o) \Omega (2N)^{1/2}]^{2N} \right) \tau^{-2N} \xi_n^{-2N} \sum_{m=1}^{S_n} \sum_{s=1}^{S_n} \|A(w_s^o)\mu\|^{2N} \\
&\leq C_3' S_n \delta_{\max} (\Omega(2N))^N \left(\max_m \delta_{\max} [P(w_m^o)]^{2N} \right) \tau^{-2N} \xi_n^{-2N} \sum_{s=1}^{S_n} [nR_n(w_s^o)]^N \\
&\leq C_3' S_n \left(\max_m \delta_{\max} [P(w_m^o)]^{2N} \right) \tau^{-2N} \xi_n^{-2N} \sum_{s=1}^{S_n} [nR_n(w_s^o)]^N \to 0 \quad \text{as } n \to \infty,
\end{aligned}$$

where C_3, C_3' are constants, the fifth inequality follows from Lemma C.1-5, and the last line follows

from Equation (B.8). Also, for Equation (B.13)

$$
\begin{aligned}
&\Pr\left\{\sup_{w\in\mathcal{W}}\left|\frac{\|P(w)\epsilon\|^2-\mathrm{tr}[\Omega P(w)'P(w)]}{nR_n(w)}\right|>\tau\right\}\\
&\le \Pr\left\{\sup_{w\in\mathcal{W}}\sum_{t=1}^{S_n}\sum_{s=1}^{S_n} w_t w_s|\epsilon' P'_{(t)}P_{(s)}\epsilon-\mathrm{tr}[\Omega P'_{(s)}P_{(t)}]|>\tau\xi_n\right\}\\
&\le \Pr\left\{\max_{1\le t\le S_n}\max_{1\le s\le S_n}|\epsilon' P'_{(t)}P_{(s)}\epsilon-\mathrm{tr}[\Omega P'_{(s)}P_{(t)}]|>\tau\xi_n\right\}\\
&\le \sum_{t=1}^{S_n}\sum_{s=1}^{S_n} E\left\{\frac{[\langle\Omega^{-1/2}\epsilon,\Omega^{1/2}P(w_t^o)'P(w_s^o)\Omega^{1/2}\Omega^{-1/2}\epsilon\rangle-\mathrm{tr}(\Omega P(w_t^o)'P(w_s^o))]^{2N}}{\tau^{2N}\xi_n^{2N}}\right\}\\
&\le C_4\tau^{-2N}\xi_n^{-2N}\sum_{t=1}^{S_n}\sum_{s=1}^{S_n}\mathrm{tr}(P(w_t^o)'P(w_s^o)\Omega(4N)P(w_s^o)'P(w_t^o))^N\\
&= C_4\tau^{-2N}\xi_n^{-2N}\delta_{\max}(\Omega(4N))^N\sum_{t=1}^{S_n}\sum_{s=1}^{S_n}\mathrm{tr}(P(w_t^o)'P(w_s^o)P(w_s^o)'P(w_t^o))^N\\
&= C_4\tau^{-2N}\xi_n^{-2N}\delta_{\max}(\Omega(4N))^N\sum_{t=1}^{S_n}\sum_{s=1}^{S_n}\mathrm{tr}(P(w_t^o)P(w_t^o)'P(w_s^o)P(w_s^o)')^N\\
&\le C_4' S_n\left(\max_s\delta_{\max}[P(w_s^o)]^{2N}\right)\tau^{-2N}\xi_n^{-2N}\sum_{t=1}^{S_n}[nR_n(w_t^o)]^N\to 0\quad\text{as } n\to\infty,
\end{aligned}
$$

where C_4 and C_4' are constants. Thus we obtain Equation (B.12) and Equation (B.13) from Equation (B.8), which completes the proof. □

Lemma 5. *Under Assumption 5, for every* $s=1,2,\ldots,S_n$,

$$
\mathrm{tr}(P_{(s)})=O_p\left(K_{(s)}+\sum_{l=1}^{r_s}\lambda_{l,(s)}\cdot K_{(s)}^2\right).
$$

Proof of Lemma 5. We follow the notations in the proof of Lemma 4.

$$
\begin{aligned}
\operatorname{tr}(P) &= \sum_{i=1}^{n} \rho_{ii} \\
&= \sum_{i=1}^{n} \mathcal{B}_i' (\mathcal{B}' \mathcal{L}_{Z_i} \mathcal{B})^{-1} \mathcal{B}_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathcal{B}_i' \left(S_i + \sum_{l=1}^{r} \lambda_l T_{l,i} + O(||\lambda||^2)\, U \right)^{-1} \mathcal{B}_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathcal{B}_i' \left(I_k + \sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U \right)^{-1} S_i^{-1} \mathcal{B}_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \operatorname{tr} \left[S_i^{-1} \mathcal{B}_i \mathcal{B}_i' \left(I_k + \sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U \right)^{-1} \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \operatorname{tr} \left(S_i^{-1} \mathcal{B}_i \mathcal{B}_i' \right) + \frac{1}{n} \sum_{i=1}^{n} \operatorname{tr} \left[S_i^{-1} \mathcal{B}_i \mathcal{B}_i' \left(\left(I_k + \sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U \right)^{-1} - I_k \right) \right] \\
&\equiv \Phi_{1,n} + \Phi_{2,n}.
\end{aligned}
$$

From Equation (B.5), $\Phi_{1,n} = O(K)$, furthermore

$$
\begin{aligned}
\Phi_{2,n} &= \frac{1}{n} \sum_{i=1}^{n} \operatorname{tr} \left[S_i^{-1} \mathcal{B}_i \mathcal{B}_i' \left(\left(I_k + \sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U \right)^{-1} - I_k \right) \right] \\
&\leq \frac{K}{n} \sum_{i=1}^{n} \delta_{\max} \left[S_i^{-1} \mathcal{B}_i \mathcal{B}_i' \left(\left(I_k + \sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U \right)^{-1} - I_k \right) \right], \text{ by Lemma C.2} \\
&\leq \frac{K}{n} \sum_{i=1}^{n} \delta_{\max} \left(S_i^{-1} \mathcal{B}_i \mathcal{B}_i' \right) \delta_{\max} \left(\left(I_k + \sum_{l=1}^{r} \lambda_l S_i^{-1} T_{l,i} + O(||\lambda||^2)\, S_i^{-1} U \right)^{-1} - I_k \right) \\
&= \frac{K}{n} \sum_{i=1}^{n} \delta_{\max} \left(S_i^{-1} \mathcal{B}_i \mathcal{B}_i' \right) \cdot O \left(\sum_{l=1}^{r} \lambda_l \right), \text{ by Lemma C.3-2 and Equation (B.2)} \\
&= O \left(\left(\sum_{l=1}^{r} \lambda_l \right) K^2 \right),
\end{aligned}
$$

where the last inequality is similar to the proof of the rate of $\Theta_{2,n}$ in the proof of Lemma 4. Hence, $\operatorname{tr}(P) \leq \Phi_{1,n} + \Phi_{2,n} \leq O\left(K + \left(\sum_{l=1}^{r} \lambda_l\right) K^2\right)$. This completes the proof of Lemma 5. □

Proof of Theorem 2. The proof is similar to Theorem 1 in [69]. Let $\rho = \max_{1\le s\le S_n} \max_{1\le i\le n} \rho_{ii}^{(s)}$. Let $K_n = \max_{1\le s\le S_n} K_{(s)}$. By Lemma 5, Assumption 7 and Assumption 4, we know

$$\operatorname{tr}(P_{(s)}) = O\left(K_{(s)} + \sum_{l=1}^{r_s} \lambda_{l,(s)} \cdot K_{(s)}^2\right) = O(K_{(s)}). \tag{B.14}$$

Thus, by Assumption 8,

$$\rho = O(n^{-1}K_n). \tag{B.15}$$

In addition, by Assumption 7 and Assumption 4,

$$O(n^{-1}K_{(s)}^3) = O(n^{3/(1+2\alpha_s)-1}) = O(n^{(2-2\alpha_s)(1+2\alpha_s)}) = o(1). \tag{B.16}$$

Obviously,

$$\widehat{C}_n(w) = C_n(w) + 2n^{-1}\operatorname{tr}[P(w)\widehat{\Omega}(w)] - 2n^{-1}\operatorname{tr}[P(w)\Omega].$$

Therefore, Equation (3.10) holds if

$$\sup_{w\in\mathcal{W}} |\operatorname{tr}[P(w)\widehat{\Omega}(w)] - \operatorname{tr}[P(w)\Omega]|/[nR_n(w)] = o_p(1).$$

Moreover, by Equation (3.6), Assumption 6, and Lemma 4, we have $nR_n(w_s^o) \ge \operatorname{tr}(\Omega P_{(s)}' P_{(s)}) \ge \bar{\sigma}^2 \operatorname{tr}(P_{(s)}' P_{(s)}) = O(K_{(s)})$. Therefore, by Assumption 4, we have

$$\xi_n \to \infty. \tag{B.17}$$

Let $H_{(s)} = \mathrm{diag}(\rho_{11}^{(s)}, \ldots, \rho_{nn}^{(s)})$ and $H(w) = \sum_{s=1}^{S_n} w_s H_{(s)}$. Then we obtain that

$$
\begin{aligned}
&\sup_{w \in \mathcal{W}} |\operatorname{tr}[P(w)\widehat{\Omega}(w)] - \operatorname{tr}[P(w)\Omega|/[nR_n(w)] \\
&= \sup_{w \in \mathcal{W}} |[y - P(w)y]' H(w)[y - P(w)y] - \operatorname{tr}[H(w)\Omega]|/[nR_n(w)] \\
&= \sup_{w \in \mathcal{W}} |[\epsilon + \mu - P(w)y]' H(w)[\epsilon + \mu - P(w)y] - \operatorname{tr}[H(w)\Omega]|/[nR_n(w)] \\
&\leq \sup_{w \in \mathcal{W}} \frac{|\epsilon' H(w)\epsilon - \operatorname{tr}[H(w)\Omega]|}{[nR_n(w)]} + 2 \sup_{w \in \mathcal{W}} \frac{|\epsilon' H(w)[P(w)y - \mu]|}{[nR_n(w)]} \\
&\quad + \sup_{w \in \mathcal{W}} \frac{|[P(w)y - \mu]' H(w)[P(w)y - \mu]|}{[nR_n(w)]} \\
&\leq \sup_{w \in \mathcal{W}} \frac{|\epsilon' H(w)\epsilon - \operatorname{tr}[H(w)\Omega]|}{[nR_n(w)]} \\
&\quad + 2 \sup_{w \in \mathcal{W}} \frac{|\epsilon' H(w)[P(w)\mu - \mu]|}{[nR_n(w)]} \\
&\quad + 2 \sup_{w \in \mathcal{W}} \frac{|\epsilon' H(w)P(w)\epsilon - \operatorname{tr}[H(w)P(w)\Omega]|}{[nR_n(w)]} \\
&\quad + 2 \sup_{w \in \mathcal{W}} \frac{|\operatorname{tr}[H(w)P(w)\Omega]|}{[nR_n(w)]} \\
&\quad + \sup_{w \in \mathcal{W}} \frac{|[P(w)y - \mu]' H(w)[P(w)y - \mu]|}{[nR_n(w)]} \\
&\equiv D_1 + D_2 + D_3 + D_4 + D_5.
\end{aligned}
$$

For any $\tau > 0$,

$$
\begin{aligned}
&\Pr(D_1 > \tau) \leq Pr\left(\sup_{w\in\mathcal{W}} |\epsilon' H(w)\epsilon - \mathrm{tr}[H(w)\Omega]| > \tau\xi_n\right) \\
&\leq \Pr\left(\max_{1\leq s\leq S_n} |\epsilon' H(s)\epsilon - \mathrm{tr}[H(s)\Omega]| > \tau\xi_n\right) \\
&\leq \Pr\left(\{|\epsilon' H(w_1^o)\epsilon - \mathrm{tr}[H(w_1^o)\Omega]| > \tau\xi_n\}\bigcup\cdots\bigcup\{|\epsilon' H(w_{S_n}^o)\epsilon - \mathrm{tr}[H(w_{S_n}^o)\Omega]| > \tau\xi_n\}\right) \\
&\leq \sum_{s=1}^{S_n} \Pr\left(|\epsilon' H_{(s)}\epsilon - \mathrm{tr}[H_{(s)}\Omega]| > \tau\xi_n\right) \\
&\leq \tau^{-2N}\xi_n^{-2N}\sum_{s=1}^{S_n} E\left[|\epsilon' H_{(s)}\epsilon - \mathrm{tr}[H_{(s)}\Omega]|^{2N}\right] \\
&\leq C_5\tau^{-2N}\xi_n^{-2N}\sum_{s=1}^{S_n}\left(\mathrm{tr}(\Omega(4N)^{1/2}H_{(s)}\Omega(4N)H'_{(s)}\Omega(4N)^{1/2})\right)^N \text{ by [80]}, \\
&\leq C_5\tau^{-2N}\xi_n^{-2N}\delta_{\max}(\Omega(4N))^{2N}S_n\max_{1\leq s\leq S_n}\left(\mathrm{tr}(H_{(s)}^2)\right)^N \\
&\leq C_5'\tau^{-2N}\xi_n^{-2N}S_n O((n^{-1}K_n^2)^N) \\
&= \xi_n^{-2N}S_n O((n^{-1}K_n^2)^N) \text{ by Equation (B.17) and Equation (B.16)},
\end{aligned}
$$

where one can obtain the seventh inequality by recognizing that $\mathrm{tr}(H_{(s)}^2) \leq \mathrm{tr}(H_{(s)})\delta_{\max}(H_{(s)}) \leq \mathrm{tr}(P_{(s)})\rho$, and then using Equation (B.14) and Equation (B.15). C_5, C_5' are constants.

Similarly,

$$
\begin{aligned}
&\Pr(D_3/2 > \tau) = Pr\left(\sup_{w\in\mathcal{W}} \frac{|\epsilon' H(w)P(w)\epsilon - \mathrm{tr}[H(w)P(w)\Omega]|}{[nR_n(w)]} > \tau\right) \\
&\leq \sum_{s=1}^{S_n} \Pr\left(|\epsilon' H(s)P(s)\epsilon - \mathrm{tr}[H(s)P(s)\Omega]| > \tau\xi_n\right) \\
&\leq \tau^{-2N}\xi_n^{-2N} \sum_{s=1}^{S_n} E\left[|\epsilon' H(s)P(s)\epsilon - \mathrm{tr}[H(s)P(s)\Omega]|^{2N}\right] \\
&\leq C_6\tau^{-2N}\xi_n^{-2N} \sum_{s=1}^{S_n} \left(\mathrm{tr}(\Omega(4N)^{1/2}H(s)P(s)\Omega(4N)P(s)'H(s)'\Omega(4N)^{1/2})\right)^N \text{ by [80]}, \\
&\leq C_6\tau^{-2N}\xi_n^{-2N}\delta_{\max}(\Omega(4N))^{2N} S_n \max_{1\leq s\leq S_n} \delta_{\max}(P(s)P(s)')^N \max_{1\leq s\leq S_n} \left(\mathrm{tr}(H(s)^2)\right)^N \\
&\leq C_6'\tau^{-2N}\xi_n^{-2N} S_n O(K_n^N) O((n^{-1}K_n^2)^N) \\
&= \xi_n^{-2N} S_n O(n^{-1}K_n^3))^N = o(1) \text{ by Equation (B.17) and Equation (B.16)},
\end{aligned}
$$

where the last inequality applies the result $\mathrm{tr}(H_{(s)}^2) = O(n^{-1}K_n)$ from the proof of D_1, along with $\delta_{\max}(P(s)P(s)') \leq \delta_{\max}^2(P(s)) = O(K_{(s)})$ by Lemma 4. C_6, C_6' are constants.

From Equation (3.6) we have $nR_n(w) \geq \|P(w)\mu - \mu\|^2$. Along with the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
D_2 &\leq 2 \sup_{w\in\mathcal{W}} \left(\|\epsilon\|^2\|H(w)\|^2\|P(w)\mu - \mu\|^2/(nR_n(w))^2\right)^{1/2} \\
&\leq \|\epsilon\|\xi_n^{-1/2} \max_{1\leq s\leq S_n} \|H_{(s)}\| \\
&= \|\epsilon\|\xi_n^{-1/2}\rho \\
&= O(n^{-1/2})\xi_n^{-1/2}O(n^{-1}K_n) \text{ by Equation (B.15)}, \\
&= o(1) \text{ by Equation (B.17) and Equation (B.16)}.
\end{aligned}
$$

Next,

$$
\begin{aligned}
D_4 &\leq 2\xi_n^{-1} \max_{1\leq s\leq S_n} \delta_{\max}(H_{(s)})\delta_{\max}(\Omega)\operatorname{tr}(P_{(s)}) \text{ by Lemma C.1-3,} \\
&\leq 2\xi_n^{-1} \max_{1\leq s\leq S_n} \rho O(1)\operatorname{tr}(P_{(s)}) \\
&\leq \xi_n^{-1} O(n^{-1}K_n) \max_s \operatorname{tr}(P_{(s)}) \text{ by Equation (B.15),} \\
&= \xi_n^{-1} O(n^{-1}K_n^2) \text{ by Lemma 5,} \\
&= o(1) \text{ by Equation (B.17) and Equation (B.16).}
\end{aligned}
$$

Finally,

$$
\begin{aligned}
D_5 &\leq \rho \sup_{w\in\mathcal{W}} [(P(w)y-\mu)'(P(w)y-\mu)/(nR_n(w))] \\
&= \rho \sup_{w\in\mathcal{W}} [L_n(w)/R_n(w)] \\
&= O(n^{-1}K_n)O(1) = o(1),
\end{aligned}
$$

where the second equality comes from Equation (B.11) and Equation (B.15). □

Proof of Theorem 3. The proof is similar to Theorem 2 in [59]. First observe that

$$
\widehat{C}_n^*(w) = C_n(w) + 2n^{-1}\operatorname{tr}[P(w)\widehat{\Omega}^*] - 2n^{-1}\operatorname{tr}[P(w)\Omega].
$$

Thus, Equation (3.12) holds if

$$
\sup_{w\in\mathcal{W}} |\operatorname{tr}[P(w)\widehat{\Omega}^*] - \operatorname{tr}[P(w)\Omega]|/[nR_n(w)] = o_p(1).
$$

Recall $H_{(s)}$ and $H(w)$ defined in the proof of Theorem 2. We have

$$
\begin{aligned}
&\sup_{w\in\mathcal{W}} |\operatorname{tr}[P(w)\widehat{\Omega}^*] - \operatorname{tr}[P(w)\Omega|/[nR_n(w)] \\
&= \sup_{w\in\mathcal{W}} |[y - P_{(s^*)}y]'H(w)[y - P_{(s^*)}y] - \operatorname{tr}[H(w)\Omega]|/[nR_n(w)] \\
&= \sup_{w\in\mathcal{W}} |[\epsilon + \mu - P_{(s^*)}\mu - P_{(s^*)}\epsilon]'H(w)[\epsilon + \mu - P_{(s^*)}\mu - P_{(s^*)}\epsilon] - \operatorname{tr}[H(w)\Omega]|/[nR_n(w)] \\
&\le \sup_{w\in\mathcal{W}} \frac{|\epsilon'(I_n - P_{(s^*)})'H(w)(I_n - P_{(s^*)}\epsilon) - \operatorname{tr}[(I_n - P_{(s^*)})'H(w)(I_n - P_{(s^*)})\Omega]|}{[nR_n(w)]} \\
&\quad + 2\sup_{w\in\mathcal{W}} \frac{|\epsilon'(I_n - P_{(s^*)})'H(w)(I_n - P_{(s^*)})\mu|}{[nR_n(w)]} \\
&\quad + \sup_{w\in\mathcal{W}} \frac{|\mu'(I_n - P_{(s^*)})'H(w)(I_n - P_{(s^*)})\mu|}{[nR_n(w)]} \\
&\quad + \sup_{w\in\mathcal{W}} \frac{|\operatorname{tr}[P'_{(s^*)}H(w)P_{(s^*)}\Omega]|}{[nR_n(w)]} \\
&\quad + 2\sup_{w\in\mathcal{W}} \frac{|P'_{(s^*)}H(w)\Omega|}{[nR_n(w)]} \\
&\equiv \tilde{D}_1 + \tilde{D}_2 + \tilde{D}_3 + \tilde{D}_4 + \tilde{D}_5.
\end{aligned}
$$

We borrow the notation of $A(w)$ from the proof of Theorem 1, and ρ, K_n in the proof of

Theorem 2. Let $B(w) = (I_n - P_{(s^*)})'H(w)(I_n - P_{(s^*)}) = A(w^o_{(s^*)})'H(w)A(w^o_{(s^*)})$, then

$$
\begin{aligned}
\Pr(\tilde{D}_1 > \tau) &\leq \Pr\left(\sup_{w\in\mathcal{W}} |\epsilon' B(w)\epsilon - \mathrm{tr}[B(w)\Omega]| > \xi_n\tau\right) \\
&\leq \tau^{-2N}\xi_n^{-2N}\sum_{s=1}^{S_n} E\left[|\epsilon' B(w)\epsilon - \mathrm{tr}[B(w)\Omega]|^{2N}\right] \\
&\leq C_7\tau^{-2N}\xi_n^{-2N}\sum_{s=1}^{S_n}\left[\mathrm{tr}\left(\Omega(4N)^{1/2}B(w)'\Omega(4N)B(w)\Omega(4N)^{1/2}\right)\right]^N \\
&\leq C_7\tau^{-2N}\xi_n^{-2N}\delta_{\max}^{2N}(\Omega(4N))\sum_{s=1}^{S_n}\left[\mathrm{tr}(B(w)'B(w))\right]^N \\
&\leq C_7'\tau^{-2N}\xi_n^{-2N}S_n\rho^{2N}\left[\mathrm{tr}(A(w^o_{(s^*)})'A(w^o_{(s^*)})A(w^o_{(s^*)})'A(w^o_{(s^*)}))\right]^N \\
&\leq C_7'\tau^{-2N}\xi_n^{-2N}S_nO(n^{-1}K_n)^{2N}\left[\delta_{\max}\left(A(w^o_{(s^*)})'A(w^o_{(s^*)})\right)\mathrm{tr}\left(A(w^o_{(s^*)})'A(w^o_{(s^*)})\right)\right]^N \\
&\leq C_7'\tau^{-2N}\xi_n^{-2N}S_nO(n^{-1}K_n)^{2N}\left[O(K_{(s^*)})\,\mathrm{tr}\left(A(w^o_{(s^*)})'A(w^o_{(s^*)})\right)\right]^N \\
&\leq C_7'\xi_n^{-2N}O(n^{-1}K_n)^{2N}O(K_n)^N S_n\left(\mathrm{tr}(I_n - P_{(s^*)} - P'_{(s^*)} + P_{(s^*)}P'_{(s^*)})\right)^N \\
&\leq C_7'\xi_n^{-2N}O(n^{-2}K_n^3)^N S_n(n + O(K_{(s^*)}))^N \\
&= C_7'\xi_n^{-2N}S_nO(n^{-2}K_n^3)^N n^N \\
&= \xi_n^{-2N}S_nO(n^{-1}K_n^3)^N = o(1) \text{ by Equation (B.17) and Equation (B.16),}
\end{aligned}
$$

where the sixth inequality comes from Equation (B.15) and Lemma C.1-3, the seventh inequality is by Lemma 4, and the ninth inequality comes from Lemma 5 and Lemma 5, and $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B) \leq |\,\mathrm{tr}(A)| + |\,\mathrm{tr}(B)|$. C_7, C_7' are constants.

By the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\tilde{D}_2 &\leq 2\xi_n^{-1}\|I_n - P_{(s^*)}\mu\| \sup_{w\in\mathcal{W}}\|H(w)(I_n - P_{(s^*)})\epsilon\| \\
&\leq 2\xi_n^{-1}\|I_n - P_{(s^*)}\mu\| \cdot \rho(1 + \delta_{\max}(P_{(s^*)}))\|\epsilon\| \\
&\leq 2\xi_n^{-1}(nR_n(w))^{1/2}O(n^{-1}K_n)O(K_n^{1/2})O(n^{1/2}) \\
&\leq \xi_n^{-1/2}O(n^{-1}K_n^3)^{1/2} = o(1) \text{ by Equation (B.17) and Equation (B.16),}
\end{aligned}
$$

where the second inequality is by Lemma C.1-5, and the third inequality comes from Equation (3.6), Equation (B.15) and Lemma 4. Similarly,

$$\begin{aligned}
\tilde{D}_3 &\leq \xi_n^{-1}\rho\|(I_n - P_{(s^*)})\mu\|^2 \\
&\leq \xi_n^{-1}\rho\|(I_n - P_{(s^*)})\mu\|(1 + \delta_{\max}(P_{(s^*)}))\|\mu\| \\
&\leq \xi_n^{-1/2}O(n^{-1}K_n)O(K_n^{1/2})O(n^{1/2}) \\
&= o(n^{-1}K_n^3)^{1/2} = o(1),
\end{aligned}$$

where we use Equation (3.6), Equation (B.15), Lemma 4, and Assumption 9 in the third inequality. Next,

$$\begin{aligned}
\tilde{D}_4 &\leq \xi_n^{-1}\delta_{\max}(\Omega)\rho \cdot \mathrm{tr}(P'_{(s^*)}P_{(s)}) \\
&\leq \xi_n^{-1}O(n^{-1}K_n)O(K_n), \text{ by Equation (B.15) and Lemma 4} \\
&= o(n^{-1}K_n^2) = o(1).
\end{aligned}$$

Finally,

$$\begin{aligned}
\tilde{D}_5 &\leq 2\xi_n^{-1}\delta_{\max}(\Omega)\rho \cdot \mathrm{tr}(P_{(s^*)}) \\
&\leq \xi_n^{-1}O(n^{-1}K_n)O(K_n), \text{ by Equation (B.15) and Lemma 4} \\
&= o(n^{-1}K_n^2) = o(1).
\end{aligned}$$

□

B.2 Additional Simulation Results

We summarize the mean model average weights for Case (I) and Case (II) outlined in Section 3.3.

Table B.1: Case (I) MMA Weight Summary (Mean).

n	σ	w_1	w_2	w_3	w_4	w_5	w_6
50	0.25	0.01	0.02	0.10	0.11	0.11	0.66
	0.50	0.02	0.03	0.12	0.12	0.12	0.59
	1.00	0.05	0.06	0.16	0.14	0.14	0.45
	2.00	0.09	0.10	0.14	0.16	0.16	0.35
100	0.25	0.00	0.00	0.09	0.08	0.08	0.75
	0.50	0.01	0.01	0.10	0.10	0.10	0.68
	1.00	0.02	0.02	0.14	0.14	0.15	0.53
	2.00	0.07	0.07	0.16	0.17	0.17	0.36
200	0.25	0.00	0.00	0.08	0.03	0.03	0.86
	0.50	0.00	0.00	0.10	0.05	0.05	0.79
	1.00	0.01	0.01	0.12	0.11	0.11	0.64
	2.00	0.04	0.04	0.16	0.17	0.17	0.42
400	0.25	0.00	0.00	0.05	0.02	0.02	0.92
	0.50	0.00	0.00	0.08	0.03	0.03	0.86
	1.00	0.01	0.01	0.12	0.06	0.06	0.75
	2.00	0.02	0.02	0.16	0.14	0.14	0.53

Table B.2: Case (II) MMA Weight Summary (Mean).

n	σ	w_1	w_2	w_3	w_4	w_5	w_6
50	0.25	0.01	0.01	0.09	0.09	0.09	0.70
	0.50	0.02	0.02	0.11	0.11	0.11	0.63
	1.00	0.05	0.05	0.16	0.14	0.13	0.46
	2.00	0.09	0.09	0.16	0.17	0.16	0.33
100	0.25	0.00	0.00	0.06	0.05	0.05	0.83
	0.50	0.00	0.00	0.08	0.09	0.09	0.73
	1.00	0.02	0.02	0.13	0.14	0.14	0.56
	2.00	0.07	0.07	0.17	0.17	0.17	0.36
200	0.25	0.00	0.00	0.04	0.02	0.02	0.92
	0.50	0.00	0.00	0.09	0.04	0.04	0.83
	1.00	0.01	0.01	0.13	0.10	0.10	0.66
	2.00	0.04	0.04	0.17	0.17	0.16	0.43
400	0.25	0.00	0.00	0.02	0.01	0.01	0.97
	0.50	0.00	0.00	0.06	0.02	0.02	0.90
	1.00	0.00	0.01	0.12	0.05	0.05	0.77
	2.00	0.02	0.02	0.15	0.13	0.13	0.55

B.3 Supplementary Materials

Lemma C.1. *For two matrices A and B, we have*

1. $\delta_{\max}(AB) \leq \delta_{\max}(A)\delta_{\max}(B)$,
2. $\delta_{\max}(A+B) \leq \delta_{\max}(A) + \delta_{\max}(B)$,
3. *for any symmetric matrix A and positive semi-definite matrix B,* $\mathrm{tr}(AB) \leq \delta_{\max}(A)\,\mathrm{tr}(B)$,
4. $\delta_{\max}(A) \leq \|A\|$,
5. $\|Ax\| \leq \delta_{\max}(A)\|x\|$, *where x is a $k \times 1$ vector, and $\|x\|$ denotes the Euclidean norm for a vector.*

Proof: See [81], page 288-290.

Lemma C.2. *For a $k \times k$ matrix A,* $\mathrm{tr}(A) \leq k \cdot \delta_{\max}(A)$.

Proof of Lemma C.2.

$$\mathrm{tr}(A) = \sum_{i=1}^{k} \nu_i(A) \leq \sum_{i=1}^{k} \delta_i(A) \leq k \cdot \delta_{\max}(A), \tag{C.1}$$

where the first inequality is from [81], page 291. □

Lemma C.3. *For a $k \times k$ matrix A with $\delta_{\max}(A) = o(1)$, we have*

1. $\delta_{\max}((I_k + A)^{-1}) = O(1)$,
2. $\delta_{\max}((I_k + A)^{-1} - I_k) = O(\delta_{\max}(A))$,
3. $\delta_{\max}((I_k + A)^{-2} - I_k) = O(\delta_{\max}(A))$.

Proof of Lemma C.3. By the result of the Neumann series, for a $k \times k$ matrix A, if $\delta_{\max}(A) < 1$, then $(I_k + A)$ is invertible and

$$(I_k + A)^{-1} = \sum_{j=0}^{\infty} (-1)^j A^j. \tag{C.2}$$

With Lemma C.1-1, Equation (C.2), and $\delta_{\max}(A) = o(1) < 1$, for Lemma C.3-1 we have

$$\begin{aligned}
\delta_{\max}((I_k + A)^{-1}) &= \delta_{\max}\left(\sum_{j=0}^{\infty}(-1)^j A^j\right) \\
&\leq \sum_{j=1}^{\infty}\delta_{\max}(A^j) \\
&\leq \sum_{j=1}^{\infty}\delta_{\max}(A)^j \\
&= \frac{1}{1-\delta_{\max}(A)} \\
&= O(1).
\end{aligned}$$

For Lemma C.3-2, we first observe that

$$\sum_{j=0}^{\infty}(-1)^j A^j - I_k = -A + A^2 - A^3 + \cdots$$

Then, similar to the proof of Lemma C.3-1,

$$\begin{aligned}
\delta_{\max}((I_k + A)^{-1} - I_k) &= \delta_{\max}\left[\left(\sum_{j=0}^{\infty}(-1)^j A^j\right) - I_k\right] \\
&= \delta_{\max}\left(\sum_{j=1}^{\infty}(-1)^j A^j\right) \\
&\leq \sum_{j=1}^{\infty}\left(\delta_{\max}(A)^j\right) \\
&= \frac{\delta_{\max}(A)}{1-\delta_{\max}(A)} \\
&= O(\delta_{\max}(A)).
\end{aligned}$$

For Lemma C.3-3, notice that,

$$\begin{aligned}\left(\sum_{j=0}^{\infty}(-1)^j A^j\right)^2 - I_k &= (I_k - A + A^2 - A^3 + \cdots)^2 - I_k \\ &= (I_k - 2A + 3A - 4A + \cdots) - I_k \\ &= \sum_{j=1}^{\infty}(-1)^j (j+1) A^j.\end{aligned}$$

Hence,

$$\begin{aligned}\delta_{\max}((I_k + A)^{-2} - I_k) &= \delta_{\max}\left(\left(\sum_{j=0}^{\infty}(-1)^j A^j\right)^2 - I_k\right) \\ &= \delta_{\max}\left(\sum_{j=1}^{\infty}(-1)^j (j+1) A^j\right) \\ &\leq \sum_{j=1}^{\infty}\delta_{\max}\left((-1)^j (j+1) A^j\right) \\ &\leq \sum_{j=1}^{\infty}(j+1)\left(\delta_{\max}(A)^j\right) \\ &= \frac{\delta_{\max}(A)(2 - \delta_{\max}(A))}{(1 - \delta_{\max}(A))^2} \\ &= O(\delta_{\max}(A)).\end{aligned}$$

□

In the next two lemmas, we leave out the subscript (s) for notational ease.

Lemma C.4. *Let $S_z = n^{-1}\sum_{i=1}^{n}\mathcal{B}_i\mathcal{B}_i'\,\mathbf{1}(Z_i = z)$. Under Assumption 5, for any $z \in \mathcal{M}_Z$,*

$$0 < \underline{\eta}_z \leq \delta_{\min}(S_z) \leq \delta_{\max}(S_z) \leq \overline{\eta}_z < \infty$$

in probability.

Proof of Lemma C.4. Let $\widehat{\mathcal{B}}(X, \widehat{z}; z) = Q_z^{-1/2}\mathcal{B}(x)\,\mathbf{1}(\widehat{z} = z)$, and $\widehat{\mathcal{B}}_{i,z} = \widehat{\mathcal{B}}(X_i, Z_i; z) = Q_z^{-1/2}\mathcal{B}_i\,\mathbf{1}(Z_i =$

z). Since the matrices considered in this proof are all symmetric and positive semi-definite, their eigenvalues and singular values coincide. Therefore,

$$\delta_{\max}(Q_z^{-1/2}) = (\delta_{\max}(Q_z^{-1}))^{1/2} = (\delta_{\min}(Q_z))^{-1/2} = \underline{\eta}_z^{-1/2}.$$

Then by Assumption 5-1, $E[\widehat{\mathcal{B}}_{i,z}\widehat{\mathcal{B}}'_{i,z}] = I_k$. In addition, by Lemma C.1-5 and Assumption 5-2,

$$\begin{aligned}
\sup_{X\in\mathcal{M}_X, Z\in\mathcal{M}_Z} \|\widehat{\mathcal{B}}(X,Z;z)\| &\leq \delta_{\max}(Q_z^{-1/2}) \sup_{X\in\mathcal{M}_X, Z\in\mathcal{M}_Z} \|\mathcal{B}(x)\,\mathbf{1}(Z=z)\| \\
&\leq \underline{\eta}_z^{-1/2} \sup_{X\in\mathcal{M}_X} \|\mathcal{B}(x)\| \\
&\leq \underline{\eta}_z^{-1/2}\zeta_0(K) \equiv \zeta_z(K).
\end{aligned}$$

Letting $R_z = n^{-1}\sum_{i=1}^n \tilde{\mathcal{B}}_{i,z}\widehat{\mathcal{B}}'_{i,z}$, similar to (A.1) in [67] we have

$$\begin{aligned}
E\left[\|R_z - I_k\|^2\right] &= \sum_{l=1}^K \sum_{j=1}^K E\left[\left(\sum_{i=1}^n \widehat{\mathcal{B}}_{i,z,l}\widehat{\mathcal{B}}_{i,z,j}/n - I_{jl}\right)^2\right] \\
&= \sum_{l=1}^K \sum_{j=1}^K E\left[\left(\sum_{i=1}^n \widehat{\mathcal{B}}^2_{i,z,l}\widehat{\mathcal{B}}^2_{i,z,j}\right)\right]/n \\
&= E\left[\sum_{l=1}^K \widehat{\mathcal{B}}^2_{i,z,l} \sum_{j=1}^K \widehat{\mathcal{B}}^2_{i,z,j}\right]/n \\
&\leq \sup_{X,Z} \|\widehat{\mathcal{B}}(X,Z;z)\|^2 \, E\left[\sum_{l=1}^K \widehat{\mathcal{B}}^2_{i,z,l}\right]/n \\
&\leq \zeta_z(K)^2 K/n \to 0.
\end{aligned}$$

Hence, $\|R_z - I_k\| = o_p(1)$. And using $0 \leq \delta_{\min}(A) \leq \delta_{\max}(A) \leq \|A\|$ for any matrix A, we have $\delta_{\min}(R_z) \xrightarrow{p} 1$, and $\delta_{\max}(R_z) \xrightarrow{p} 1$. Observing that $R_z = Q_z^{-1}S_z$, with Lemma C.1-(i),

$$\delta_{\max}(S_z) \leq \delta_{\max}(Q_z)\,\delta_{\max}(R_z) = \overline{\eta}_z\delta_{\max}(R_z) \xrightarrow{p} \overline{\eta}_z < \infty.$$

$$\delta_{\min}(S_z) \geq \delta_{\min}(Q_z)\,\delta_{\min}(R_z) = \underline{\eta}_z \delta_{\min}(R_z) \xrightarrow{p} \underline{\eta}_z > 0.$$

□

Lemma C.5. *Under Assumption 5-1 and 5-2,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_i'\mathcal{B}_i = O_p(K).$$

Proof of Lemma C.5. With Assumptions 5-1 and 5-2, by the result in [67], we have $\|\frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_i\mathcal{B}_i' - I_K\| = o_p(1)$. In addition, for a $k \times k$ matrix A,

$$\operatorname{tr}(A) \leq |\operatorname{tr}(A)| \leq \sum_{i=1}^{k} |a_{ii}| \leq \left(\sum_{i=1}^{k} |a_{ii}|^2\right)^{1/2} \cdot \sqrt{k} \leq \|A\| \cdot \sqrt{k}.$$

Therefore,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_i'\mathcal{B}_i &= \operatorname{tr}\left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_i\mathcal{B}_i'\right) \\
&= \operatorname{tr}\left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_i\mathcal{B}_i' - I_K\right) + \operatorname{tr}(I_K) \\
&\leq \left|\operatorname{tr}\left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_i\mathcal{B}_i' - I_K\right)\right| + K \\
&\leq K + o_p(1)\cdot\sqrt{K} = O_p(K). \qquad \text{(C.3)}
\end{aligned}$$

□

APPENDIX C

APPENDIX FOR THE THIRD ESSAY

C.1 Proofs of Main Theorems

C.1.1 Proof of Theorem 4

We first provide proofs for the first two variations ($j = 1, 2$) for each type (KS, CvM, AD) estimators. Notice that for a given w, the univariate KS, CvM, AD type objective functions, are all norms of $J_{G,i}(r, w)$ with respect to $r \in [0, 1]$. In particular, the KS objective function is the sup-norm $\|\cdot\|_\infty$, the CvM objective function is the L_2-norm $\|\cdot\|_2$, and the AD objective function is an adjusted L_2-norm $\|\cdot\|_{2,AD}$ (for a proof, see Appendix C.2). Therefore, our PIT-based objective functions can be written as

$$F_G^{(1)}(w) = \max_i \|J_{G,i}\|_F, \;\; F_G^{(2)}(w) = \|J_{G,1}\|_F + \|J_{G,2}\|_F, \tag{C.1}$$

where F_G is either K_G, C_G, or A_G, and the norm $\|\cdot\|_F$ is the corresponding norm for each type. Recall that the population counterpart of $J_{G,i}(r, w)$ is $J_{0,i}(w) = G^{-1}\sum_{t=1}^{G} E[V_{ti}(r, w)]$. The population counterparts of the objective functions are

$$F_0^{(1)}(w) = \max_i \|J_{0,i}\|_F, \;\; F_0^{(2)}(w) = \|J_{0,1}\|_F + \|J_{0,2}\|_F. \tag{C.2}$$

Following Appendix A in [75], under Assumption 10 and 11 we have

$$\sup_{w\in\mathcal{W}} \left| \|J_{G,i}\|_F - \|J_{0,i}\|_F \right| \xrightarrow{a.s.} 0, i = 1, 2. \tag{C.3}$$

Then, for the first variation ($j = 1$) of each type (F=KS,CvM,AD), using the triangular in-

equality of norm and the switch of ordering in supremum, we have

$$\begin{aligned}
\sup_{w\in\mathcal{W}}\left|F_G^{(1)}(w)-F_0^{(1)}(w)\right| &= \sup_{w\in\mathcal{W}}\left|\max_i\|J_{G,i}\|_F-\max_i\|J_{0,i}\|_F\right| \\
&\leq \sup_{w\in\mathcal{W}}\max_i\left|\|J_{G,i}\|_F-\|J_{0,i}\|_F\right| \\
&= \max_i\sup_{w\in\mathcal{W}}\left|\|J_{G,i}\|_F-\|J_{0,i}\|_F|\right| \\
&\xrightarrow{a.s.} 0. \qquad \text{(C.4)}
\end{aligned}$$

For the second variation ($j=2$), using the triangular inequality of norm and supremum, we have

$$\begin{aligned}
\sup_{w\in\mathcal{W}}\left|F_G^{(2)}(w)-F_0^{(2)}(w)\right| &= \sup_{w\in\mathcal{W}}\left|\|J_{G,1}\|_F+\|J_{G,2}\|_F-\|J_{0,1}\|_F-\|J_{0,2}\|_F\right| \\
&\leq \sup_{w\in\mathcal{W}}\left(\left|\|J_{G,1}\|_F-\|J_{0,1}\|_F\right|+\left|\|J_{G,2}\|_F-\|J_{0,2}\|_F\right|\right) \\
&\leq \sup_{w\in\mathcal{W}}\left|\|J_{G,1}\|_F-\|J_{0,1}\|_F\right|+\sup_{w\in\mathcal{W}}\left|\|J_{G,2}\|_F-\|J_{0,2}\|_F\right| \\
&\xrightarrow{a.s.} 0. \qquad \text{(C.5)}
\end{aligned}$$

Following the argument in [75], results (C.4), (C.5) along with Assumption 10-4 lead to $\widehat{w}\xrightarrow{a.s.}w^*$, where w^* is the unique minimzier of $F_0^{(j)}, j=1,2$.

Now we turn to prove the third variation ($j=3$). Note that $F_G^{(3)}(w)=\|J_{G,1}+J_{G,2}\|_F$, and

$F_0^{(3)}(w) = \|J_{0,1} + J_{0,2}\|_F$. Using the triangular inequality of norm, we have

$$\begin{aligned}
\sup_{w\in\mathcal{W}}\left|F_G^{(3)}(w) - F_0^{(3)}(w)\right| &= \sup_{w\in\mathcal{W}}\left|\|J_{G,1} + J_{G,2}\|_F - \|J_{0,1} + J_{0,2}\|_F\right| \\
&\leq \sup_{w\in\mathcal{W}}\left|\|J_{G,1} + J_{G,2} - J_{0,1} - J_{0,2}\|_F\right| \\
&\leq \sup_{w\in\mathcal{W}}\left|\|J_{G,1} - J_{0,1}\|_F + \|J_{G,2} - J_{0,2}\|_F\right| \\
&\leq \sup_{w\in\mathcal{W}}\left|\|J_{G,1}\|_F - \|J_{0,1}\|_F\right| + \sup_{w\in\mathcal{W}}\left|\|J_{G,2}\|_F - \|J_{0,2}\|_F\right| \\
&\xrightarrow{a.s.} 0. \qquad \text{(C.6)}
\end{aligned}$$

C.1.2 Proof of Theorem 5

Following the proof of Theorem 2 in [75], under Assumption 11 and 12 we have

$$\sup_{w\in\mathcal{W}}\left|\,\mathrm{KLIC}_{G,j}(w) - \mathrm{KLIC}_{0,j}(w)\right| \xrightarrow{a.s.} 0,\ \ j = 1, 2, 3. \qquad \text{(C.7)}$$

The result of $j = 3$ is immediately following the proof in [75]. For $j = 1$, using triangular inequality and the switch of ordering of supremum, we have we have

$$\begin{aligned}
\sup_{w\in\mathcal{W}}\left|H_G^{(1)}(w) - H_0^{(1)}(w)\right| &= \sup_{w\in\mathcal{W}}\left|\,\max_i \mathrm{KLIC}_{G,i}(w) - \max_i \mathrm{KLIC}_{0,i}(w)\right| \\
&\leq \sup_{w\in\mathcal{W}}\max_i\left|\,\mathrm{KLIC}_{G,i}(w) - \mathrm{KLIC}_{0,i}(w)\right| \\
&= \max_i \sup_{w\in\mathcal{W}}\left|\,\mathrm{KLIC}_{G,i}(w) - \mathrm{KLIC}_{0,i}(w)\right| \\
&\xrightarrow{a.s.} 0. \qquad \text{(C.8)}
\end{aligned}$$

For $j = 2$, using triangular inequality, we have

$$\begin{aligned}
\sup_{w\in\mathcal{W}} \left|H_G^{(2)}(w) - H_0^{(2)}(w)\right| &= \sup_{w\in\mathcal{W}} \left| \mathrm{KLIC}_{G,1}(w) + \mathrm{KLIC}_{G,2}(w) - \mathrm{KLIC}_{0,1}(w) - \mathrm{KLIC}_{0,2}(w)\right| \\
&\leq \sup_{w\in\mathcal{W}} \Big(\left| \mathrm{KLIC}_{G,1}(w) - \mathrm{KLIC}_{0,1}(w)\right| \\
&+ \left| \mathrm{KLIC}_{G,2}(w) - \mathrm{KLIC}_{0,2}(w)\right| \Big) \\
&\leq \sup_{w\in\mathcal{W}} \left| \mathrm{KLIC}_{G,1}(w) - \mathrm{KLIC}_{0,1}(w)\right| \\
&+ \sup_{w\in\mathcal{W}} \left| \mathrm{KLIC}_{G,2}(w) - \mathrm{KLIC}_{0,2}(w)\right| \\
&\xrightarrow{a.s.} 0. \qquad \text{(C.9)}
\end{aligned}$$

Similar to the proof in Theorem 4, results (C.8), (C.9) along with Assumption 12-5 lead to $\widehat{w} \xrightarrow{a.s.} w^*$, where w^* is the unique minimzier of $H_0^{(j)}, j = 1, 2$.

C.2 Supplementary Materials

In this appendix I show that Anderson-Darling Statistic is a Norm. The key is to check the triangular inequality. For $f : \mathcal{X} \to \mathbb{R}$, let

$$\|f\| = \left(\int_{\mathcal{X}} \frac{f^2(x)}{x(1-x)} dx \right)^{1/2}.$$

Then, for $1 \leq p < \infty$,

$$\begin{aligned}
\int \frac{|f(x)+g(x)|^p}{x(1-x)} dx &\leq \int \frac{(|f(x)|+|g(x)|)|f(x)+g(x)|}{x(1-x)} dx \\
&= \int \frac{|f(x)|}{[x(1-x)]^{1/p}} \frac{|f(x)+g(x)|^{p-1}}{[x(1-x)]^{(p-1)/p}} dx \\
&+ \int \frac{|g(x)|}{[x(1-x)]^{1/p}} \frac{|f(x)+g(x)|^{p-1}}{[x(1-x)]^{(p-1)/p}} dx \\
&= A + B. \qquad \text{(C.10)}
\end{aligned}$$

Using Holder inequality,

$$\begin{aligned} A &\leq \left[\int \frac{|f(x)|^p}{x(1-x)}\,dx\right]^{\frac{1}{p}} \left[\int \left(\frac{|f(x)+g(x)|}{[x(1-x)]^{1/p}}\right)^{(p-1)\frac{p}{p-1}} dx\right]^{1-\frac{1}{p}} \\ &= \left[\int \frac{|f(x)|^p}{x(1-x)}\,dx\right]^{\frac{1}{p}} \left[\int \left(\frac{|f(x)+g(x)|^p}{x(1-x)}\right) dx\right]^{1-\frac{1}{p}}. \end{aligned} \tag{C.11}$$

Similar argument works for B. Then let $p = 2$, applying simple algebra, we obtain the result $\|f+g\| \leq \|f\| + \|g\|$.