Semi-parametric instrument-free demand estimation: relaxing optimality and equilibrium assumptions*

Sungjin Cho, *Seoul National University*†
Gong Lee, *Georgetown University*‡
John Rust, *Georgetown University*§
Mengkai Yu, *Georgetown University*¶

May 13, 2019

Abstract

In most markets, consumer demand results from a compound arrival/choice process: consumers arrive to a market stochastically and make independent discrete choices over which item to purchase, including the "outside good". Market demand results from an aggregation of individual consumer choices, and in general is more accurately modeled a a price-dependent probability distribution rather than a linear demand curve that is traditionally studied in the literature on demand estimation. We consider the problem of identification of consumer preferences and arrivals when market prices are endogeneously determined but the implied distribution of demand (including mean demand) is nonlinear in prices and there are no relevant instrumental variables to deal with price endogeneity. In addition, most data sets are subject to censoring: we typically do not observe the number of arriving customers, or those who chose the outside good. Recent studies have shown that the hypotheses of a) optimality, and b) equilibrium constitute powerful identifying restrictions that enable demand estimation in the presence of a variety of endogeneity and censoring problems. In this paper we focus whether it is possible to identify demand when assumptions a) and b) are relaxed. We establish the global, non-parametric identification of preferences and consumer arrival probabilities in a simplified static setting but show via examples that the identification of unobserved types of consumers is very challenging, in contrast to the more optimistic conclusions from theoretical analyses that prove that the random coefficient logit model (which is a component of our overall model of demand) is non-parametrically identified.

Keywords identification, maximum likelihood estimation, semi-parametric estimator, instrument-free, structural estimation, identification of mixture models, price discrimination, information matrix

*__Acknowledgements:__ We thank Michael Keane and two anonymous referees for helpful suggestions that lead to us to develop a new estimator that relaxes a key assumption of optimal dynamic hotel pricing that we relied on in the previous version of this paper. We are also grateful to Yuichi Kitamura for helpful correspondence on the topic of identification of mixture models, and to the revenue manager of hotel 0 (who must remain anonymous due to confidentiality restrictions) for providing the reservation data that made this study possible. We also acknowledge STR for market level hotel occupany data it provided that proved key to the empirical identification of our demand estimation. We thank Georgetown University for research support, and Rust gratefully acknowledges financial support provided by the Gallagher Family Chair in Economics.

†Department of Economics, Seoul National University e-mail: sungcho@snu.ac.kr
‡Department of Economics, Georgetown University, e-mail: gl430@georgetown.edu
§Department of Economics, Georgetown University, e-mail: jr1393@georgetown.edu
¶Department of Economics, Georgetown University, e-mail: my262@georgetown.edu

1 Introduction

In most markets consumer demand results from a compound arrival/choice process: consumers arrive to a market over time stochastically and make independent discrete choices over which item to purchase, including the 'outside good'. Market demand results from an aggregation of individual consumer choices, and is more appropriately modeled as a nonlinear price and state-dependent stochastic process rather than a linear demand curve that is traditionally studied in the literature on demand estimation. We consider the problem of identification of consumer preferences and arrivals when market prices are endogeneously determined but the implied stochastic process for demand is nonlinear in prices and there are no relevant instrumental variables to deal with price endogeneity. In addition, most data sets are subject to censoring: we typically do not observe the number of arriving customers, or those who chose the outside good.

The motivation for our paper is an empirical analysis of hotel pricing in a specific luxury hotel market in a major US city, see Cho, Lee, Rust, and Yu (2018). Demand shocks in the face of the fixed room capacity for the 7 hotels in this local market lead to strong positive correlation between hotel prices and room occupancy: the hotels raise their prices to ration their available capacity on days where the demand for rooms in this market is high, but lower their prices significantly in to compete for market share on days when demand is low and there are a significant number of unsold rooms. This leads to strong co-movement in prices and occupancy rates in the hotels in this market. OLS estimation of room occupancy on hotel prices results in positively sloped "demand curves" for hotel rooms due to the endogeneity in hotel pricing in response to fluctuating demand. Price endogeneity also arises as a result of unobserved characteristics of the hotels in this market: although all seven hotels are classified as "luxury hotels" there are unobserved characteristics that make customers willing to pay more to stay in the top tier hotels in this local market, and their higher willingness to pay enables these hotels to charge more.

Though it has long been recognized that it is possible to deal with the latter form of endogeneity in certain types of nonlinear models using market share data only without the benefit of observing the discete choices of individual customers (see e.g. Berry, Levinsohn, and Pakes (1995)), these approaches depend on the ability to invert market shares to obtain transformed equations that are linear in prices, to which instrumental variable approaches can be applied. However it is unclear how to apply this approach in our hotel market example, where we typically observe only observe the occupancy of a single hotel in the market, but not occupancies at competing hotels. As a result, we do not observe market shares that the BLP estimator inverts to form the regression equations to which instrumental variables can be applied. Even in situations where we observe the joint occupancies of all of the hotels, it is still the case that we almost never observe 1) the total number of consumes "arriving"' to consider making a purchase in this

market, or 2) the number of arriving consumers who choose the "outside good' (i.e. to not stay in any of the hotels in this market). Without information on arrivals and the outside good, we cannot construct the market shares necessary to apply BLP. Even in the rare situations where the number of arrivals and/or the number of consumers choosing the outside good is observed, we show that the formulas for "market shares" are mixtures of censored multinomial distributions that cannot be inverted to obtain the linear-in-price estimating equation required by BLP.

On top of this, in many situations such as our hotel example, relevant instrumental variables do not exist. An instrument is an observed variable that causes exogenous shifts in a hotel's price but does not enter the hotel's pricing strategy. It is hard to think of such variables in the hotel market example. A typically used instrumental variable, the "Hausman instrument", is the price of hotels in different but "nearby" hotel markets. However many demand shocks are seasonal in nature and these seasonal effects result in strong co-movement of hotel prices and occupancy rates not only in the market we study but also in nearby markets. Thus, the Hausman instrument is not a good instrumental variable to deal with the classical "simultaneous equations" endogenity in hotel prices, even after including seasonal dummies in the regression equations. Besides seasonal dummies we do observe other "demand shifters" x that help hotels forecast demand, such as knowledge of big events, conventions, etc. But these x variables, which may be exogenous demand shifters, are "included variables" that hotels use to predict occupancy and set prices for their rooms and thus are not valid instruments either.

These problems motivate a search for new approaches to demand estimation. A recent paper by MacKay and Miller (2018) introduces a novel "instrument-free" approach to demand estimation: "Our main result is that price endogeneity can be resolved by interpreting an OLS estimate through the lens of a theoretical model. With a covariance restriction, the demand system is point identified, and weaker assumptions generate bounds on the structural parameters. Thus, causal demand parameters can be recovered without the availability of exogenous price variation." (p. 32). Their approach exploits additional restrictions implied by the assumptions of a) equilibrium and b) optimality, and depends on functional form assumptions similar to BLP, namely that demand, after a suitable transformation, is a linear (or semi-linear) function of price, as well as covariance restrictions between production cost shocks and additively separable unobserved demand shocks and unobserved product characteristics. Unfortunately we cannot apply their innovative approach here: we are not aware of any relevant "cost shocks" in the hotel market, and there is no transformation of occupancy rates that results in the linear equations needed to implement the MacKay and Miller (2018) estimator. Most importantly, our goal is to see if it is possible to identify the model when we also relax maintained assumptions a) and b).

In this paper we adopt a structural semi-parametric approach that explicitly models and attempts to identify the probability distribution governing the arrival of potential customers that constitutes the fundamental "demand shock" that leads to the co-movement of prices and occupancy in this market. We also seek to identify heterogeneous consumer preferences for the different hotels and their willingness to pay to stay in them. In particular, a leading text on pricing and revenue management, Phillips (2005), notes that successful hotel pricing strategies depend on a detailed understanding of customer heterogeneity: "There is both art and science to price differentiation. The art lies in finding a way to divide the market into different segments such that higher prices can be charged to the high willingness to pay segments and lower prices to the low-willingness to pay segments." (p. 558).

Dynamic structural models of demand and firm price setting provide an attractive alternative to traditional linear instrumental variables approach to demand estimation because they enable us to more directly model the dynamics of demand in real world markets, such as the hotel market we study in this paper. Identification of arrival probabilities and heterogeneous consumer preferences can be obtained without the use of instrumental variables or the use of covariance restrictions as in MacKay and Miller (2018), however the structural approach is heavily dependent on three key assumptions: 1) parametric functional forms for consumer preferences and the stochastic process governing arrival probabilities, 2) firms set prices optimally, and 3) firms are in a dynamic equilibrium, i.e. their prices mutually satisfy conditions for a dynamic Markov Perfect equilibrium or some related solution concept.

Together these three assumptions are often powerful enough to secure identification of the unknown parameters of consumer preferences and the stochastic process for arrival of customers to the market. The structural approach requires modeling the entire market and incorporating observed and unobserved variables that capture the demand shocks and unobserved product characteristics that result in the endogeneity in the prices we observe. By explicitly modeling the endogenous determination of prices under the assumptions of optimality and equilibrium, dynamic structural models are able to bring to bear "cross equation" identifying restrictions that imply that equilibrium prices are an implicit function of firms' beliefs about the stochastic process of customer arrivals and preferences, as well as each others' price-setting and strategic behavior. Demand is "downward sloping" in a well formulated structural model, since upward sloping demand would result in price dynamics that are at odds with what we actually observe.

However a big drawback of structural models is the computational demands of solving and simulating a dynamic Markov Perfect Equilibrium for an entire market. This is a daunting task even for a relatively small local hotel market consisting of 7 luxury hotels that we analyze in this paper. Fortunately, recent work has shown that it is often possible to identify demand in the presence of endogeneity in a framework

that relaxes the equilibrium assumption so long as the optimality assumption is still imposed. The idea is that the behavior of competing firms or agents can be flexibly modeled using semi-parametric estimators by treating the pricing strategies of competing firms as infinite-dimensional "nuisance parameters."

For example, Merlo, Ortalo-Magne, and Rust (2015) studied optimal dynamic strategies of home sellers in the London housing market. Endogeneity arises in this market due to the presence of unobservable characteristics of houses that make some more attractive to most buyers. Homes that are superior in unobservable dimensions (even after controlling for a large set of observed hedonic neighborhood and home characteristics) experience a high rate of arrival of offers and sell for more. Thus observed housing demand appears to be "upward sloping" in the list price of a home if we fail to control for price endogeneity. Yet there were few instrumental variables the authors could find to do this. The alternative was to estimate a dynamic structural model, but computing a full dynamic equilibrium in the London housing with thousands of competing buyers and sellers was out of the question. Merlo et al. (2015) were able to flexibly model the arrival of buyers and the dynamic bargaining process they employed. These can be regarded as the infinite-dimensional nuisance parameters in their estimation problem. However by assuming home sellers follow an optimal dynamic pricing and bargaining strategy in response to these beliefs, the authors were able to structurally estimate their model and obtain a sensible "downward sloping" demand for housing. The hypothesis of optimality restricts demand to be downward sloping in price, since if it were upward sloping, then it would be optimal for sellers to set far higher list prices than we actually observe.

In this paper we wish to take this approach one step further, to see if it is possible to identify demand by relaxing the hypothesis of optimality as well as equilibrium in the hotel market. Though the assumption of optimality is a powerful identifying assumption, it is also a potentially dubious one that could distort our estimates of demand if firms do not behave optimally. Herbert Simon introduced the concept of *bounded rationality* as a key reason why organizations and firms fail to optimize in complex environments: see Rust (2019) for further discussion and evidence in support of Simon's view that many firms *satisfice* rather than optimize. Cho et al. (2018) discuss a multi-billion *revenue management industry* that is experiencing very rapid growth by helping hotels, airlines, and other firms in the hospitality industry set better prices. If all of the hotels, airlines and other firms were already optimizing (the typical default assumption in most economic models), then there would be little need for for the revenue management industry. Yet, Phillips (2005) notes that despite the fact that pricing decisions "are usually critical determinants of profitability" "pricing decisions are often badly managed (or even unmanaged)." (p. 38).

We discuss the identification of a model demand in the presence of endogeneity, truncation and censoring when we relax the assumptions of equilibrium and optimality. We treat the price setting strategies

of firms as infinite-dimensional nuisance parameters and discuss semi-parametric maximum likelihood estimator that can estimate the parametric components of consumer preferences and arrival rates at the usual $\sqrt{N}$ rates (where N is the sample size). There is a cost to relaxing any maintained assumption, and relaxing the optimality and equilibrium assumptions has the consequence that the estimated pricing strategies are no longer implicit functions of firms' beliefs about consumer preferences and arrival rates. That is, the semi-parametric maximum likelihood estimator we propose no longer benefits from the "cross equation restrictions" that link consumer preferences and arrival rates to the pricing strategies of firms. Even though actual pricing strategies undoubtedly do depend on firms' beliefs about their customer preferences and arrival rates, the use of semi-parametric methods to estimate our *econometric model* comes at the cost of a loss of information from failing to impose the optimality and equilibrium restrictions. At a minimum this loss of information is reflected in larger asymptotic standard errors for the parameters, but in the worst case the demand model parameters may not be identifiable, and thus cannot be consistently estimated using any econometric estimation method that relaxes the optimization and equilibrium assumptions.

On the other hand, structural estimators that impose optimality and equilibrium restrictions will generally result in inconsistent demand estimates if actual firm behavior violates these assumptions. Thus, it is desirable to develop estimators that can identify demand without imposing these strong assumptions. If this is possible, it opens up a potentially powerful new avenue for econometrics to be useful for policy making, by enabling us to test hypotheses about firm behavior without imposing them *a priori.* For example, Harrington (2017) and Ezrachi and Stucke (2016) raise the specter of "algorithmic collusion" by sophisticated revenue management systems (RMS). The nature of "deep learning" algorithms from the artificial intelligence and reinforcement learning literatures makes it difficult to actually inspect the computer code used by commercial RMS to determine if it been explicitly designed to collude, or whether the algorithms "learn to collude" through repeated interaction. As we noted earlier, there is strong co-movement of prices and occupancy rates in the hotel market we analyzed, and some analyses might interpret such co-movement as a telltale sign of algorithmic collusion. Further, a structural model that imposed the hypothesis of collusion may result in distorted estimates of demand to help "rationalize" the maintained assumption of collusion. If it is possible to estimate demand without imposing strong assumptions about the type of equilibrium in this market, we can use the estimated demand model to solve for equilibrium under different equilibrium concepts, such as collusive pricing or Bertrand (competitive, non-collusive) pricing, and compare the predicted behavior to non-parametric estimates of the pricing strategies firms are actually using in this market. This may allow us to reject the hypothesis of algorithmic collusion in favor

of a model of competitive, Bertrand pricing where the price and occupancy co-movements are a natural response of a competitive market with inelastic supply of rooms that is subject to variable demand shocks.

Revenue management systems are proprietary so we do not know what sort of optimization principles they use and what types of data and econometric methods they employ. McAfee and te Veld (2008) note that "At this point, the mechanism determining airline prices is mysterious and merits continuing investigation because airlines engage in the most computationally intensive pricing of any industry." (p. 437). Phillips (2005) notes that "The tools that pricers use day to day are far more likely to be drawn from the fields of statistics or operations research than from economics." (p. 68) and he credits marketing (which he regards as a subfield of operations research and management science) noting that "marketing science has brought some science to what was previously viewed as a 'black art"' (p. 70). Yet "there remains a gap between marketing science models and their use in practice. The reasons for this gap are numerous. Many marketing models have been build on unrealistically stylized views of consumer behavior. Other models have been build to 'determine if what we see in practice can happen in theory.' Other models seem limited by unrealistically simplistic assumptions." (p . 70).

The empirical analysis of hotel pricing in Cho et al. (2018) demonstrates that it is possible to identify a realistic stochastic model of hotel demand that relaxes equilibrium and optimality assumptions. They focus on a single hotel that uses the IdeaS^{TM} RMS, a subsidiary of the SAS satistical software company. This hotel, which we refer to as "hotel 0" due to a non-disclosure agreement that prevents us from revealing its identity, follows the price recommendations of the IdeaS RMS approximately 60% of the time. On other occasions the revenue manager at hotel 0 deviates and chooses her own price. Though we are unable to observe which prices are the IdeaS recommended prices and which are set by the human revenue manager, we do know the information hotel 0 uses to set its prices, including the information in its own reservation database and real time information on the prices of its competitors from the *Market Vision^{TM}* pricing service. Hotel 0 cannot access the reservation databases of its competitors, and it generally does not know their occupancy rates, either *ex ante* (i.e. the number of rooms booked so far) or *ex post* (the actual or realized occupancy on a day by day basis). Though our identification results in section 3 demonstrate that it is in principle possible to identify the demand model parameters without observing occupancy rates of competing hotels, our ability to identify demand was greatly assisted by auxiliary data we obtained from the company STR Global which collects price and occupancy data on over 63,000 hotels worldwide.

Using the estimated demand model parameters, Cho et al. (2018) use dynamic programming (DP) to calculate counterfactual optimal dynamic hotel pricing strategies. In essence, their econometric demand model and DP algorithm constitute their own "RMS" and prediction of optimal recommended prices.

They find that the optimal dynamic prices from our model deviate significantly from the prices that hotel 0 actually charged. Via stochastic simulations of the estimated model, they find that adopting the counterfactual optimal dynamic hotel pricing strategy would have increased revenues of hotel 0 by approximately 12%, and resulted in higher occupancy rates. The counterfactual optimal pricing strategy typically results in significantly higher prices being charged for reservations made more than 20 days in advance of occupancy, but cuts price significantly as the occupancy date approaches, resulting in additional reservations and revenues compared to hotel 0's existing pricing strategy.

Overall, the analysis of Cho et al. (2018) suggests that far from engaging in "algorithmic collusion" we can reject the hypothesis that hotel 0 is even using a dynamically optimal pricing strategy (i.e. a best response to its competitors). Thus we can also reject the hypothesis that the firms in this market are setting prices in accordance with a dynamic Bertrand-Nash equilibrium. This conclusion is broadly consistent with Herbert Simon's work on satisficing behavior by firms. Since we do not observe which prices charged by hotel 0 are those recommended by the IdeaS RMS and which were chosen by its revenue manager, we cannot determine whether the source of hotel 0's suboptimality is due to its RMS or decisions by its human revenue manager.

A limitation of the analysis of Cho et al. (2018) is that their model of consumer demand did not allow for the presence of an outside good. However this has the huge drawback that under a calculated collusion counterfactual the colluding hotels should drive their prices to infinity, which is clearly not a plausible conclusion. Though a demand model without an outside good appears to fit the data quite well under "normal circumstances" where there is strong competition that drives hotel prices down to reasonable levels, it seems clear that an outside good must be present in the demand model if we want to obtain reasonable predictions about what prices might look like in more extreme counterfactuals such as collusion.

Thus an important question that we address in this paper is, "is it possible to identify demand parameters if we do not observe the total number of consumers arriving to the market, nor do we observe the number of consumers choosing the outside good?" We show that at least in a simplified static demand setting, the answer to this question is affirmative. Perhaps surprisingly, we also show that if we are willing to make parametric restrictions on consumer preferences (e.g. that their probabilities of choosing different hotels is given by a multinomial logit model and arrivals are governed by a negative binomial distribution), then the models parameters are identified even in situations where we observe occupancies of a single hotel in the market.

The general approach developed in this paper, i.e. of using a semi-parametric estimator to estimate

demand parameters while relaxing the assumption of optimality, has been used previously. Hall and Rust (2019) studied the pricing and inventory investment decisions by a firm that trades ("speculates") in the steel market. In their application, they had to confront different censoring and edogeneity problem. They refer to this as a problem of "endogenous sampling" due to censoring in the wholesale prices of steel that the the firm buys steel at. That is, the firm only records the price of steel on days it purchases steel. Hall and Rust (2019) estimated a dynamic structural model of optimal steel price speculation and an "unrestricted" model that relaxes the assumption of optimality using the Method of Simulated Moments (MSM, McFadden (1998)). They showed that it is possible to consistently estimate the parameters of the wholesale price process by censoring simulated data in the same way that actual data are censored. Using a Hausman test, they tested and rejected the assumption that the firm's steel purchases were governed by an optimal (S,s) inventory investment strategy. Using stochastic simulations, they also showed that if the firm had adopted an optimal (S,s) strategy, it would have earned significantly higher trading profits (as much at 40% higher) over the sample period.

Our confidence in counterfactual predictions from structural models rests on our confidence in our ability to identify consumer preferences that are part of a realistic model of demand in the presence of endogeneity and various types of censoring that we typically find even in the very best types of business data sets. The focus of this paper is to see whether it is theoretically possible to identify preferences and arrival probabilities in a simple static setting.

Section 2 summarizes the key features of our data set on hotel pricing, providing a concrete illustration of the price endogeneity and censoring problems we face. Section 3 discusses the identification of demand in a simplified static setting where the key ideas underlying our approach to identification can be explained more clearly. We prove a global, non-parametric identification result that appears to provide ground for optimism about our ability to identify and estimate demand under weak assumptions. However we also provide illustrative calculations of the asymptotic variances of parameters of consumer utility functions for hotels in a setting where there is unobserved heterogeneity among consumers. Contrary to the optimistic theoretical results of Fox, Kim, Ryan, and Bajari (2012) we find that the asymptotic variances of the distribution of random coefficients blow up exponentially fast as we increase the number of consumer types in our specification. Our results provide some insight into the commonly observed finding in empirical studies that it is very difficult to identify more than 2 or 3 types of consumers, even in very large data sets. Section 4 provides some conclusions and suggested directions for future research.

2 Hotel Data

This section describes the hotel market that motivates the questions about demand estimation and identification that we attempt to answer in this paper, and in particular the simple static model of hotel demand that we introduce in section 3. As we noted in the introduction, due to a non-disclosure agreement with the hotel that provided the data for our study, we are unable to provide too much detail about the local market in which hotel 0 operates to guarantee the anonymity of the hotel and the owner. We can say that it is a luxury hotel located in a highly desirable downtown location of a major US city.

Hotel 0 is one of seven luxury hotels operating in a well defined local market that is recognized by online travel agencies (OTAs) and other travel agents. Though customers can book at other luxury hotels in other parts of this city, the locations of these other luxury hotels are sufficiently far from this particular desirable area that they are not regarded as relevant substitutes for customers who wish to stay in this specific area of the city. We consider any choice of another hotel outside the seven hotels in this local hotel market, including the decision not to stay in any hotel, as a choice of the outside good.[1]

Table 1 lists some summary information about the seven hotels: all are 4-star or higher rated luxury hotels. To avoid identifying the hotels we show only the relative capacity, where we normalize the capacity of the largest hotel to 1. However our model uses all relevant information including the actual capacity, which we will show is an important factor in hotel pricing. The customers of the hotel are both business/government customers who mainly stay in the hotel on weekdays and tourists who typically stay on weekends. Since business customers and government customers are reimbursed for their travel expenses, we can expect them to be more price inelastic than tourists. On the other hand, many government agencies and large corporations that do frequent business in this city have negotiated government and corporate discounted rates with this hotel. These discounted rates are typically a fixed percentage, often 15 to 20%, off the currently quoted price that is called the *best available rate* (BAR). The revenue manager of hotel 0 is in charge of updating an array of BARs for different room classes and different future arrival dates and posting these prices to the web via the Global Distribution System (GDS, a network of computer connections that give travel agents access to a hotel's reservation database to check availability and reserve rooms) and via its own website.

Customers generally book hotel rooms in advance, and generally only a small fraction of customers (approximately 8%) book their rooms on the same day that they intend to occupy them. Hotel 0's reservation database provides information on when each hotel room was booked and through which *channel.* A

[1]There is also limited capacity of private residences in this area, so alternatives such as AirBnB is a minor factor in this market, and so we also lump this option into the catch-all category, "outside good."

Table 1: Hotels in the local market in our study

Property	Avg. BAR	Star	Class	Chained Brand	Rate	Relative Capacity	Distance to mass transit	Cancel Policy
hotel 0	$ 293.26	4	Luxury	No	4.4	79%	3 min	1 day before
hotel 1	$ 282.64	4.5	Upper Up	No	4.4	81%	5 min	3 day before
hotel 2	$ 285.16	4	Upper Up	No	4.4	63%	3 min	1 day before
hotel 3	$ 338.29	4	Upper Up	Yes	4.2	99%	8 min	2 day before
hotel 4	$ 397.09	4	Luxury	No	4.6	100%	10 min	Strict
hotel 5	$ 253.51	4	Upper Up	No	4.2	47%	8 min	3 day before
hotel 6	$ 454.30	5	Luxury	Yes	4.7	52%	10 min	1 day before

hotel room can be booked over the phone, via hotel 0's website, via a traditional travel agency, or via an *online travel agency* (OTA) such as Priceline or Expedia. The decentralized nature by which consumers search for hotels and book rooms implies that there is no single site or source that observes all consumers who "arrive" to book a room in a particular market during a particular point in time. A hotel will know how many consumers have booked one of its own rooms at every possible future arrival date, but there is no entity that observes all consumers searching for rooms or the number of bookings made in all of the competing hotels in a given market for arrival at any given future date. So this is the sense in which *arrivals are unobserved.*

Similarly, no single entity will know how many of the customers who arrive to book a room in a market will choose the outside good, i.e. to either choose to stay at a hotel outside this market or other form of accommodation such as AirBnB. Thus, from hotel 0's perspective, suppose that 10 customers book a room on a particular day for occupancy at some future date. Hotel 0 cannot distinguish between situations a) and b):

a) 100 customers arrived and 10 of these customers chose hotel 0, 50 of them chose one of its competitors, and the remaining 40 chose the outside good

b) 70 customers arrived, 10 booked at hotel 0, 50 booked at one of its competitors and only 10 chose the outside good.

In both cases a) and b) hotel 0 observes only the 10 customers who booked one of its rooms, but it has no information on the number of customers arriving in total, the number choosing the outside good, or even the number of customers who book a room at one of its competitors. In view of this, it is very difficult for a hotel to determine its *market share* on any particular day. Hotels do, of course, observe each

others *capacities* and they can obtain (at a cost) historical data on occupancies of their competitors from firms such as STR, but hotel 0 does not have real time information on the bookings and occupancies of its competitors.

The revenue manager uses a *uniform price strategy* and does not sell blocks of rooms to wholesalers under contracts that give wholesalers discretion to set their own prices for the blocks of rooms they purchase. Thus, there is no ability to "arbitrage" prices of rooms for hotel 0 by searching different OTAs. However hotel 0 does pay a significant commission, ranging from 15 to 25%, for reservations that are made via OTAs such as Expedia. The GDS that hotel 0 uses allows the revenue manager to change prices as frequently as she desires, though there is a short lag before the prices are propagated everywhere on the Internet including the leading OTAs. However for hotel 0's own website and reservation system, price changes take place instantaneously, and hotel 0 has its own loyalty program that provides discounts to customers who are members of the program. There are other groups that include weddings that involve a larger group of guests that are typically individually negotiated with the hotel revenue manager, but the discounts to these groups are typically quoted as a percentage discount off the BAR similar to corporate and government contract rates.

As we noted above, hotel 0 subscribes to the IdeaS RMS that provides recommended prices. The hotel revenue manager uses her own discretion to select a relatively small number of different possible BARs (effectively, she discretizes the pricing space) which are treated as a predefined choice set that is entered into the RMS. Based on a proprietary algorithm that considers remaining availability, seasonal effects, cancellation rates and competitors' prices, the RMS communicates a recommended BAR to the revenue manager at the start of each business day. Even though the revenue manager has some control over the prices the RMS can recommend via her choice of a predefined finite set of possible BARs, she often ignores the recommended price from the RMS and instead sets her own BARs based on her own experience, judgement and intuition. Unfortunately, our data do not specify which prices were ones recommended by IdeaS and which are ones she set herself, but she told us that she estimated she used the recommended prices approximately 60% of the time.

Thus, we do not know to what extent the IdeaS RMS is able to observe and adapt to the knowledge that the revenue manager is disregarding their recommended prices. This would seem to be important information that any RMS would want to collect, including the revenue manager's feedback about the overall quality of the recommended prices from the system. We can imagine that manual "price overrides" are common for newly launched hotels where the RMS may initially not have enough data to form good predictions about demand, or when there are unexpected changes to demand or entry/exit of other hotels

Table 2: Data sources used in this study

Data	The first day of occupancy	The last day of occupancy	Observations	Description
market vision	2010-09-21	2014-08-13	609,181	competitors' price
reservation raw	2009-09-01	2013-10-31	201,176	reservations detail information
cancellation raw	2009-09-01	2013-10-31	29,241	cancel detail information
daily pick-up report	2010-09-16	2014-05-21	475,187	daily revenue report
STR market data	2010-01-01	2014-12-31	1,731	competitors' occupancy
Data range	**2010-10-01**	**2013-10-31**		**37 months**

in the local market. In these cases we might expect that the recommended prices from the RMS would be less trustworthy until sufficient data are accumulated to enable the RMS to provide an updated model of customer demand that provides accurate predictions for the local market in question.

Hotel 0 provided us information from its reservation database that enabled us to track all bookings, cancellations, and prices for a 37 month period between September 2010 and October 2013. In addition, we were provided aggregate daily reports and their competitive daily rates of hotel 0's six competitors from a service called *Market Vision* provides quotes from hotel 0's six competitors for several room rate categories several times per day. While Market Vision provides excellent data on prices, as we noted above, it provides no information on the number of bookings or occupancies at hotel 0's competitors. This information does not seem to be readily available, but we were able to obtain data on the occupancy of hotel 0's competitors on a daily basis thanks to data provided by STR. Table 2 summarizes the data sources we used for our study.

Our data are unique in the level of detail we have on reservations and cancellations. Our reservation database contains the full history of each individual booking, including the channel through which the booking was made. Each booking is identified with a unique reservation identification number that is created when the reservation is initiated and becomes the permanent identifier for each reservation along with time stamps and dates of arrival and departure and amounts actually paid including incidental charges.

Hotel 0 has two basic categories of rooms: regular rooms and higher tier rooms such as luxury suites, but 95% of the rooms in the hotel are regular rooms. Thus we focus on the demand and prices of regular rooms. We rarely observe overbooking of the regular rooms, though on the few occasions where this happens the overflow customers are automatically upgraded to a room in one of the the higher tiers.

There are around 200 rate codes which can be broken into 14 categories. To simplify our analysis, we divided the codes into two; transient and group bookings. Transients are individual travelers who pay the

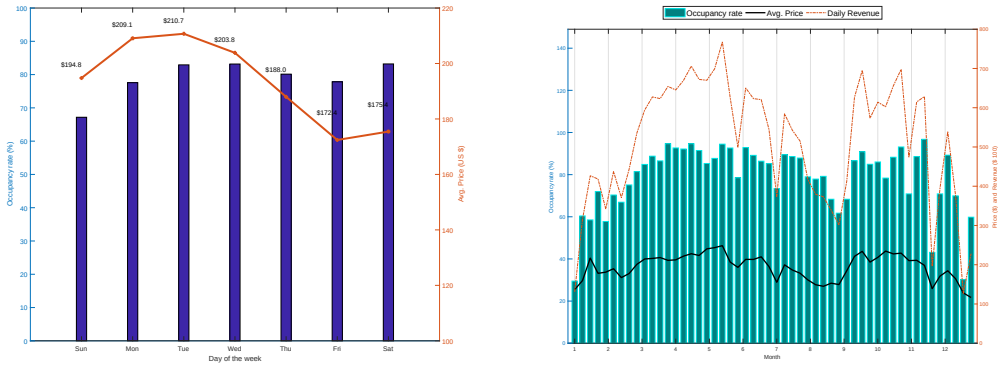
BAR or discounted BAR. Although the net of commission price that hotel 0 receives differs depending on which channel was used to do the booking (i.e. an OTA versus hotel 0's own website), transient customers themselves pay the same price regardless of channel, namely the BAR in effect at the time they booked. Group bookings are also generally based on the BAR in effect when they booked, however it will vary by pre-negotiated contract discount rate that differs from different groups (rate codes).

Although the data we have on hotel 0 provides an incredible level of detail, as we show in the next section, our model requires more data about the reservation/cancellation quantity dynamics of hotel 0's competitors that are not provided in the Market Vision data, which provide only competitors' prices. As we shall see, information on the total number of consumers who "arrive" and book rooms at one of the seven hotels in this local market is critical for our inferences about customer demand, and especially how customers respond to daily fluctuations in the relative BARs of the seven competing hotels. Unfortunately we do not have access to the reservation databases of hotel 0's competitors, so we are unable to observe the total number of new reservations that are made in at all the hotels and at which prices (including group, corporate discounts, etc) besides hotel 0. However as we show in the next section, it is possible to make inferences on the booking and reservation/cancellation dynamics of hotel 0's competitors given their prices if we can at least observe the total final occupancy rates of its competitors. Fortunately we were able to obtain this information from STR via an academic research contract it has with Georgetown University. In addition to total occupancy at each competing hotel on a daily basis, the STR data provide information on the competitors' ADRs and total revenue. The STR data turn out to be crucial for our ability to estimate a credible demand model.

2.1 Data summary

As we noted in the introduction, there is strong co-movement in prices and occupancy rates in this market. Figure 1 illustrates the cyclical pattern of occupancy and prices, both over a given week and over the year, reflecting seasonal variations in the demand for hotels. The bars in the left hand panel of figure 1 show a typical weekly cycle of occupancy for hotel 0 where the lowest occupancy is on Sunday, but a peak occupancy on Saturday, and a midweek peak occupancy on Tuesdays and Wednesdays. The ADR peaks on Tuesday, and the higher rates during the weekdays reflects price discrimination for less price elastic business guests, whereas the lower rates on Fridays and Saturdays are designed to attract more price elastic tourists. Occupancy is lowest on Sundays when tourists are checking out to return home for work on Monday, whereas a typical business guest checks in during the middle of the week and departs before the weekend. The right hand panel of figure 1 shows the price and occupancy dynamics over the year. Occupancy rates are the highest in the spring and early fall, and are lowest around holidays such

Figure 1: Booking and price dynamics over the week and year



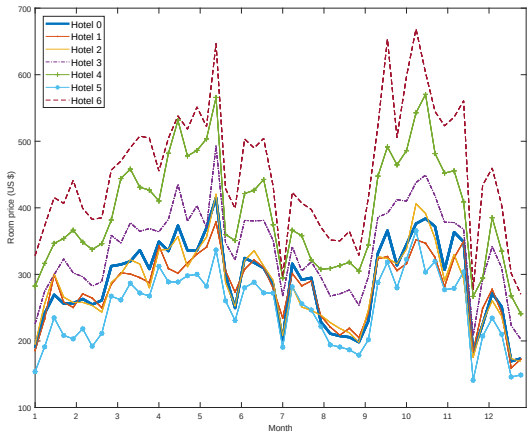
as Thanksgiving, Christmas and New Year's. The black line in the figure plots hotel 0's ADR and total revenues, and we see that both of these move in sync with the ups and downs in occupancy rates. This suggests that prices and revenues at hotel 0 are highly "demand driven".

Figure 2 compares the price dynamics for hotel 0 to those of its six competitors over the year. It plots the weekly average BAR from October 2010 to October 2013 for same-day reservations using the Market vision data, though we would obtain similar results if we plot a time series of ADRs using the STR data. The bold line plots the average BAR of hotel 0 while the other lines indicate BAR of six competitor hotels. We see strong co-movement in the prices of the seven hotels, and that they follow similar cyclical fluctuations, though hotel 0 tends to underprice its competitors with the exception of hotel 5. Similar the prices in figure 1 we find that prices are highest in the spring and the fall with peaks in early May and mid-September and October. Prices are lowest at the key holidays: Thanksgiving, Christmas, New Year's, as well as early July and August. During peak periods the average BAR of hotel 0 can be over $350 per night, whereas in the lowest periods it averages about $200.

The pattern of co-movement in the prices in this market might be described as "price following" and given the fact that most hotels use RMS and have extensive knowledge of their competitors' prices from services such as Market Vision, it could raise concerns about the possibility that the RMS enable these hotels to engage in algorithmic collusion. The price troughs following price peaks might be interpreted as "price wars" that are designed to punish hotels that deviate from the recommended prices that are highest when prices are peaking. However, as we will see, we do not think this is the correct conclusion to draw from these price patterns.

Figure 3 plots the time series of ADRs and occupancy rates for all seven hotels in this market for the first half of 2010 using the STR data. The top left panel plots the occupancy rate for hotel 0 versus the

Figure 2: Annual price dynamics for all seven hotels



occupancy rate of its competitors, where the competitor occupancy rate is defined as the total occupancy at the six competing hotels divided by the total room capacity of those hotels. With few exceptions, we see that occupancy follows the same weekly cycle at all of the hotels that we illustrated in the left panel of figure 1 for hotel 0, as well as the seasonal fluctuations (i.e. higher in the spring but lower at end of June) that we observed in the right panel of figure 1. The top right panel of figure 3 shows that all seven hotels also have strong weekly cycles in their ADRs and the reasons are likely to be much the same as we conjecture for hotel 0: higher mid-week prices to discriminate against less price elastic business guests and lower weekend rates to try to attract the more price elastic tourists.

The lower two panels of Figure 3 plot the cycles in occupancy rates (red lines) and ADRs (blue lines) for hotel 0 (right hand panel) versus its competitors (left hand panel). The data suggests that the weekly price cycles are driven not only by different compositions of guests (business versus tourists) but also to ration scarce capacity, since these hotels tend to be fully booked midweek but not on weekends. Both hotel 0 and its competitors follow similar weekly occupancy and price cycles, as well as similar seasonal price/occupancy cycles. For example we see that ADRs for both hotel 0 and its competitors peaked in mid April 2010, during a period where occupancy was close to 100% both mid-week and on the weekends.

It is natural to ask the question: which motive is more important for hotel 0? That is, does the revenue manager increase prices mainly to ration scarce capacity, or to try to exploit the more inelastic demand of business travelers who are more likely to be staying in the hotel midweek? Or, is hotel 0 simply following the prices of its competitors? If so, is this price following behavior a sign that all of the hotels are following the recommended prices from their RMS, and could this be evidence of tacit collusion mediated by the RMS?

Figure 3: Co-movement in ADR and cccupancy rates for all seven hotels

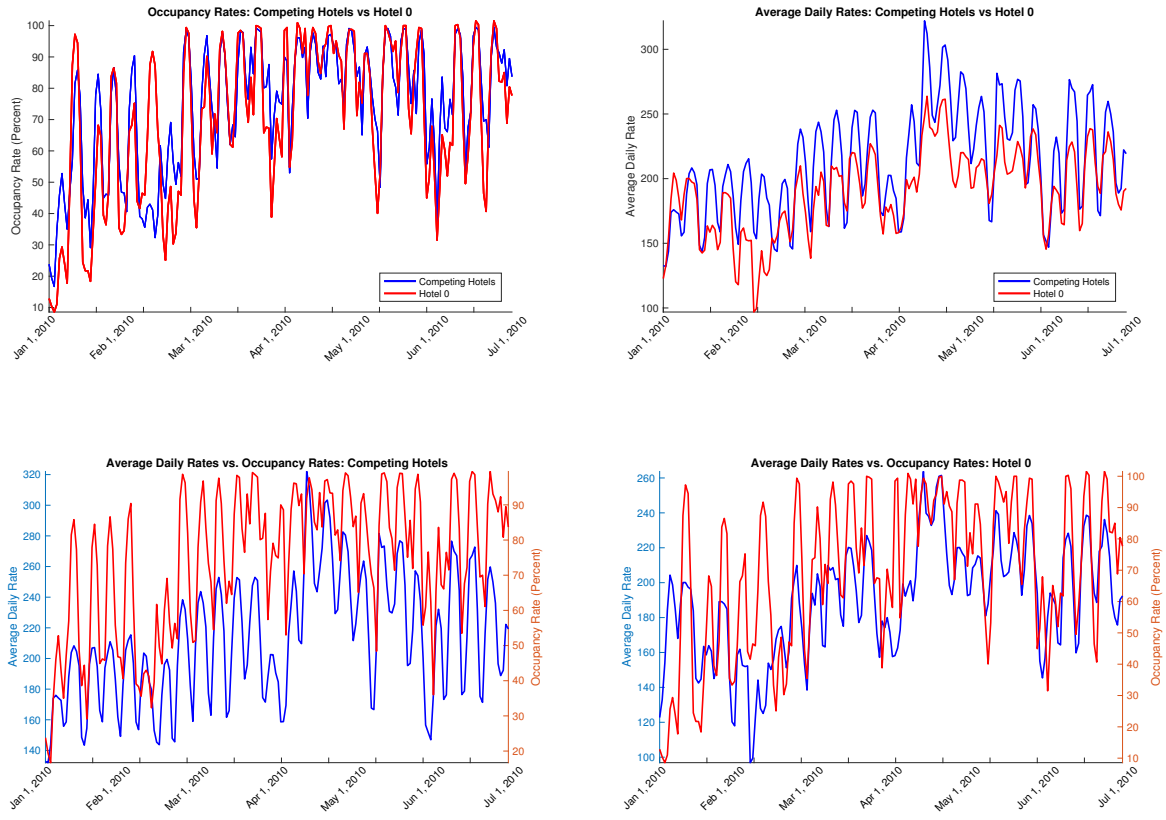


Table 3 provides some insight into this question by presenting the results of a simple OLS regression of the logarithm of hotel 0's ADR on the average ADR of its six competitors and on its own and competitors' occupancy rates. This simple model results in an R^2 of 86% when we also add dummies for different days of the week and months of the year to capture the weekly and seasonal price cycles.

Note that the occupancy also affects hotel 0's pricing but in a counterintuitive fashion: hotel 0's occupancy rate has a negative coefficient, but the occupancy rate of its competitors has a much larger positive coefficient. We may suspect that the co-movement in occupancy rates leads to a collinearity issue but hotel 0's own occupancy has a negative coefficient even after we remove the occupancy of the competing hotels from the regression. The coefficient estimate for Hotel 0's own occupancy rate only turns positive when we remove the ADR of the competing hotels, but then the fit of the model drops precipitously, to an R^2 of 0.17.

The regression findings suggest that the effect of occupancy on hotel 0's pricing decisions are second order relative to the dominant effect of the prices set by its competitors. To a first approximation, hotel

Table 3: Ordinary least squares regression with dependent variable ADR_0

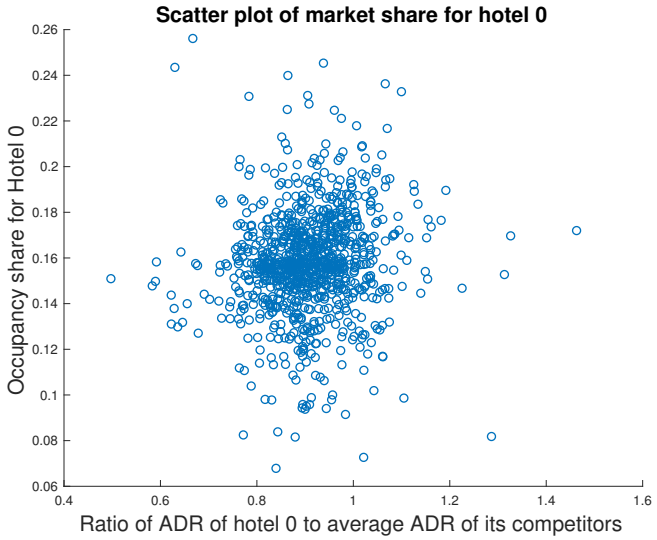
Variable	Estimate	Standard Error
constant	27.93	2.24
ADR_c	0.73	0.01
OCC_0	-0.09	0.027
OCC_c	0.273	0.044
$N = 1277, R^2 = 0.83$		

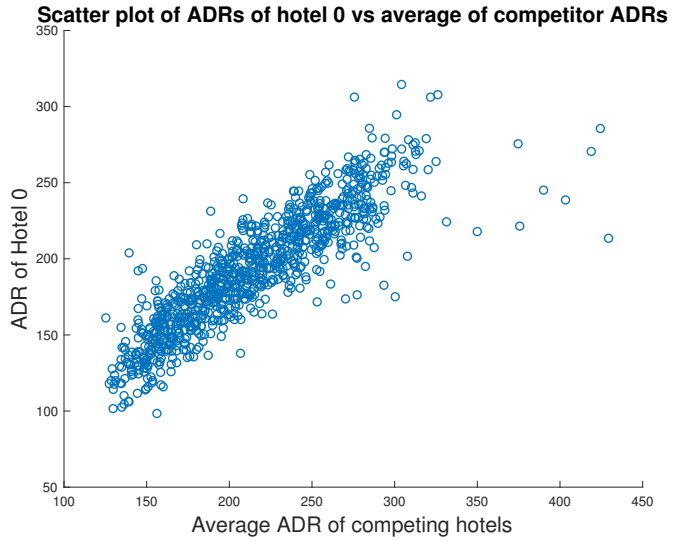
0 sets its prices at 70% of the average of its competitors' prices. The R^2 drops to 0.69 when we remove ADR_c from the regression but retain occupancy variables and daily and seasonal dummies. Overall, the regression results suggest that the revenue manager is setting prices in accordance with a "price following" strategy, and that knowledge of her competitors' prices is the most important piece of information besides the day of the week and the season of the year that she uses to set her own prices. The fact that hotel 0's own occupancy appears to have only a second order effect on its price setting once we condition on the prices of competitors suggests that raising prices to ration scarce capacity is not an important motive for hotel 0.

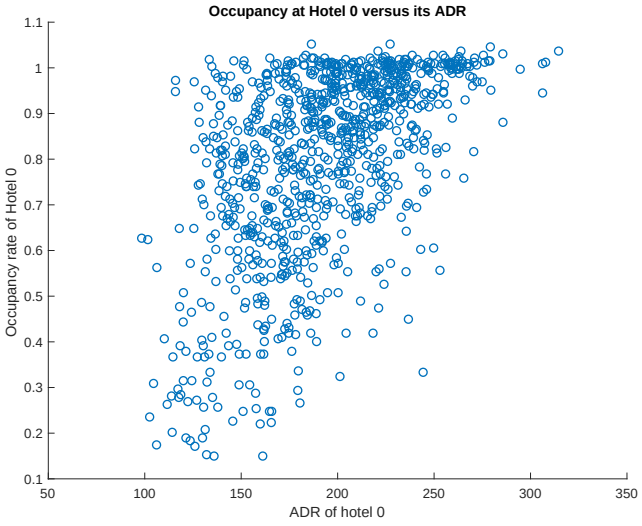
On the other hand it is not clear whether the fact that hotel 0's behavior is well approximated by "price following strategy" is evidence in favor of "algorithmic collusion" that Ezrachi and Stucke (2016) and Harrington (2017) discuss. Even if demand for rooms cycles in a systematic way during the week versus weekends, it is not clear that collusive prices would necessarily follow the same cyclical pattern that we observe in this market. In particular, we would expect that if the hotels in this market operated as a cartel, their prices would rise sufficiently high that there would be excess capacity even during the peak weekday periods, and the excess capacity would serve in part as a credible threat to engage in a price war that would deter any of the hotels that contemplated deviating from the collusive recommended prices, see Benoit and Krishna (1987) and Davidson and Deneckere (1990).

An alternative hypothesis is that this market is best approximated by a dynamic competitive equilibrium in a market characterized by strong Bertrand price competition subject to fixed capacity constraints. Stochastic shocks to demand lead to the price cycles we observe, with prices peaking to ration the available capacity in periods where demand exceeds available supply, but prices falling significantly as predicted by Bertrand price competition in periods of low demand where there is excess capacity. In this paper we will argue that the latter explanation is more likely to be closer to the truth, especially given what we have already reported about hotel 0's disinclination to follow the recommended prices of its RMS, combined with the fact that the revenue manager believes that the recommended prices are too low.

Figure 4: Price scatterplots for hotel 0 and its competitors







Regardless of the interpretation, the strong co-movement of hotel 0's prices with the prices of its competitors creates real difficulties for demand estimation. We can see the problem in figure 4. The right hand panel lots the ADR of hotel 0 against the average ADR of its competitors. The co-movement in prices is evident it the positive correlation in prices we see in the figure, something already captured in the regression results in table 3. However we might expect that on days where the *relative price* of hotel 0 is higher that there should be fewer guests booking its rooms. But the left hand panel of figure 4 shows that there is little evidence in favor of this hypothesis. The scatterplot of hotel 0's share of total occupancy on the ratio of hotel 0's ADR to the average ADR of its competitors is roughly a circle of dots, which explains why we do not obtain a negative coefficient on hotel 0's price in an OLS regression of its market share on the ADR of hotel 0 relative to its competitors.

The final panel of figure 4 is a scatterplot of hotel 0's own occupancy against its ADR over the period of our sample. This figure shows an *upward sloping* relationship between price and occupancy, which encapsulates the endogeneity problem we discussed in the introduction. We believe the endogeneity is emblematic of the classic Cowles Commission simultaneous equations type of endogeneity between prices and quantities. If the market is well approximated as a dynamic Bertrand equilibrium but subject to large stochastic demand shocks, then we would expect to see high prices set to ration demand when demand is high but low prices as the hotels compete for the available demand in periods where there is excess supply of rooms. This type of competition will generate a positively sloped scatterplot of prices similar to what we observe in figure 4 and generally a positively sloped relationship between ADRs and occupancy for hotels individually. Thus, simple OLS regressions will result in *positively sloped demand curves* in this market.

There are no obvious instrumental variables that can solve this endogeneity problem. One possible instrumental variable is a decrease in capacity of the hotel. If we regard the hotel as setting prices to ration demand, then in periods where there is a reduction in available rooms for exogenous reasons (such as a bursted pipe or other problems that remove rooms from service temporarily, or planned upgrades to rooms that take rooms out of service for a period of time or permanently, such as when the hotel converted 23 of its standard rooms to deluxe rooms), then the decrease in supply of rooms may serve as an instrumental variable that may allow us to estimate a negatively sloped demand curve.

Unfortunately when we tried to use available capacity as an instrument we find highly unreliable and generally insignificant results. Depending on the subsample we use, that estimated coefficient for 2SLS of the log of the ratio of the ADR of hotel 0 to the average ADR of its competitors ranges from -4.02 to 7.76 but the maximum t-statistic for any of these subsamples is 1.2. Most likely the capacity instruments are

weak instruments since the F-statistic in the first stage regressions ranges from 0.03 to 5.43. There is not enough exogenous variation in hotel 0's available capacity to make this a good instrument for estimating the effect of hotel 0's price on demand. An additional complication is that the model of demand we introduce in the next section does not result in a linear demand equation. Instead the model is a *price-dependent probability distribution for demand* that results from a micro-aggregation of individual choices of consumers who arrive at random to book a room at one of these hotels and choose the best option (including the outside good) given the BARs quoted by these hotels at the time they make their decisions.

3 Identification of a Static Model of Hotel Demand

We analyze the identification problem in the context of a simple static model of hotel demand, i.e. a model where there are no advance bookings of hotel rooms. The static setting allows us to illustrate our approach and key issues with less notational complexity. We focus on a particular local hotel market and assume that the L hotels in this market do not take advance bookings but rather on each day t the hotels set prices (simultaneously), then a random number of customers arrive and choose which hotel to stay at, after observing the vector of prices p_t that the hotels set that day. One of the options available to all consumers is the "outside good" i.e. to not stay in any of the hotels. We develop a model that is rich enough to reflect the censoring and endogeneity problems that we observe in our hotel data set, and thus provides a simple initial framework to convey the basic ideas underlying our approach.

The fundamental challenge is one of identification: the econometrician does not observe $\tilde{A}$, the number of customers arriving to book rooms, nor does the econometrician observe the number of arriving customers who choose the outside good. In the statistics literature this is known as a problem of *truncation.* We find this statistical term confusing since our data is also subject to a related problem of *censoring:* at each hotel we only observe the minimum of its actual demand for rooms and the hotel's room capacity. A further data problem is that the econometrician may not observe the occupancy rates at competing hotels. We also consider the possibility of *unobserved heterogeneity* i.e. the econometrician does not observe an individual's preferences for the various hotels or their degree of price sensitivity, which is crucial information necessary to determine the customer's willingness to pay to stay in various hotels in this market.

The econometrician can observe the vector of prices p chosen each day by the hotels, and the realized occupancies, d, (an $L \times 1$ vector) though as we noted, these occupancies are censored at the capacities C (also an $L \times 1$ vector) at each of the hotels. When a hotel reaches capacity it turns customers away. We assume that when this happens, customers are randomly turned away, and all of them choose the outside

good rather being directed to a competing hotel in this market. Thus, with probability 1 the hotel capacity constraint $d \leq C$ is enforced. Let d_t be the total number of consumers who are booked in one of the hotels on day t where e is an $L \times 1$ vectors of 1's. We have with probability 1, $\tilde{A}_t \geq d_t' e$, and the difference, $\tilde{A}_t - d_t' e$ is the number of customers who end up with the outside good, either via a voluntary choice, or because the hotel they chose was full.

We assume the hotels (and the econometrician) observe a vector of demand shifters x_t that helps them predict the number of arrivals A. There may also be other idiosyncratic variables z_t that are specific to each hotel that each hotel observes, that affect their pricing decision. We assume the econometrician does not observe the vector z_t, which we assume for simplicity is an $L \times 1$ vector where the l^{th} component z_{tl} represents a scalar idiosyncratic shock that is observed only by hotel l and reflects factors specific to hotel l how sets its price that the other hotels (and the econometrician) do not observe.

The timing of decisions in each market day t is as follows

1. At the start of the day, all L hotels observe the demand shifter x_t that help them predict the number of customers who will arrive to book rooms later that day. Each hotel l also observes idiosyncratic factors that affect its own pricing decision, z_{tl}, but not the idiosyncratic shocks observed by each of its competitors $\{z_{tl'}\}$ for $l' \neq l$. Customers may observe x_t but no customer nor the econometrician, observes the idiosyncratic shocks z_{tl}, $l \in \{0, \ldots, L-1\}$.

2. Based on the information (x_t, z_t) the firms set their prices, so we can write $p_t = p(x_t, z_t)$ is the price set at the start of day t prior to the arrival of customers. The price set by hotel l depends only on its own idiosyncratic realization z_{tl}, so the pricing rule for hotel l is independent of $\{z_{tl'}\}$, $l' \neq l$ and so can be written as

$$p_{tl} = p_l(x_t, z_{tl}), \quad l \in \{0, \ldots, L-1\}. \tag{1}$$

   The hotel pricing also reflects the common knowledge of an $L \times 1$ vector ξ of characteristics of each of the hotels that constitute attributes of each of the hotels that affect customer preferences that the hotels and customers observe but the econometrician does not observe. We assume that customers observe ξ and these characteristics affect their preferences for the hotels. Similarly the hotels' perception of customer preferences in turn affects their pricing decisions. However we do not include the unobservable variables ξ as an explicit argument of the pricing function (1), though our model does allow prices to be an implicit function of their perceptions of consumer preferences which may in turn depend on the time-invariant unobserved attribute vector ξ.

3. Customers observe the demand shifter x_t and the characteristics of the competing hotels (and the

outside good) ξ, but the number of customers arriving to book a room in hotel market is a random process that does not depend on prices p_t given x_t. We will let $H(A|x)$ be the distribution of consumers who arrive when market conditions are summarized by the observed demand shifter x but assume that the distribution of arrivals does not depend on ξ or prices p since customers only learn about ξ and p once they arrive at the market and consider the available alternatives.

4. We assume that each customer who arrives on day t to book a room observes ξ and the price vector p_t and chooses to stay in one of the L hotels based on a simple static utility maximization decision, where the hotel chosen by customer $a \in \{1,\ldots,\tilde{A}\}$ can be any of the L hotels or $l = \emptyset$ denotes the choice of the outside good (to not stay in any of the L hotels and go to some other hotel outside of this local market). We assume there are a finite number of possible types of consumers indexed by τ and there are *IID* random utility shocks that affect each consumer's choice of which hotel to stay at. We assume that the choice of hotel by consumer j on day t, l_{tj}, is given by

$$l_{tj} = \underset{l\in\{\emptyset,0,1,\ldots,L-1\}}{argmax} [u_\tau(l, p_{tl}, x) + \varepsilon_{tj}(l)] \tag{2}$$

The utility each consumer obtains from the different hotels is an implicit function of the hotels' attributes ξ, so in this sense, the set of consumer utility functions $\{u_\tau\}$ is a sufficient statistic for the set of hotel attributes ξ and in order to set prices hotels should have a good knowledge of consumer preferences. Knowledge of how hotel attributes ξ affect customer preferences is only relevant for longer term decisions by hotels, such as investment in hotel upgrades, etc.

5. We assume the $(L+1)\times 1$ vectors of the random utility components $\varepsilon_{tj} \equiv \{\varepsilon_{tj}(l) | l \in \{\emptyset, 0, 1, \ldots, L-1\}$ are continuously distributed with unbounded support over R^{L+1} and are independently distributed across different customers who arrive in this market, the behavior of a consumer of type τ can be represented by a conditional choice probability $P_\tau(l|p,x)$ which provides the probability that the consumer will choose hotel l (or the outside good if $l = \emptyset$) given the price vector p when the observed demand shifter is x. Since choices of different customers are made independently of each other, the total *potential demand* by the A customers who arrive on t given by a multinomial distribution with parameters $(A_t, \pi_\emptyset(p,x), \pi_0(p,x), \ldots, \pi_{L-1}(p,x))$ where

$$\pi_l(p,x) = \sum_{\tau=1}^{\mathcal{T}} P_\tau(l|p,x) g(\tau|x), \quad l \in \{\emptyset, 0, 1, \ldots, L-1\} \tag{3}$$

where $g(\tau|x)$ is the fraction of consumers of are of type τ on a day where the observed demand shifters equal x.

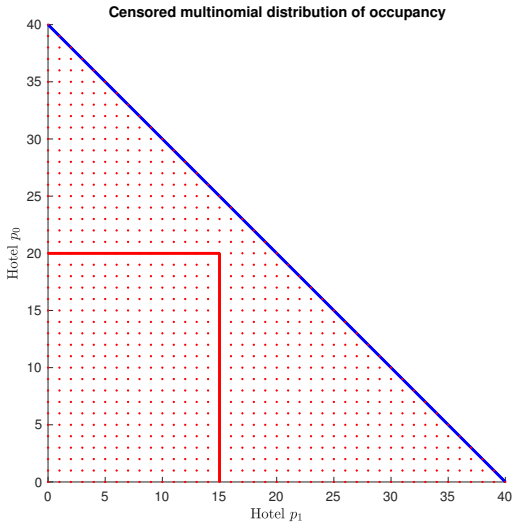


Figure 5: Censored multinomial distribution of occupancy

6. Let $f(d|A,p,x)$ be the conditional distribution of *realized occupancy* in the L hotels, where $d = (d_0, d_1, \ldots, d_{L-1})'$ is an $L \times 1$ vector of occupancies in the L hotels that satisfies each hotel's capacity constraint, $d_l \leq C_l$, where $l \in \{0, \ldots, L-1\}$. This distribution is a *censored multinomial distribution* given by

$$f(d|A,p,x) = \begin{cases} M(d|A,p,x) & \text{if } d_l < C_l,\ l \in \{0,\ldots,L-1\} \\ \sum_{d' \in D^c(d)} M(d'|A,p,x) & \text{otherwise} \end{cases} \tag{4}$$

where

$$D^c(d) = \left\{ d' | d'_l \geq d_l \text{ if } d_l = C_l,\ d'_l = d_l \text{ if } d_l < C_l,\ \text{where} \sum_l d'_l \leq A \right\} \tag{5}$$

where C_l is the room capacity of hotel l and $M(d|A,p,x)$ is the uncensored multinomial density of potential demand

$$M(d|A,p,x) = \frac{A!}{(A - \sum_{l=0}^{L-1} d_l)! d_0! d_1! \cdots d_L!} \pi_\emptyset(p,x)^{(A - \sum_{l=0}^{L-1} d_l)} \pi_0(p,x)^{d_0} \pi_1(p,x)^{d_1} \cdots \pi_{L-1}(p,x)^{d_{L-1}} \tag{6}$$

where $d_l \in \{0, 1, \ldots, A\}$, $l \in \{0, \ldots, L-1\}$ and $\sum_l d_l \leq A$.

Figure 5 illustrates the rectangular support of the censored multinomial distribtion in the case of $L = 2$ hotels, one with a capacity of $C_0 = 20$ rooms and the other with a capacity of $C_1 = 15$ rooms. The triangular region illustrates the support of the uncensored trinomial distribution when there are $A = 40$ arriving customers. Thus, each pair of potential demands (d_0, d_1) satisfying $d_0 + d_1 \leq 40$ is in the support of the trinomial distribution, with the residual customers $d_\emptyset = 40 - d_0 - d_1$ taking the outside good. Realized occupancies (d_0, d_1) must statisfy the capacity constraints $d_0 \leq C_0$ and $o_1 \leq C_1$ with probability 1. For

example the demand $(d_0, d_1) = (20, 20)$ is not feasible since the demand for hotel 1 is $d_1 = 20$ which exceeds its capacity, $C_1 = 15$. In such a case we assume the hotels serve customers on a first come, first served basis until their capacity is reached and all "excess customers" choose the outside good. Thus the realized occupancies equal $(20, 15)$ in this case, and the excess 5 customers who could not be accommodated at hotel 2 are assumed to choose the outside good. When we calculate the probability of any realized occupancy where one or more hotels is sold out, we sum the multinomial probabilities over all possible potential demands d' that are consistent with the specified hotels being at capacity, i.e. over all indices $d'_l \geq C_l$ for for those hotels l which are at capacity, $d_l = C_l$. For example the probability of the occupancy pair $d = (d_0, d_1) = (0, 15)$ is the sum over all potential demands $d' = (d'_0, d'_1)$ in the set $D^c(d)$ which is the set of all indices d' where $d'_0 = 0$ and $d'_1 \geq 15 = C_1$ and $d_\emptyset = 40 - d'_1$, i.e. the sum of the probabilities of all integer coordinates on the x axis of figure 5 from hotel 1's capacity of $C_1 = 15$ to the total number of arrivals, $A = 40$.

In the case both hotels are sold out, $d = (C_0, C_1)$, then $D^c(d)$ is the set of all integer lattice points in the northeast quadrant above corner point $d = (15, 20)$ and below the blue line in figure 5. That is, $D^c(d) = \{d' | d'_0 \geq 20,\ d'_1 \geq 15,\ \text{and}\ d'_0 + d'_1 \leq 40\}$. All possible realizations of hotel demand in the set $D^c(d) - \{15, 20\}$ correspond to 5 customers who ended up with the outside good, either voluntarily or because their preferred hotel was full. For example the point $d = (16, 21)$ corresponds to a case where 3 customers chose the outside good, 21 chose to book at hotel 0, and 16 at hotel 1. Since this demand exceeds the capacity of the two hotels, the 2 "excess customers" one from hotel 0 and the other from hotel 1 are diverted to the outside good.

We now state the key assumptions about the timing of decisions and price setting and demand in our stylized static model of the hotel market given above.

Assumption 0 (Endogeneity) *Hotels set their prices prior to knowing the number of customers A who arrive to book rooms in the market, however due to the common dependence on x, prices p and arrivals A will generally be positively correlated, and thus also positively correlated with occupancy d. Hotel prices will also generally depend on the unobserved characteristics of hotels, ξ, which affect consumer preferences and willingness to pay for different hotels.*

Assumption 1 (Stationarity and independence) *The pricing rule that hotels use to determine prices in equation (1) is time-invariant. The demand shifters $\{x_t\}$ that enter the pricing rule may be serially correlated, but are distributed independently of the idiosyncratic shocks $\{z_t\}$ affecting firm prices. The process $\{x_t, z_t\}$ is strictly stationary and is $\{z_t\}$ is IID with components that are continuously and independently distributed: i.e. if $l \neq l'$, then z_{tl} and $z_{tl'}$ are independent continuous random variables for each t.*

Assumption 2 (Conditional Independence) *The censored multinomial conditional distribution of occupancy given in equation (4) is time-invariant and independent of z given (A, p, x). That is, we have:*

$$f(d|A, p, x, z) = f(d|A, p, x), \tag{7}$$

where $f(d|A, p, x)$ is the censored multinomial conditional distribution of hotel occupancy given in equation (4).

Assumption 2 is the key to our semi-parametric identification strategy. The unobserved "price shocks" z create random variability in prices that enables us to identify the effect of price on demand for hotel rooms after controlling for x, which is the observable demand shifter that is the fundmental source of endogeneity in this model. We might also refer to Assumption 2 as *conditional exogeneity of prices* since conditional on x, the remaining variation in prices is random, so we have price variation that is akin to a randomized experiment that helps us to identify the effect of prices on consumer demand for hotels.

We also think of Assumption 2 as analogous to the conditional independence assumption in the treatment effects literature, where the assignment of a "treatment" is assumed to be independent of the potential outcomes, conditional on a vector of covariates x. Thus, assignment of a treatment can be treated as a virtual random assignment, given x in the sense that prices can be regarded as "randomly assigned" given x due to the effect of unobserved idiosyncratic factors z affecting how the hotels set their prices. The conditional distribution $G(p|x)$ captures the variability in prices due to the effect of the z shocks and the conditional distribution $H(A|x)$ reflects the uncertainty about the number of arrivals given only knowledge of the observed demand shifter x.

Note that Assumptions 0 to 2 are completely agnostic on how hotels set prices in the market, and in particular there is no assumption about hotels setting prices optimally or in a manner consistent with a price equilibrium in this market. Let $G(p|x)$ be the conditional distribution over the prices set by the hotels given the observed demand shifter x that are induced by the idiosyncratic price shocks z. In effect, this distribution is an infinite dimensional "nuisance parameter" since our primary interest is to infer consumer preferences and the arrival probability $H(A|x)$.

With enough data on a given market, it is possible to non-parametrically estimate the distribution $G(p|x)$. Note that prices may reflect the effect of unobserved characteristics of hotels ξ, and we will show that prices can reflect endogeneity due to classical simultaneous equations bias. That is, variations in the number of arriving customers across different market days with different *ex ante* values of x that constitute observed demand shifters is enough in the face of the limited capacity of the hotels in this market to result a) a strong co-movement in prices among the various hotels in the market, and b) an upward sloping relationship between price and occupancy at individual hotels. Note, however, that since

we can non-parametrically estimate $G(p|x)$ we do not have to take a stand on how individual hotels set their prices. We will now show that Assumption 2 is sufficient to identify the effect of price on the demand for hotel rooms without requiring us to develop a detailed model of equilibrium in the hotel market, or even to assume anything about how individual hotels set their prices as a function of x and z, such as the assumption that individual hotels set their prices optimally as a best response to their beliefs of the prices set by their competitors.

For simplicity, we will assume that the domain of possible demand shifters x is a finite set X, and the maximum number of customers arriving for any $x \in X$ is also finite. We assume that the idiosyncratic shocks z are continuously distributed and affect hotel prices in a continuous fashion so that prices are continuous random variables whose support is a subset of the positive orthant of R^L.

Figure 6 illustrates the joint occupancy distribution for two hotels ("hotel 0" and "hotel c") whose capacities are $C_0 = 30$ and $C_1 = 50$ respectively. We assume that the number of arrivals is $A = 120$ and the two hotels have accurate signals of the number of arrivals and set their prices accordingly. Since hotel 0 has a smaller capacity of $C_0 = 30$, it sets a price of $p_0 = 169$ and hotel c with the larger capacity of $C_1 = 50$ sets a price of $p_1 = 181$. We see that there is significant probability that hotel 0 sells out, while the chance that hotel c sells out is close to zero due to its higher capacity.

For purposes of illustrating identification, our behavioral assumption is that the hotels choose prices that constitute a "quasi-Bertrand equilibrium." That is, for a given value of the demand shifter x we assume hotel 0 chooses price $p_1^*(x)$ such that

$$p_0^*(x) = \underset{p_0}{argmax}\,[\min(C_0, E\{A|x\}\pi_0(p_0, p_1^*(x)))(p_0 - c_0)] \tag{8}$$

where $c_0 = 50$ is the marginal cost of serving a room, $E\{A|x\}$ is the expected number arrivals given x, and $p_1^*(x)$ is the price chosen by hotel c, which is given by

$$p_1^* = \underset{p_1}{argmax}\,[\min(C_1, E\{A|x\}\pi_1(p_0^*(x), p_1))(p_1 - c_1)] \tag{9}$$

where $c_1 = 50$ is hotel 2's cost of servicing a room.

We describe the solution $(p_0^*(x), p_1^*(x))$ to the hotels' profit maximization problems as a "quasi Bertrand equilibrium" that differs from a true Bertrand-Nash equilibrium in prices. This is because we assume the hotels invoke approximations to simplify the calculation of expected profits. Actual expected profits for hotel 0 should be calculated as $E\{\min(C_0, \tilde{d_0})(p_0 - c_0)|p_0, p_1^*(x), x\}$ given by

$$E\{\min(C_0, \tilde{d_0})(p_0 - c_0)|p_0, p_1^*(x), x\} = \sum_{d_0, d_1} \min(C_0, d_0)(p_0 - c_0) f(d_0, d_1|p_0, p_1^*(x), x) \tag{10}$$

where $f(d|p,x)$ is the conditional distribution of censored demand (occupancy) for the two hotels that does not condition on the number of arrivals A since the hotels do not observe A at the start of each day

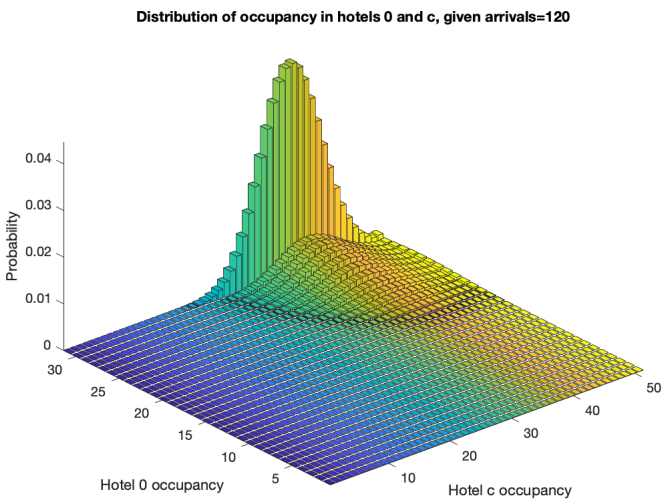


Figure 6: Occupancy distribution, $A = 120$

when they set their prices. This distribution is given in equation (13) below. Since the firms are not calculating their expected profit functions correctly, but instead use a computational shortcut, it follows that the pair of prices $(p_0^*, p_1^*) = (169, 181)$ are not Bertrand equilibrium prices. They can be regarded as a "behavioral approximation" to Bertrand pricing that reflects an unwillingness or imperfect ability to optimize, since both firms employ a convenient but incorrect shortcut approximation when they calculate their expected profit functions.

Note that our analysis of identification does not exploit any knowledge about how the hotels set their prices. Indeed, in the next section we show that to estimate parametric versions of the model by semi-parametric maximum likelihood, where we treat observed prices $\{p_t\}$ as data, so the estimator does not depend on any assumptions or an explicit structural model of how firms set their prices.

We assume that on any given day and for any given demand shock x, unspecified information or idiosyncratic shocks z cause the firms to deviate slightly from this quasi-Bertrand equilibrium pair of prices $(p_0^*, p_1^*) = (169, 181)$. Thus to generate simulated price and occupancy data, we randomly perturbed these prices by up to plus or minus 5% with uniformly distributed idiosyncratic pricing shocks. Figure 7 presents simulated price scatterplots that show that our simple static model is capable of generating endogeneity and price co-movements similar to what we observe in the actual data that we illustrated in figure 4 of section 2. Each of the panels of figure 7 should be compared to the corresponding panels of figure 4.

We generated the simulated data as described above, by randomly perturbing the quasi-Bertrand equilibrium prices for 4 different values of x that correspond to the means of trinomial distributions $H(A|x)$ for arrival of ustomers, where $x \in \{60, 100, 120, 140\}$. We assume that given x actual arrivals equal x

with probability 1/2, or $x-10$ or $x+10$ with probability 1/4, respectively. Thus, we assume that the firms receive fairly accurate signals of the number of customers arriving on different days and price accordingly.

We simulated data for a hypothetical market, with 50 observations from each of the 4 possible values of the demand shifter x, and plot the results in figure 7. When firms expect the number of arrivals to be 100 or fewer, neither expects to sell out, and hotel 0 underprices its competitor since it believes customers regard its competitor to be a superior hotel. In these cases the quasi-Bertrand equilibrium prices are $(p_0^*, p_1^*) = (167, 180)$. However as we noted above, if the hotels expect 120 arriving customers, then the quasi-Bertrand equilibrium involves hotel 0 charging slightly higher prices since hotel 0 now expects to sell out. On days where the hotels believe $A = 140$ customers will arrive, both hotels expect a sufficiently high chance of selling out that they can charge even higher prices. Due to the smaller capacity of hotel 0, it rations its scarcer capacity more tightly by charging a price $p_0^* = 188$ that is now even higher than the price charged by hotel c, $p_1^* = 183$. At these prices hotel 0 expects to sell out but hotel c expects that 45 of its 50 rooms will be occupied. Thus, the hotels expect that 45 of the 120 arriving customers will choose the outside good.

The simulated prices and occupancies from this simple model reveal patterns of price endogeneity that are remarkably similar to those we observed in the actual data in figure 4. The top panel of figure 7 shows that there is no clear relationship between the market share of hotel 0 and the ration of its price to that of its competitor, very similar to what we see in the actual data. The second panel reveals the strong positive association in the prices of the two competing hotels and this is driven largely by the demand shocks, x that cause both hotels to raise prices to ration their scarce capacity when they expect more customers to arrive, even though a significant share of arriving customers choose the outside good. Finally, the final panel of figure 7 reveals a positive correlation between price and occupancy rates at hotel 0, very similar to what we see in the actual data in (see bottom panel of figure 4). Again, this positive association results from an endogenous price and quantity response to shifts in demand, and is not a sign of a perverse "positively sloping demand curve."

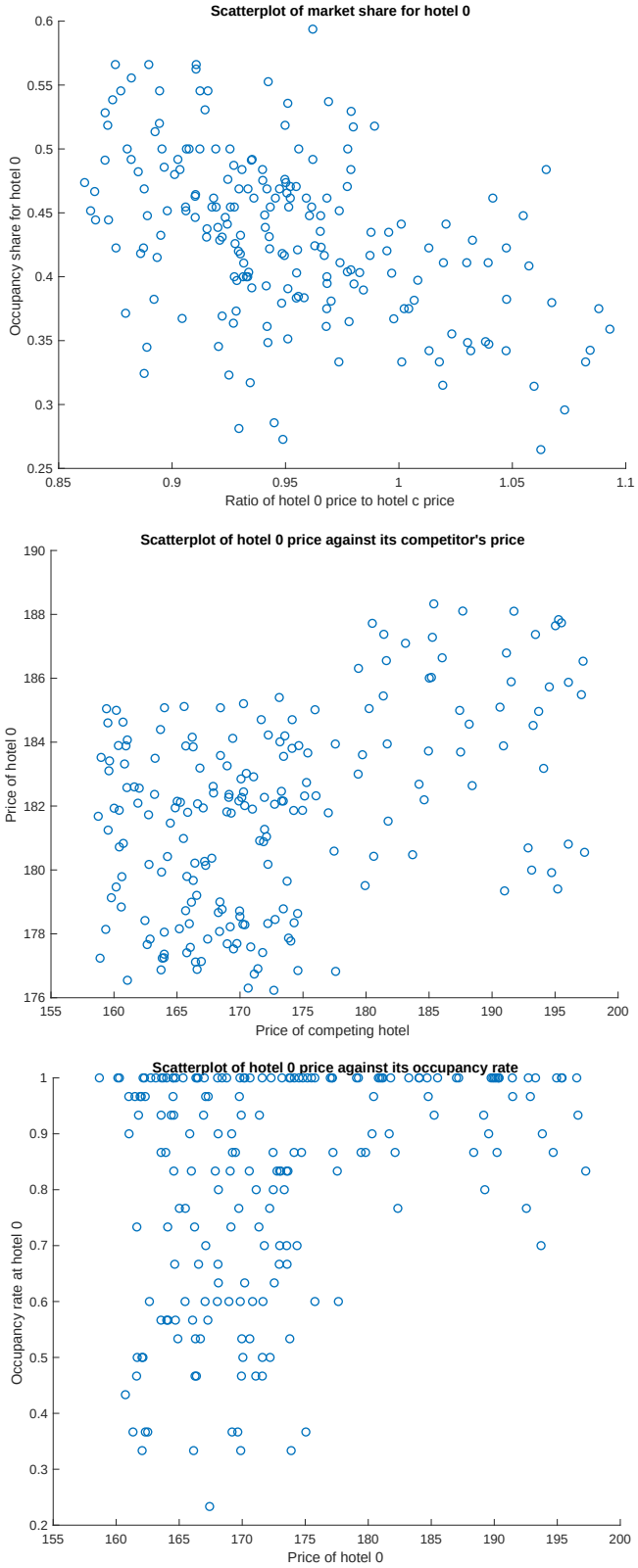
We now consider the problem of identifying consumer preferences and arrivals from potentially endogenously generated makret price and occupancy data.

Definition 1 *The structure of the hotel pricing problem consists of the objects*

$$\Gamma = \{\{g(\tau|x)|\tau \in \{1,\ldots,\mathcal{T}\}\}, \{P_\tau(l|p,x)|\tau \in \{1,\ldots,\mathcal{T}\},\ l \in \{0,1,\ldots,L\}\}, H(A|x)\},\ x \in X. \qquad (11)$$

We exclude the conditional distribution $G(p|x)$ from the elements of the structure from the problem since we wish to relax the assumptions of 1) equilibrium (the firms' set prices in accordance with a Bertrand-Nash equilibrium), and 2) optimality (firms set optimal prices, given possibly non-equilibrium

Figure 7: Simulated price scatterplots for hotel 0 and its competitor



beliefs about the prices charged by their competitors). However we consider the choice probabilities $P_\tau(l|p,x)$, the distribution of consumer types $g(x|\tau)$ and the arrival probability $H(A|x)$ as structural objects that would not change under different assumptions about how the hotels set their prices, such as if they set prices optimally and in accordance with a Bertrand Nash equilibrium. Thus, if we can identify the structural objects, we can in principle solve for a Bertrand-Nash equilibrium and compare the distribution of prices $G^*(p|x)$ arising under a Bertrand-Nash equilibrum to the distribution $G(p|x)$ that could potentially be identified under the *status quo.* Thus, G is not invariant, and depends on what assumptions we make about how the hotels set their prices.

Definition 2 *The identified objects for the hotel pricing problem are given by*

$$\Lambda = \{f(d|x,p), G(p|x)\}, \quad x \in X. \tag{12}$$

We assume that we can observe the hotels over a sufficiently long period of time where the stationarity assumption holds to consistently estimate the conditional distribution of prices given x, $G(p|x)$ so we can take this conditional distribution as "known" for purposes of the analysis of identification. Since we can consistently estimate $G(p|x)$ without imposing the assumption of equilibrium or optimality, we can be agnostic about firm behavior. Similarly, given sufficient data, we assume we can estimate the conditional distribution of occupancy given (x,p) via non-parametric methods, $f(d|x,p)$.

The identification problem concerns the question as to whether there is an invertible mapping from the identified objects Λ to the structure Γ. Define a mapping $\Psi(\Gamma)$ from the structure to the first component of the identified objects Λ by

$$f(d|x,p) = \sum_A f(d|A,x,p)H(A|x) \equiv \Psi(\Gamma), \tag{13}$$

where $f(d|A,x,p)$ is the censored multinomial distribution of hotel occupancy given the number of arrivals A that we introduced in equation (4). Note that equations (4) and (6) imply that $f(d|A,x,p)$ is itself a function of the other components for the structure Γ, i.e. the distributions over consumer types $g(\tau|x)$ and the consumer choice probabilities $P_\tau(l|p,x)$, for $\tau \in \{1,\ldots,\mathcal{T}\}$ and $x \in X$.

Definition 3 *Two structures $\Gamma \neq \Gamma'$ are observationally equivalent if $\Psi(\Gamma) = \Psi(\Gamma')$.*

Definition 4 *The hotel model with identified objects Λ is identified if there is a structure Γ satisfying $\Psi(\Gamma) = \Lambda$ and there is no other structure $\Gamma' \neq \Gamma$ that is observationally equivalent to Γ.*

3.1 Non-parametric identification

The identification problem reduces to a question on the identification of a mixture models, as can be seen in equation (13), where we are interested in "inverting" the distribution of occupancy given (x,p) given by

$f(d|x,p)$ to uniquely determine the "component distributions" $f(d|A,x,p)$ and the conditional distribution of arrivals $H(A|x)$. Actually we have a problem of identification of a *nested mixture model*, since in addition to identifying the component distributions $\{f(d|A,x,p\}$ and the mixing distribution $H(A|x)$, we are also interested in identifying the choice probabilities $\pi_l(p,x)$ and from them, the mixing distribution representation of unobserved heterogeneity in consumers given in equation (3). This is also clearly a problem of identification of mixtures, but in this case we presume knowledge of the type-unconditional choice probabilities $\pi_l(p,x)$, $l \in \{\emptyset, 0, 1, \ldots, L-1\}$ and from these identify the component probabilities $\{P_\tau(l|p,x)\}$ and the mixing distribution $g(\tau|x)$ for all possible values of (p,x).

We will assume that maximum number of consumers arriving to the market is uniformly bounded with a known upper bound (even though the actual support of $H(A|x)$ may be unknown):

Assumption 3 (Finite support) *Let $|A|(x)$ be the size of the support of the integer valued distribution $H(A|x)$. We have $|A| \equiv \max_{x \in X} |A|(x) < \infty$ where the upper bound $|A|$ is known a priori.*

Assume for the moment that we can identify the choice probabilities $\pi_l(p,x)$ of the censored multinomial representation of $f(d|A,p,x)$ given in equation (4) in the "upper level" mixture identification problem. To identify the distribution of types $g(\tau|x)$ and type-specific choice probabilities $\{P_\tau(l|p,x)\}$ in the "lower level" mixture identification problem in (3) we need to impose some additional structure and assumptions.

Assumption 4 (Mixed logit) *The utility functions for consumers given in equation (2) are linear in parameters*

$$u_\tau(l, p_l, x) = \alpha(l,x) - \beta_\tau(x) p_l, \tag{14}$$

where for the outside good, $l = 0$ we impose the normalization $u_\tau(0, p_0, x) = 0$, and we assume the error terms $\varepsilon(l)$ are standardized Type 1 extreme value (i.e. have mean zero and scale parameter 1). This implies that the type-specific choice probabilities $P_\tau(l|p,x)$ take the standard multinomial logit form

$$P_\tau(l|p,x) = \frac{\exp\{\alpha(l,x) - \beta_\tau(x) p_l\}}{1 + \sum_{l'=1}^{L} \exp\{\alpha(l',x) - \beta_\tau p_{l'}\}}. \tag{15}$$

Given this we can apply the non-parametric identification result of Fox et al. (2012) to establish that both the number of unobserved types $\mathcal{T}$ and the conditional distribution of types $g(\tau|x)$ as well as the "random coefficients" $\{\beta_\tau(x)\}$ for τ in the support of the discrete distribution $g(\tau|x)$, as well as the (non-type-specific) intercept terms $\{\alpha(l,x)\}$, $l \in \{0, 1, \ldots, L-1\}$ for each $x \in X$.

Thus, the non-parametric identification of the model hinges on whether it is possible to separately identify the component distributions $f(d|A,x,p)$ and the conditional distribution of arrivals (mixing distribution) $H(A|x)$ in equation (13). Promising recent progress on the identification of mixture models by

Kitamura and Laage (2018) suggests that this may be possible. However we cannot directly apply their key result, Proposition 6.1, since the structure of our problem is not nested within the class of mixture models that they consider. Specifically, they consider the identification of mixture models that can be written as a regression equation for an observed dependent variable y given covariates x

$$y = f(x) + \varepsilon \tag{16}$$

where the actual observations are drawn from a mixture of regression models

$$y_j = f_j(x) + \varepsilon_j, \quad j \in \{1, \ldots, J\} \tag{17}$$

with probability $\lambda_j \geq 0$ with $\sum_{j=1}^{J} \lambda_j = 1$. In this case the $\{\lambda_j\}$ are the mixing distributions and the $\{f_j(x)\}$ are the component distributions. Proposition 6.1 of Kitamura and Laage (2018) establishes the non-parametric identification of this mixture model, i.e, given knowledge of the regression $f(x)$ and the distribution of ε, they establish that the number of mixture components J and the mixing probabilities $\{\lambda_j\}$ and the component regression functions $\{f_j(x)\}$ are identified. However their result requires the error terms $\{\varepsilon_j\}$ are univariate random variables that are assumed to be continuously distributed and independent of the regressor x, and *IID* across the different types j. In addition their result relies on additional assumptions that guarantee that the regression functions $f_j(x)$ are "non-parallel" as well as other technical assumptions about the moment generating functions of the $\{\varepsilon_j\}$.

In our case we can write occupancies as a multivariate system of regressions

$$d = E\{d|p,x\} + \varepsilon \tag{18}$$

where

$$E\{d|p,x\} = \sum_{A} E\{d|A,p,x\} H(A|x) \tag{19}$$

so it is tempting to try to apply Proposition 6.1 to our setting. However the component regressions in our case are

$$d_A = E\{d|A,p,x\} + \varepsilon_A \tag{20}$$

(i.e. where the number of arrivals A index the mixture components) but the error terms in our case, $\varepsilon_A = d_A - E\{d|A,p,x\}$ are multivariate random variables that are not continuously distributed or *IID* when considered as indexed over different values of arrivals A.[2] Thus, we cannot directly apply Proposition 6.1 of Kitamura and Laage (2018) to establish the non-parametric identification of our hotel model. Further,

[2]If the capacities of the hotels are not symmetric, e.g. if $C_l \neq C_{l'}$ for $l \neq l'$, then the different components of ε_A will have different distributions, and the overall vector ε_A will have different distributions for different values of the arrivals A.

they provide counter-examples showing the mixture model (16) is non-identified when the error terms $\{\varepsilon_A\}$ are heteroscedastic, as they are in our case.

Instead we establish the identification of mixtures of censored multinomials via a direct argument. First, observe that if we fix the continuous price regressor p and the demand shifter x, we can consider the key equation (13) as nonlinear system of equations. It is actually a polynomial system of equations in $(\pi_\emptyset, \pi_0, \ldots, \pi_{L-1})$ and a linear system in $H(A|x)$ given π and (p,x). Let $|D| = \prod_{l=1}^{L}(C_i+1)$ be the size of the support of $f(d|x,p)$. Then $|D|$ indexes the number of left hand side "known values" in equation (13), whereas $\{f(d|A,x,p), H(A|x)\}$ on the right hand side are the "unknowns." Let $|A|(x)$ be an *a priori* known upper bound on the support of the number of arrivals. Without imposing any further special structure on the system of equations (13) identification would seem to be hopeless since it constitutes a system of $|D|$ equations in at most $|A|(1+|D|)$ unknowns, and thus in principle there could be far more unknowns than equations. However there is substantial *special structure* to the hotel problem in view of the fact that $f(d|A,x,p)$ takes the form of a censored multinomial distribution in (4). For fixed (p,x), this special structure reduces the problem to a nonlinear system of equations with $|D|$ equations in $L+|A|(x)-1$ unknowns. If $|D| > L+|A|(x)-1$, then equation (13) will be a an over-determined system, i.e. it will have more equations than unknowns. We can consider this to be a basic "rank condition" for identification.

Note that for fixed (p,x), if we treat the component distributions $f(d|A,p,x)$ as known, equation (13) can be viewed as a system of linear equations $f = f_A \times H$ where f_A is a matrix of dimension $|D| \times |A|$ formed with the densities $f(d|A,p,x)$ arrays as its columns. If the matrix f_A has full rank, then there is a unique mixing distribution $H(A|x)$ that solves (13) when the component distributions $f(d|A,p,x)$ are fixed at their true values. However we note that (13) is a *nonlinear* system of equations when we consider $\{H(A|x), \{\pi_l(p,x)\}, l \in \{\emptyset, 0, 1, \ldots, L-1\}\}$ as the full set of unknowns. Therefore a different argument is required to establish identification of these functions.

Our identification results below will consider two possible information structures:

- **Full information** The econometrician can observe the number of customers choosing the outside good, or the number of arrivals, or both.
- **Limited information** The econometrician cannot observe the number of customers choosing the outside good, or the number of arrivals.

We will now provide sufficient conditions for the non-parametric identification of the model under both the full information and limited information structures. We start by providing a lemma that establishes that the model is partially identified under either information structure.

Lemma *Under Assumptions* $0,\ldots,4$ *if* $C_l \geq 1$*,* $l \in \{0,\ldots,L\}$ *then the ratios of the choice probabilities,* $r_l(p,x) = \pi_l(p,x)/\pi_0(p,x)$ *are identified for* $l = 1,\ldots,L-1$ *for any* (p,x) *such that* $\pi_0(p,x) > 0$*.*

Proof Note first that if $C_l \geq 1$ for $l = 0,\ldots,L-1$, then $|D| \geq 2^L > L$ so the rank condition for identification is satisfied. Let e_l be an $L \times 1$ vector whose elements equal 0 except for element l which equals 1. Then $f(e_l|p,x)$ is the probability that hotel l has exactly 1 customer occupying one of its rooms. By assumption this probability is known and positive for each l. We have

$$f(e_l|p,x) = \pi_l(p,x)\sum_A A\pi_\emptyset^{A-1}H(A|x), \quad l \in \{0,1,\ldots,L-1\} \tag{21}$$

From equation (21) it immediately follows that the ratios $r_l(p,x)$ given by

$$r_l(p,x) = \frac{f(e_l|p,x)}{f(e_0|p,x)} = \frac{\pi_l(p,x)}{\pi_0(p,x)}, \quad l \in \{1,\ldots,L-1\} \tag{22}$$

are identified. □

We can write the choice probabilities $\pi_l(p,x)$ in terms of the identified ratios of choice probabilities, $r_l(p,x)$ as $\pi_l(p,x) = \pi_0(p,x)r_l(p,x)$ and use the fact that the choice probabilities sum to 1 to write the probability of the outside good, $\pi_\emptyset(p,x)$ in terms of $(\pi_0(p,x),\ldots,\pi_{L-1}(p,x))$, to reduce the identification problem to the solution of a system of $|D|-L$ equations in $|A|$ unknowns, $(\pi_0(p,x),\{H(A|x)\})$.

Theorem 0 *Suppose we have full information on arrivals and the outside good. Then if* $C_l \geq 1$*,* $l = 0,\ldots,L-1$ *the hotel model is non-parametrically identified.*

Proof Under full information, the hotels (and the econometrician) can observe all arrivals and all consumers who choose the outside good (though observing one enables us to deduce the other via the identity

$$A = d_\emptyset + \sum_{l=0}^{L-1} d_l. \tag{23}$$

Since arrivals are observed, it follows that $H(A|x)$ is identified for each x (since we presume for the purposes of the analysis of identification we have infinitely many observations and thus can consistently estimate the discrete distribution $H(A|x)$ from the empirical distribution). So the question of identification reduces to the identification of the probability $\pi_0(p,x)$. Define $r_0(p,x) = 1$. Since the choice probabilities sum to 1 for all (p,x), we have

$$\pi_\emptyset(p,x) = 1 - \sum_{l=0}^{L-1}\pi_l(p,x) = 1 - \pi_0(p,x)\sum_{l=0}^{L-1} r_l(p,x) \tag{24}$$

where the $r_l(p,x)$ are known functions of (p,x) by the partial identification lemma above. Let 0 be an $L \times 1$ vector of 0's, so $f(0|p,x)$ is the probability of zero occupancy in all L hotels given (p,x), which is also a known function given our assumption that $f(d|p,x)$ is identified. We have

$$f(0|p,x) = \sum_A \pi_\emptyset(p,x)^A H(A|x) = \sum_A \left[1 - \pi_0(p,x)\sum_{l=0}^{L-1} r_l(p,x)\right]^A H(A|x) \tag{25}$$

by equation (24). Note that equation (25) is a polynomial equation in $\pi_0(p,x)$ and we know it has at least 1 solution in the unit interval, where at least one root is the true value $\pi_0(p,x)$ that customers choose hotel 0. Define the polynomial $P(y) : R \to R$ by

$$P(y) = \sum_A \left[1 - y \sum_{l=0}^{L-1} r_l(p,x)\right]^A H(A|x). \tag{26}$$

Notice that $P(0) = 1$ and furthermore, we have

$$P'(y) = -\left(\sum_{l=0}^{L-1} r_l(p,x)\right)\left[\sum_A A\left(1 - y\sum_{l=0}^{L-1} r_l(p,x)\right)^{A-1} H(A|x)\right] < 0 \quad y \in [0,1]. \tag{27}$$

Since $f(0|p,x) \in (0,1)$ and we know there is one solution of the equation $P(y) = f(0|p,x)$ in the unit interval (i.e. the true probability $\pi_0(p,x)$), equations (26) and (27) imply that there is only one solution in the unit interval, i.e. $\pi_0(p,x)$ is identified, and thus the entire model $\{(\pi_\emptyset(p,x),\ldots,p_{L-1}(p,x)), H(A|x)\}$ is identified. □

In the limited information case, we do not observe the number of consumers who arrive in the hotel market, nor the consumers who choose the outside good. We can only observe the occupancy in each of the hotels, and with sufficient data, this enables us to consistently estimate $f(d|p,x)$, the joint distribution of occupancy given (p,x). Identification is more difficult in this case since we cannot directly recover the distribution of arrivals, $H(A|x)$, which was the first key step to the proof of Theorem 0 for the case where we have full information (e.g. we observe arrivals). However the intuition that when $|D| > L + |A|(x) - 1$ we have more equations than unknowns and so the rank order for identification is satisfied does not automatically lead to a proof of identification. Though we conjecture that the model is identified when this rank condition holds, at this point we require additional conditionsl to prove identification, given in Theorem 1 below.

Theorem 1 *Suppose Assumptions* $0,\ldots,4$ *hold, and* $C_l \geq 1$, $l = 0,1,\ldots,L-1$. *Further, suppose that for each* $x \in X$ *there exists a price* p *such that* $f(C|p,x) = 0$ *where* $C = (C_0,C_1,\ldots,C_{L-1})$ *is the vector of capacities of the hotels in this market that also satisfies* $\pi_\emptyset(p,x) = \pi_0(p,x)$. *Then the hotel pricing problem is identified.*

Proof: Take any demand shifter $x \in X$. Let $|C| = \sum_{l=0}^{L-1} C_l$ be the total capacity of the L hotels in this market. By our assumption that there exists a p such that $f(C|p,x) = 0$ it is clear that $|A|(x) < |C|$, where $|A|(x)$ is the largest number of arrivals in the support of $H(A|x)$. The largest number of arrivals when the demand shifter is x is given by

$$|A|(x) = \sup_d \{|d| | f(d|p,x) > 0\} \tag{28}$$

where $|d| = \sum_{l=0}^{L-1} d_l$. Thus, the maximum number of arrivals can be identified by finding the vector d in the support of $f(d|p,x)$ for which the total occupancy of the hotels in this market, $|d|$, is the largest.

Now, by our assumption that $\pi_\emptyset(p,x) = \pi_0(p,x)$, we can solve for $\pi_0(p,x)$ via the equation

$$1 - \pi_\emptyset(p,x) = 1 - \pi_0(p,x) = \pi_0(p,x) \left[\sum_{l=0}^{L-1} r_l(p,x) \right] \tag{29}$$

or

$$\pi_0(p,x) = \pi_\emptyset(p,x) = \frac{1}{1 + \sum_{l=0}^{L-1} r_l(p,x)}, \tag{30}$$

and hence the choice probabilities $(\pi_\emptyset(p,x), \ldots, \pi_L(p,x))$ are identified. Now we show how to identify $H(|A|(x)|x)$, i.e. the probability of the maximum number of arrivals $|A|(x)$ when the the demand shifter is x. First, the identification of the choice probabilities $(\pi_\emptyset(p,x), \ldots, \pi_L(p,x))$ implies that for any $A \geq 0$, the censored multinomial distribution $f(d|A,p,x)$ given in equation (4) is identified. Since $|A|(x)$ is the maximal number of arrivals in state x, then for any d satisfying $f(d|p,x) > 0$ and $|d| = |A|(x)$ we have

$$f(d|p,x) = H(|A|(x)|x) f(d|A,p,x) \tag{31}$$

so $H(|A|(x)|x)$, the probability of $|A|(x)$ arrivals in state x, is identified.

Now we show by induction that $H(A|x)$ is identified for all $A < |A|(x)$. Suppose the arrival probabilities $\{H(|A|(x)|x), H(|A|(x)-1|x), \ldots, H(A|x)\}$ are identified. We show that $H(A-1|x)$ is identified as follows. Let d_A be any joint occupancy in the support of $f(d|p,x)$ satisfying: a) $|d_A| = A$ and b) $f(d_A|p,x) > 0$. Let d_{A-1} be an occupancy vector satisfying $d_{A-1,l} = d_{A,l}$ for $l = 1, \ldots, L-1$ and $d_{A-1,0} = d_{A,0} - 1$. Then we have $|d_{A-1}| = A - 1$ and we have

$$f(d_{A-1}|p,x) = f(d_{A-1}|A-1,p,x) H(A-1|x) + \sum_{A'=A}^{|A|(x)} f(d_{A-1}|A',p,x) H(A'|x). \tag{32}$$

By our inductive hypothesis, the sum on the right hand side of equation (32) is identified. Since $f(d_{A-1}|A-1,p,x) > 0$, it follows that we can solve this equation for $H(A-1|x)$ and so it is identified as well. Thus we conclude for each $x \in X$ that $H(A|x)$ is identified.

To complete the proof, we need to show that the choice probabilities $(\pi_\emptyset(p',x), \pi_0(p',x), \ldots, \pi_{L-1}(p'x))$ are identified not only for the particular (p,x) for which $\pi_\emptyset(p,x) = \pi_0(p,x)$, but also for any p' in the support of $G(p|x)$ that may not satisfy the restriction that $\pi_\emptyset(p',x) = \pi_0(p',x)$. However by repeating the proof of Theorem 0, we see once $H(A|x)$ is identified, it follows that the choice probabilities $(\pi_\emptyset(p',x), \pi_0(p',x), \ldots, \pi_{L-1}(p'x))$ are identified for all p' in the support of $G(p|x)$. □

We believe the hotel model is identified under weaker assumptions than those assumed in Theorem 1, but we have not yet succeeded in providing a proof of this. We conclude this section by showing that

once we are able to identify the choice probabilities, we can also identify the distribution of unobserved heterogeneity $g(\tau)$ and the random coefficients $\{\beta_\tau(x)\}$ in the multinomial logit type-specific choice probabilities in equation (15).

Theorem 2 *Under the assumptions of Theorem 0 with full information, or Theorem 1 with limited information, the distribution of types $g(\tau)$ and associated random coefficients $\{\beta_\tau(x)\}$ are identified.*

Proof This result follows from the identification result of Fox et al. (2012).

3.2 Parametric approach to identification

In most empirical applications, researchers are willing to make parametric functional form assumptions for consumer preferences and arrival probabilities. We have already discussed a functional form assumption on consumer preferences leading to a random coefficients version of the multinomial logit model in the previous section. Given this, it is not a huge leap for an empirical researcher to make a convenient functional form assumption for the number of arrivals, $H(A|x)$ (e.g. truncated Poisson or negative binomial). Then, except for the non-parametric "nuisance function" $G(p|x)$ we have a fully parametric model that can be estimated by maximum likelihood. As we noted, it is not actually necessary to estimate $G(p|x)$ since we can treat the realized prices set by the hotels $\{p_t\}$ as "data" and estimate the parameters θ of the model by maximum likelihood using the likelihood function $L(\theta)$ given by

$$L_T(\theta|\{d_t,p_t,x_t\}) = \prod_{t=1}^{T}\left[\sum_A f(d_t|A,p_t,x_t,\theta)H(A|x_t,\theta)\right], \tag{33}$$

where T is the sample size and θ is a vector of parameters that includes any unknown parameters of the arrival distribution $H(A|x,\theta)$ and the finite mixture model of consumer preferences $\{\alpha_\tau(x),\beta_\tau(x),g(\tau|x)\}$. Under suitable regularity conditions, as $T\to\infty$, $\log(L_T(\theta|\{d_t,p_t,x_t\}))/T$ converges in probability to its expectation $-D(\theta,\theta^*)$, the Kullback-Leibler distance between the parameter vector θ and the true parameter vector θ^* given by

$$D(\theta,\theta^*) = \sum_x \int_p \sum_d \log\left(\frac{\sum_A f(d|p,A,\theta^*)H(A|x,\theta^*)}{\sum_A f(d|p,A,\theta)H(A|x,\theta)}\right)\sum_A f(d|p,A,\theta^*)H(A|x,\theta^*)G(p|x)K(x) \tag{34}$$

where $K(x)$ is the (discrete) distribution of the demand shifters x. Clearly we have $D(\theta^*,\theta^*)=0$. The parametric demand model is identified if the only value of θ that minimizes $D(\theta,\theta^*)$ over θ is θ^*, i.e. if and only if for any $\theta\neq\theta^*$ we have $D(\theta,\theta^*)>0$.

Let $\nabla D(\theta,\theta^*)$ be the gradient of the Kullback-Leibler distance with respect to its first argument θ when the second argument θ^* is fixed at the true value. Under standard regularity conditions it is easy to show that $\nabla D(\theta^*,\theta^*)=0$, and $\nabla^2 D(\theta^*,\theta^*)=I(\theta^*)$, where $I(\theta^*)$ is the information matrix given by

$$I(\theta^*) = \sum_x \int_p \sum_d \left(\frac{[\nabla_\theta \sum_A f(d|p,A,\theta^*)H(A|x,\theta^*)][\nabla_\theta \sum_A f(d|p,A,\theta^*)H(A|x,\theta^*)]'}{\sum_A f(d|p,A,\theta)H(A|x,\theta)}\right)G(p|x)K(x) \tag{35}$$

Via a second order Taylor series expansion, it follows that in a neighborhood of θ^* the Kullback-Leibler distance is approximately a quadratric form in $I(\theta^*)$ and centered at θ^*

$$D(\theta, \theta^*) \simeq (\theta - \theta^*)' I(\theta^*)(\theta - \theta^*), \tag{36}$$

so following the analysis of Rothenberg (1971), a sufficient condition for the *local identification* of θ^* is that $I(\theta^*)$ has full rank. However to establish global identification, we need to show that the the *only* minimizer of $D(\theta, \theta^*)$ is $\theta = \theta^*$, but as Rothenberg noted, "Unfortunately it is more difficult to obtain global results." (p. 832). Global identification can be established for some parametric families of distributions, such as expontential families. However our model is not a member of an exponential family and we are not aware of a separate general argument to establish the global identification of the model.

Though we have already proved that our model is globally and non-parametrically identified in Theorem 1, and this clearly implies the global identification of the model in the special case where $H(A|x)$ is a parametric distribution, we believe that we gain a lot of additional insight by analyzing the "information content" of different assumptions about data we observe and parametric functional form assumptions via their impact on the information matrix $I(\theta^*)$. Further, empirical researchers often do not have the benefit of global identification theorems and even when they are available, these abstract theorems that presume access to infinite amounts of data may not always be such good guides to what can be identified in practice. We will illustrate this point below.

For a model to be "well identified" in practice, an empirical researcher needs more than an abstract identification theorem: they need to check the following things to be confident that they have identified the "true model": a) they must be confident that the model they are estimating is correctly specified, b) they must guarantee that they have been able to find the global maximum of the likelihood function (or whatever statistical objective function they are using to estimate the model), and c) they need to be able to invert the information matrix to obtain the asymptotic standard errors of the maximum likelihood estimator (or compute standard errors via some other statistical criterion such as GMM or MSM) and demonstrate that these standard errors are not unacceptably large. Of course in practice, empirical researchers rarely know for sure if their models are correctly specified, though there do exist a variety of model specification tests to check this, such as the random cell Chi-square test of Andrews (1988). The most difficult task for an applied researcher to verify is that the algorithm that they used to maximize the likelihood function has indeed found the global maximum of the the likelihood. However checking identification by verifying that the information matrix is invertible is easy to do.

We illustrate the "practical identification" of our semi-parametric model of hotel pricing by calculating the information matrix for specific examples. In each example we show that the information matrix is

invertible under surprisingly weak assumptions on the degree of censoring, including an extreme case where we only observe the occupancy at a single hotel but not at its competitors, and we do not observe the outside good. We show below that we can identify the parameters of consumers' utility functions (subject to a normalization that sets the utility of outside good to 0), including the preference parameter for the competing hotel even though we only observe occupancy of hotel 0. We also show that it is possible to identify the parameters of the distribution of arrivals of customers, even though we do not observe arrivals, and thus we have no way of knowing if a customer who did not book a room at hotel 0: a) arrived, b) arrived and chose the outside good, or c) arrived and chose to book a room at a competing hotel.

However at the same time, the inverse of the information matrix provides the asymptotic variances of the parameters of the model, and we show there is a substantial loss in information due to the censoring, and this loss of information is reflected in huge increases in the asymptotic variances of the parameters. This implies that while the model is technically "identified", we require substantially more data in order to be able to estimate the parameters of the model with reasonable precision when face a situation where there is censoring in the table.

Table 5 provides the asymptotic variances of a model where $\theta \in R^5$. The five parameters of this model include θ_1, the preference "intercept" for consumers who choose hotel 0, θ_2, the preference intercept for consumers who choose to stay at the competing hotel, θ_3, the coefficient on the price of the hotel (i.e. the key parameter for deriving the demand elasticity for the two hotels), and (θ_4, θ_5) the parameters for a trinomial logit model of the arrivals of customers.

$$H(A|x,\theta_4,\theta_5) = \begin{cases} \frac{1}{1+\exp\{\theta_4\}+\exp\{\theta_5\}} & \text{if } A = x-10 \\ \frac{\exp\{\theta_4\}}{1+\exp\{\theta_4\}+\exp\{\theta_5\}} & \text{if } A = x \\ \frac{\exp\{\theta_5\}}{1+\exp\{\theta_4\}+\exp\{\theta_5\}} & \text{if } A = x+10 \end{cases} \tag{37}$$

The parameter values we chose as the true parameters θ^* in table 5 below result in the same arrival probability $H(A|x,\theta_4,\theta_5)$ and consumer choice probabilities that we used to generate the simulated arrival and hotel choice data above. Note in this case, the observed demand shifter x is fairly informative on the number of actual arrivals, and thus is an unbiased and fairly accurate predictor that helps the hotels when setting their prices.

Table 4 displays the asymptotic variances for a fully structural model of the hotel market under the assumption we have *correctly modeled* the price determination process. That is, if we had correctly guessed that prices in this market were determined by the "quasi-Bayesian-Nash" equilibrium, we could model the "cross-equation" restrictions that imply that in equilibrium the distribution of prices charged by the hotels, $G(p|x)$, becomes a parametric distribution since prices become an implicit function of the

Table 4: Asymptotic variances of parameters, one type model, cross-equation restrictions on prices

Parameter	True value	Full information	Joint occupancy data	Hotel 0 occupancy data only
θ_1	1.7	1.29	1.85	149.73
θ_2	2.05	1.19	1.68	1240.48
θ_3	0.008	0.00004	0.0005	0.001
θ_4	$\log(2)$	0.117	0.232	31.17
θ_5	0.0	0.145	0.295	121.55

Table 5: Asymptotic variances of parameters, one type model

Parameter	True value	Full information	Joint occupancy data	Hotel 0 occupancy data only
θ_1	1.7	3.96	10.52	97.89
θ_2	2.05	3.95	10.48	231.56
θ_3	0.008	0.000057	0.00012	.0003
θ_4	$\log(2)$	6	64.16	655.12
θ_5	0.0	8	132.32	976.12

model parameters θ once we impose this restriction on the likelihood function. In our case, we assumed that realized prices charged by hotels takes the form of proportional price shocks, so $p_t = z_t p(\theta)$, where z_t is an $L \times 1$ vector of *IID* random variables, uniformly distributed on the interval $(1-\delta, 1+\delta)$ where in our illustration we have chosen $\delta = .1$. This implies that $G(p|x,\theta)$ is a multivariate uniform distribution centered on the price vector $p(\theta)$ computed from the "quasi-Bayesian-Nash" equilibrium discussed above. The information matrix in this case is given by

$$\sum_x \int_{1-\delta}^{1+\delta} \cdots \int_{1-\delta}^{1+\delta} \sum_d \left(\frac{[\nabla_\theta \sum_A f(d|zp(\theta^*),A,\theta^*)H(A|x,\theta^*)][\nabla_\theta \sum_A f(d|zp(\theta^*),A,\theta^*)H(A|x,\theta^*)]'}{\sum_A f(d|zp(\theta^*),A,\theta^*)H(A|x,\theta^*)} \right) \frac{dz}{[2\delta]^L} K(x) \tag{38}$$

Table 4 should be compared to table 5, which provides the asymptotic variances of the parameters when we estimate the model via the semi-parametric maximum likelihood estimator that treats prices as "data" without the structural estimation of the quasi-Bertrand-Nash equilibrium as we did in table 4. We can see there is a cost to failing to impose the equilibrium restrictions, but in this case the cost of failing to do this is not huge in terms of the increase in the asymptotic variances of the parameters. This is good news, though it may be specific to our particular specification of uniformly distributed proportional idiosyncratic price shocks. Other specifications, such as a beta, exponential or lognormal distributions for the idiosyncratic components z_t, may result in a larger cost to failing to impose the full structure and cross-equation restrictions in the model.

The first column of table 5 provides the asymptotic standard errors (diagonal of the inverse of the

information matrix $I(\theta^*)$ under the assumption of "full information", i.e. under the hypothetical that all consumers book rooms in this market via a central web site booking agency which can record the total number of arrivals A and the choices of each of the A customers who arrives to book a room, including those who choose the outside good. The full information case is not feasible in practice due to the data limitations we discussed in section 2, but it provides a benchmark from which to evaluate the loss of information due to the censoring of the data that we observe in practice. The second column presents the asymptotic variances of the parameters when we observe the realized joint occupancies d of the two hotels, but not the number of arrivals nor the number of arriving customers A who chose the outside good. We see that the variances of the preference parameters $(\theta_1, \theta_2, \theta_3)$ nearly double, but the asymptotic variances of the arrival probability parameters (θ_4, θ_5) increase by more than 10-fold. This makes sense, as the loss of information on the number of arrivals makes it harder for us to infer the distribution of arrivals even when we have constrained the demand shifter x to be a fairly informative signal about realized demand, even in the worst case.

The final column of table 5 considers the case where we only observe d_0, the occupancy for hotel 0. Surprisingly, all model parameters remain identified even in the case where we have such limited information, though we see that there is a big cost to the reduced information: the asymptotic variances of the preference parameters θ_1 and θ_2 increase between 6 and 10 fold, though the variance of the price sensitivity parameter θ_3 only doubles. Perhaps not surprisingly, we see that the variances of (θ_4, θ_5) increase again by nearly 10-fold. Thus, a rough measure of the cost of censoring is roughly that 10 times as many observations are required to obtain estimated of the arrival distribution when we observe (d_1, d_2) compared to the full information case, and another 10 times (or a total of 100 times) as many observations are required to obtain comparable accuracy in our estimates of $H(A|x)$ if we only observe the occupancy of hotel 0, d_0.[3]

Table 6 displays what happens to our inferences when we attempt to identify two possible unobserved types of consumers. We follow the identification analysis of Fox et al. (2012) by assuming common intercepts in the discrete choice models, $\theta_1 = 1.7$ and $\theta_2 = 2.05$. Thus the two types of consumers agree on the relative attractiveness and features of the two hotels, which are captured in these choice-specific intercepts, but they have different levels of price sensitivity as captured by the parameter $\theta_3 = 0.008$ and $\theta_4 = 0.018$, where both coefficients θ_3 and θ_4 enter into the model with a negative sign, so consumers prefer cheaper hotels all other things equal. The type 1 consumers are thus less price elastic, and could

[3]Interestingly, if we make the observed demand shifteri x an unbiased but less accurate signal of arrivals A (e.g. by letting $A = x - 30$ with probability 1/4, $x = A$ with probability 1/2 and $A = x + 30$ with probability 1/4), then the greater variation in arrivals enables us to estimate its distribution *more accurately* in the sense that the ratio of the asymptotic variances of (θ_4, θ_5) in the case where we observe joint occupancy data are closer to the variances under full information.

Table 6: Asymptotic variances of parameters, two type model with common intercepts

Parameter	True value	Full information	Joint occupancy data	Hotel 0 occupancy data only
θ_1	1.7	1109	1755	13179
θ_2	2.05	1119	1772	2950
θ_3	0.008	0.0197	0.0329	0.774
θ_4	0.018	0.0496	0.0978	30.5
θ_5	$\log(3)$	19435	32407	152487
θ_6	$\log(2)$	6	96	682
θ_7	0	8	229	950

be thought of as "business customers" whereas the type 2 consumers are much more price-sensitive and can be thought of as "leisure customers". The parameter $\theta_5 = \log(3)$ determines the number of type 1 customers via the binary logit functional form, $g(\tau_1|x) = 1/(1+\exp\{\theta_5\})$ so with this value, we assume that $g(\tau_1|x) = /3/4$ of all customers are the less price elastic business customers.

The fact that even though only one quarter of all customers are more price elastic causes the hotels to lower their prices in equilibrium. For example with only a single type of customer, $p_0^*(x) = 299$ and $p_1^*(x) = 279$ when $x = 140$, i.e. on days the hotels expect to be the busiest and both expect to sell out. However with two types of consumers, the "quasi-Bertrand" prices are $p_0^*(x) = 238$ and $p_1^*(x) = 227$, respectively. From table 6 we see that there is a big cost in terms of the precision of the maximum likelihood estimator from the decision to relax the assumption of only a single type of consumer. Even though the problem is still fully parametric and the two type model only involves estimation of two extra parameters θ_4 and θ_5 in table 6, we see that the asymptotic variances of the price sensitivity coefficient $\theta_3 = 0.008$ increase 4-fold, and the variance of the price sensitivity of more price-sensitive leisure customers, $\theta_4 = 0.018$ is another 3-fold higher than that. Further, the asymptotic variance of the fraction of type 1 consumers, θ_5 is 32407! Thus, we would require over 8000 observations just to be able to have enough observations to reject the hypothseis that $\theta_5 = 0$. This represents over 22 years of daily data, and to our our knowledge it is rare to find a data set covering such a long span of time in the hotel industry.

Comparing the first and second columns of table 6 to the first two columns of table 5, we see that the "culprit" is the relaxation of the assumption that there is only a single consumer type: the full information asymptotic variances are generally 100 times higher, except for the arrival parameters (θ_4 and θ_5 in the one type specification, and θ_6 and θ_7 in the two type specification) since we have a "block diagonal" information matrix between the utility function parameters and the arrival probability parameters when we have full information on the choices of all consumers. The asymptotic variances of the two parameters determining the 3 arrival probabilities are always 6 and 8, respectively. When we have joint occupancy

Table 7: Asymptotic variances of parameters, two type model with type-specific intercepts

Parameter	True value	Full information	Joint occupancy data	Hotel 0 occupancy data only
θ_1	1.7	42008	67159	20035356
θ_2	2.05	26402	42230	11596538
θ_3	0.008	0.362	0.573	166
θ_4	1.6	471866	755987	184231109
θ_5	2.5	982880	1571907	393980171
θ_6	0.018	51.9	82.9	21056
θ_7	$\log(3)$	925449	1474897	408459472
θ_8	$\log(2)$	6	85	756
θ_9	0	8	197	1181

data as opposed to full information, clearly the asymptotic variances of these parameters increase significantly (by factors between 10 and 14 in the full information model and between 16 and 28 in the two type model with common intercepts). However the "cost of censoring" is not nearly as big as the "cost of heterogeneity" in terms of how much higher the asymptotic variances of the utility parameters increase in moving from the single type to the two type specification. The variances of θ_1, θ_2, and θ_3 are respectively, 47, 50 and 195 times higher in the two type case relative to the one type case.

There is a clear loss if information (and hence a cost in terms of higher asymptotic variance of parameter estimates) when we try to do inference using only occupancy data on a single hotel (e.g. hotel 0) compared to using the entire joint distribution of occupancy, amazingly the parameters are still identified even with such limited information from only a single hotel. Withough full occupancy data, for example, we cannot directly compute market shares that form the basis for many demand estimation strategies such as BLP, for example. In tables 5 and 6 we can see that there is certainly a big cost to trying to do inference using occupancy on only a single hotel in the market but it is technically possible to do given sufficient data. We also see that when we try to estimate a two type model with only occupancy data from a single hotel, we pay an correspondingly higher price: the variances of θ_1, θ_2 and θ_3 are respectively 307, 9 and 774 times higher than the values in the case where there is only a single type of consumer.

Tables 7 and 8 provide generally very depressing conclusions on our ability to make accurate inferences when, respectively, a) there are two types of consumers but we also allow for type-specific intercepts and not just type-specific price sensitivity parameters as in the identification analysis of Fox et al. (2012), and b) there are three types of consumers with common intercepts. Even though the information matrix is technically invertible in both of these cases, the astonomical values of the asymptotic variances tell us that inference is not realistically possible in either of these cases — at least in this specific example.

Table 8: Asymptotic variances of parameters, three type model with common intercepts

Parameter	True value	Full information	Joint occupancy data	Hotel 0 occupancy data only
θ_1	1.7	368	80339	2683479
θ_2	2.05	375	80420	11596537
θ_3	0.008	0.0004	0.882	5.4
θ_4	0.018	10.4	8123	36508
θ_5	0.025	0.009	26376	56165
θ_6	$\log(6)$	262245	927312297	2987774474
θ_7	0	911083	3744816360	10442804124
θ_8	$\log(2)$	6	103	756
θ_9	0	8	232	1181

This example may provide insight into why it is so infrequent to see any empirical study with more than three estimated types of consumers, among empirical researchers who use the Heckman and Singer (1984) semi-parametric approach for estimating models with unobserved heterogeneity. If researchers are properly calculating the standard errors using the estimated information matrix, our conjecture is that they will not report results when the standard errors of parameters are so large so that all parameters are estimated to be "statistically insignificant." Another way of looking at the problem is that the likelihood function becomes increasingly "flat" after three types of consumers have been incorporated into the model, in the sense that the incremental gain in the likelihood from adding a fourth or fifth type becomes negligible. We do not think this is a sign of some amazing empirical regularity that "there are at most 3 types of people" but rather a symptom of the severe *practical* problem of associated with identification of random coefficient models and unobserved heterogeneity, despite the identification theorem of Fox et al. (2012) that, taken literally, claims that it is possible to identify not only the number of unobserved types, but also *any* distribution of random coefficients, and to be able to distinguish unobserved heterogeneity that takes the form of a continuouus mixing distribution from one that is pure discrete (such as what we have considered here). In particular, it is essentially impossible in practice to distinguish unobserved heterogeneity that takes the form of a mixture of normals (and hence is a continuous specification of unobserved heterogeneity) from a situation where there are only a finite number of "point masses" such as which we have considered here. Intuitively, a mixture of normals where the variances of the normals in this mixture are small will constitute a very similar distribution of random coefficients as a model where the heterogeneity takes the form of a finite mixture of point masses.[4]

[4]We do not suggest that the identification result of Fox et al. (2012) is wrong but the practical difficulty of identification it could indicate that it is an "ill-posed inverse problem".

4 Conclusion

We have introduced a simplified static model of demand and showed that the demand parameters are identifiable and estimable under fairly weak assumptions even in the presence of econometric problems of endogeneity and censoring — and even when we relax standard strong maintained assumptions of optimality and equilibrium that are commonly imposed to help identify structural models. So this is good news.

Unfortunately, the conclusions the reader can take away from this analysis are rather mixed and inconclusive. On the positive side, we have established a new global non-parametric identification result for a demand model that can be regarded as a "nested mixture model" with an "upper level" mixture over an unobserved number of consumers arriving to the market, and an "lower level" mixture of different consumer types with different degrees of price sensitivity.

On the negative side, we have shown via illustrative calculations of the information matrix for parametric versions of the demand model that the degree of information decreases rapidly as the number of unobserved types in the model increases. Essentially, increasing the number of unobserved consumer types results in near collinearity in the information matrix, and this results in an exponential blow up in the asymptotic variance of coefficients that characterize the distribution of random coefficients. Thus, contrary to the conclusion of Fox et al. (2012) who view their theoretical proof of identication as "comforting to empirical researchers" p. 210 we think there is a big gap between the "practical" aspect of identification that empirical researchers confront when estimating their models and theoretical analyses of identification that abstract from most of the practical problems that empirical researchers face.

In our view, the most relevant practical indicator of the strength of identification is the asymptotic variance of an estimator. Identification problems will naturally manifest themselves in unreasonably large estimated standard errors for parameters, which in turn is symptomatic of an estimation criterion that is virtually "flat" in the parameters, at least along certain directions of the parameter space. Though abstract theoretical analyses of identification such as provided in this paper or in Fox et al. (2012) can serve as useful points of departure, ultimately the burden is on the empirical researcher to show that their parameter estimates are the unique global optimizer of the statistical objective function (e.g. likelihood function). This is often very difficult, if not impossible to do. In most complex nonlinear models, about the best a researcher can do is attempt a thorough search of the parameter space and show there no other parameter values that result in comparable fit. Short of that, empirical researchers are justified in claiming their models are identified if: a) they can succeed in estimating them, and b) they can calculate the standard errors accurately and show that they are not unreasonably large.

References

Andrews, D. W. K. (1988). Chi-square diagnostic tests for econometric models: Theory. *Econometrica*, *56-6*, 1419–1453.

Benoit, J., & Krishna, V. (1987). Dynamic duopoloy: Prices and quantities. *Review of Economic Studies*, *54*, 23–36.

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, *63*(4), 841–890.

Cho, S., Lee, G., Rust, J., & Yu, M. (2018). *Optimal dynamic hotel pricing.*

Davidson, C., & Deneckere, R. (1990). Excess capacity and collusion. *International Economic Review*, *31*(3), 521–541.

Ezrachi, A., & Stucke, M. E. (2016). *Virtual competitiion: The promise and perils of the algorithm-drive economy*. Harvard University Press.

Fox, J., Kim, K., Ryan, S., & Bajari, P. (2012). The random coefficients logit model is identified. *Journal of Econometrics*, *166*, 204–212.

Hall, G., & Rust, J. (2019). *Econometric methods for endogenously sampled time series: The case of commodity price speculation in the steel market* (Tech. Rep.). Georgetown University.

Harrington, J. (2017). *Developing competition law for collusion by autonomous price-setting agents* (Tech. Rep.). Wharton School of Business.

Heckman, J. J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, *52*(2), 271-320.

Kitamura, Y., & Laage, L. (2018). *Nonparametric analysis of finite mixtures.*

MacKay, A., & Miller, N. (2018). *Demand estimation in models of imperfect competition.*

McAfee, P., & te Veld, V. (2008). Dynamic pricing with constant demand elasticity. *Production and Operations Management*, *17*(4), 432–438.

McFadden, D. L. (1998). A method of simulated moments for estimation of discrete choice models without numerical integration. *Econometrica*, *57*(5), 995–1026.

Merlo, A., Ortalo-Magne, F., & Rust, J. (2015). The home selling problem: Theory and evidence. *International Economic Review*, *56*(2), 457–484.

Phillips, R. L. (2005). *Pricing and revenue optimization*. Stanford University Press.

Rothenberg, T. (1971). Identification in parametric models. *Econometrica*, *39-3*, 577–591.

Rust, J. (2019). *Has dynamic programming improved decision making?*