http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 64, No. 9, September 2018, pp. 4118–4136
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Price to Compete . . . with Many: How to Identify Price Competition in High-Dimensional Space

Jun Li,[a] **Serguei Netessine,**[b] **Sergei Koulayev**[c]

[a] Ross School of Business, University of Michigan, Ann Arbor, Michigan 48104; [b] The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [c] Consumer Financial Protection Bureau, Washington, DC 20006

Contact: junwli@umich.edu, http://orcid.org/0000-0002-9237-9147 (JL); netessin@wharton.upenn.edu, http://orcid.org/0000-0002-3587-3894 (SN); sergei.koulayev@gmail.com (SK)

Received: October 3, 2015
Revised: June 6, 2016; December 21, 2016
Accepted: March 23, 2017
Published Online in Articles in Advance: September 14, 2017

https://doi.org/10.1287/mnsc.2017.2820

Abstract. We study price competition in markets with a large number (in the magnitude of hundreds or thousands) of potential competitors. We address two methodological challenges: simultaneity bias and high dimensionality. Simultaneity bias arises from joint determination of prices in competitive markets. We propose a new instrumental variable approach to address simultaneity bias in high dimensions. The novelty of the idea is to exploit online search and clickstream data to uncover customer preferences at a granular level, with sufficient variations both over time and across competitors in order to obtain valid instruments *at a large scale*. We then develop a methodology to identify relevant competitors in high dimensions combining the instrumental variable approach with high-dimensional $l-1$ norm regularization. We apply this data-driven approach to study the patterns of hotel price competition in the New York City market. We also show that the competitive responses identified through our method can help hoteliers proactively manage their prices and promotions.

History: Accepted by Vishal Gaur, operations management.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2017.2820.

Keywords: price competition • simultaneity bias • high dimensionality • industries: hotel–motel

1. Introduction

Many markets rely on a large number of competing sellers to provide goods or services. The hospitality industry is one of them. At major travel destinations, hundreds of branded and nonbranded hotels compete with a full range of quality tiers including budget, economy, upscale, and luxury. Competition among hotels also intensifies as information—room rates in particular—becomes transparent within a few clicks online. The same trend applies to many other industries and markets mediated by online technologies, such as retailing, insurance, mortgage, and real estate. For example, over two million third-party merchants compete on Amazon.com, where their sales reached two billion units in 2014—that is, more than 40% of all items sold on Amazon (Bensinger 2015).

In markets with a large number of competitors, it is unlikely that every firm would react to the actions (e.g., price changes) of every other competitor. Managers may have only limited awareness of the identities and actions of the other players in the market. Furthermore, the costs of tracking, optimizing, and responding to the market can be substantial such that it is not worthwhile to compute best responses to competitors whose influence is perceived to be weak. Understanding such potentially sparse competition is important—a hotel manager who anticipates competitive responses can manage his pricing and promotional decisions more proactively. Classic models of competition in economics study either only a handful of players (i.e., duopoly or oligopoly) or an infinite number of players (i.e., perfect competition), such that the price equals the marginal cost and the profit is zero. Limited attention is given to the intermediate case, in which the number of players is larger than a few, but the competition is not perfect. In such markets, sellers find ways to differentiate themselves, though to a limited extent. Some models localize competition along a single dimension (e.g., Hotelling line, Salop circle). However, when sellers compete on multiple dimensions such as geography, amenities, quality of service, brand name, price, etc., where to draw the boundaries of competition is not obvious.

Using modern machine learning methods, we show how vast amounts of data generated online might be employed to characterize price competition in such a high-dimensional space. In particular, we address the following questions in this paper: Whose price is influenced by whom? What non-price factors determine the boundaries of price competition? How can firms use competitive information proactively to manage their prices and promotional decisions? The solution we propose is to construct a system of price equations, where a hotel's price is a function of all other

hotels' prices, and then identify the ones with *economically* and *statistically* significant coefficients. Although the basic approach sounds straightforward, it is subject to two key challenges: simultaneity bias and high dimensionality.

The simultaneity bias comes from joint determination of prices. A seller's price is a function of its competitors' prices, and vice versa. Since prices are jointly determined by a system of equations, without adjusting for simultaneity bias, the estimates would be biased. In other words, what we measure can be merely correlations rather than direct price influences. The common solution to simultaneity bias is to identify valid instruments, or, in this case, variables that are correlated with one price but with none of the others (not through unobservables). A valid instrument is often difficult to come up with even for a single equation but more so when there are hundreds or thousands of competitors (equations), and we need one instrument for each variable. Also note that the instrument needs to change as rapidly as prices do in order to preserve sufficient variation from the original variable. In the hotel industry, for example, prices change on a weekly or daily basis. Most commonly used instruments either do not exhibit sufficient intertemporal variation (such as cost-based instruments) or are not seller-specific (such as local events or weather) to be orthogonal to all other prices. Lagged prices, although effective in other settings (Tereyağoğlu et al. 2018), are not good instruments in our setting because of serial correlation in demand across both travel dates and booking dates.

The novel idea that we propose is to exploit demand signals observed through online search and clickstream data. Consumers searching online for hotels engage in various search actions to condense the consideration set (Koulayev 2014) in order to match their own preferences. When a customer makes a search request—say, one week prior to the desired travel date—and filters three-star hotels close to Times Square, her specific preferences are revealed. As a result, we observe which hotels enter her consideration set. By aggregating such demand for each hotel on each booking date and travel date, we obtain a hotel-specific measure of demand. For example, if there is a convention near the Financial District, we would observe it through higher than normal exposures of hotels in that region and for that date. Consumers' preferences and search strategies are often so variant that even two nearby hotels receive different demand levels as a result of differences on other search dimensions (brands, stars, amenities, etc.). Moreover, as we observe the sequence and types of search actions, this measure can be constructed such that it only contains non-price-related exposures, such that the number of exposures received by a hotel is not correlated with the price it offers.

The second challenge is high dimensionality. Given a large number of explanatory variables, chances are there will be only a few that are truly relevant. Trying to obtain an exact estimate for each variable is neither efficient nor desired. It is not efficient because collinearity will become a pressing concern, and it will affect the statistical significance of many estimates. It is not desired because one may end up overfitting a model with many variables but not discovering which are truly relevant. In other words, the methodology should strike a balance between a reasonable model fit and screening out the nonrelevant variables. Various regularization methods have been developed for the purposes of model prediction and variable selection, in which nonsignificant (or nonrelevant) variables are assigned zero coefficients. The objective function penalizes the nonzero estimates, and by selecting the right penalty parameter, one can achieve different levels of model sparsity. However, most often, regularization is used only to identify correlation and not causality. In this paper, we try to achieve both—regularization and causality—by using the instrumental variable approach in high-dimensional space.

We test the performance of our proposed method on synthetic data first. We note that because an instrumental variable only uses partial variation from the original variable, when it comes to variable selection, it could be the case that a variable is not selected not because it is irrelevant but because it contains only limited variation. Indeed, we find that an instrumental variable approach may perform poorly when combined with least absolute shrinkage and selection operator (LASSO) variable selection in high dimensions, when the endogenous variable is replaced with the first-stage predicted value (two-stage predictor substitution, or 2SPS). However, if we keep the endogenous variable while including the first-stage predicted residual (two-stage residual insertion, or 2SRI), we can achieve much higher variable selection accuracy—higher precision and higher recall at the same time. This is because with 2SRI, the original variable and the predicted residual are not necessarily selected or dropped simultaneously. When the residual is not selected but the original variable is, we retain the full variation from the original variable. Note that in this case, the endogeneity for that specific variable is not a pressing concern since the residual turns out to be nonsignificant (Hausman 1978). Finally, we used residual bootstrap to achieve higher accuracy in variable selection and in the estimation of the confidence intervals with potentially weak instruments.

We applied our method to the New York City hotel market and found several interesting patterns of price competition. First, engagement in competition-based revenue management is prevalent across brands (or non-brands) and across all quality tiers. It explains

30.2% of within-hotel price variation, in contrast to 22.3% explained by demand based variables, including booking date and travel date characteristics, advance purchases, and exposures observed through online search. This finding points out that more attention should be paid to revenue management under competition than we currently observe in the literature. Second, when choosing whose prices to follow, branded hotels are more restricted by whether another hotel is within the same quality tier and less by how distant it is compared with independent hotels. Budget and luxury hotels are more confined by quality boundaries and less by geographical boundaries compared with economy and upscale hotels. Third, when hoteliers react to competitors' prices, they tend to miss those potential competitors who are geographically farther away or are of different quality tiers (i.e., different star levels). Finally, we illustrate how such competitive response information can help hoteliers improve their revenue management practice. Using our algorithm, a hotel manager will know which other hotels benchmark their prices against his own. Therefore, when deciding what promotions to offer, for example, he would be able to anticipate and strategically take into account key competitors' responses, to achieve more realistic and accurate growth targets. We conclude by observing that our methodology can be applied to many other industrial settings with large numbers of sellers or firms and where online search and comparison are present.

2. Literature Review

A majority of competitive models can be cast in the framework of duopoly or oligopoly models, where everyone competes with everyone else directly, or models with perfect competition, where the operating profit is driven to zero (see Tirole 1988 for examples). A few models allow for localized rivalry where players only compete directly with neighbors—for example, one-dimensional spatial models (Hotelling 1929, Salop 1979, Gabszewicz and Thisse 1979). In such models, demand of each player is directly affected only by the prices of its neighbors, one on each side, but not *directly* by prices of all other more distant players. The idea of localized competition is generalizable to multidimensional space (Anderson et al. 1989). The difficulty, though, comes with the definition of market boundaries. Feenstra and Levinsohn (1995) estimate an oligopoly pricing model when products are multidimensionally differentiated, using data from the U.S. automobile market, where the market boundaries are defined by a weighted Euclidean distance of the product characteristic space where the weights are estimated. Pinkse et al. (2002) use a semiparametric approach to distinguish global versus local competition and find that competition in the U.S. wholesale gasoline market is highly localized. Olivares and Cachon (2009) estimate how local competition affects dealer inventory in the automobile industry.

Note that when the number of competitors is large, models with global competition poorly represent real-world decision processes. No market participant possesses the cognitive capability to evaluate all options, either products or competitors, nor is it necessary. Rather, firms often apply simplified decision rules to reduce dimensionality.

To understand how firms compete when the market is fragmented and when the number of potential competitors is large, we model the competitive responses directly and let the data determine which are the key competitors when it comes to price competition. We adopt the LASSO method to select competitors whose prices directly influence the price of each other hotel in our sample. Proposed by Tibshirani (1996), LASSO gained its popularity for high-dimensional estimation problems because of its statistical accuracy for prediction and variable selection coupled with its computational feasibility. Various theories have been developed since then to establish its asymptotic consistency and to expand its applications to nonlinear settings (see Bühlmann and van de Geer 2011 for an overview). Recently, it has also been applied in operations management contexts. Ang et al. (2016) propose the Q-LASSO method for emergency department wait time prediction, which combines statistical learning with fluid model estimators. Rudin and Vahn (2016) apply regularization to solve the optimal order quantity in the newsvendor problem. Ryzhov et al. (2015) identify designs that exert significant impacts on fundraising outcomes using the LASSO method.

While most theories and applications of LASSO are primarily concerned with correlations rather than causality, recently LASSO has also been used to facilitate estimation with endogeneity concerns; see Belloni et al. (2012). There are two critical differences between our approach and theirs. First, the objective of the proposed use of LASSO is different. Unlike Belloni et al. (2012), we use LASSO to select potential *endogenous variables* rather than to select *instruments*. In our setting, we face a large number of potentially endogenous variables, and our primary objective is to identify which ones are economically and statistically relevant (i.e., most prominent competing sellers). In Belloni et al. (2012), however, the authors face only a small number of endogenous variables but a large number of instruments. Therefore, their objective is to select which instruments to use rather than which endogenous variables to enter into the main equation. In other words, the objective of LASSO is to achieve relevance

in our paper, while the objective is to achieve estimation precision in theirs. The second critical difference is that while Belloni et al. (2012) has a single equation, we have a system (and a large number) of simultaneous equations. This simultaneity is a major cause of endogeneity.

Finally, our research is rooted in the revenue management and pricing literature, which took off in the 1980s with the deregulation in the airline industry (Boyd 2007). Principles and algorithms of dynamic management of inventory controls and prices (see, e.g., Belobaba 1989, Gallego and van Ryzin 1994) have been adopted by many industries—hotels, car rentals, casinos, retailing, to name a few (see more examples in Talluri and van Ryzin 2005)—albeit with varying levels of sophistication. All these industries are highly competitive, including the U.S. airline industry, with years of unprofitable performances, and the fragmented hotel industry, where the 50 largest companies generate only 45% of the revenue, according to Hoovers First Research 2017 Lodging Industry Report. Competitive pricing has only slowly gained traction in the large body of revenue management (RM) and pricing literature. Netessine and Shumsky (2005) study a stylized model of revenue management model with two players and arrive at a counterintuitive conclusion that competition in revenue management leads to higher prices because firms allocate more capacity for high-price customers under competition. Adida and Perakis (2010) analyze the joint decision of prices and inventory under competition using a continuous-time deterministic differential game. Martinez-de-Albéniz and Talluri (2011) provide a closed-form solution to the equilibrium price paths for a duopoly, in which sellers engage in Bertrand competition in all states. Gallego and Hu (2014) study dynamic pricing competition of perishable products under a deterministic arrival process, which also sheds light on the stochastic problem with random demand arrivals.

In terms of context, our paper is perhaps closest to Lederman et al. (2014), who use a random utility model to identify heterogeneous consumer segments using data from the hotel industry. Their focus is demand modeling, and our focus is to understand the competitive response patterns. While empirical research in revenue management is burgeoning, most of these studies focus on analyzing a single firm's problem (see, e.g., Vulcano et al. 2010, Newman et al. 2014, Li et al. 2014). There are very few empirical works that directly address competition in the RM setting. Fisher et al. (2018) develop and validate with field experiments a best-response pricing algorithm to competition in online retailing. However, competitive pricing in industries with fixed capacity can be more challenging because of the combination of competition and dynamic management of prices.

3. A Model of Price Competition

In this section, we first illustrate the issue of simultaneity bias and how we address it through the use of instrumental variables. We then discuss how to identify the relevant competitors among a large number of potential competitors while accounting for simultaneity bias.

3.1. A Model with N Competitors

Consider a market served by N hotels. Let P_{ijt} denote the price per night charged by hotel i for a stay on travel date j booked at booking date t.[1] A hotel determines its nightly price based on when the travel will happen, when the reservation is made, as well as competitor prices. Specifically,[2]

$$\begin{aligned}
P_{1jt} &= \alpha_1 + \beta_{12}P_{2jt} + \beta_{13}P_{3jt} + \cdots + \beta_{1N}P_{Njt} + X_{jt}\gamma_1 + \varepsilon_{1jt},\\
P_{2jt} &= \alpha_2 + \beta_{21}P_{1jt} + \beta_{23}P_{3jt} + \cdots + \beta_{2N}P_{Njt} + X_{jt}\gamma_2 + \varepsilon_{2jt},\\
&\cdots\\
P_{Njt} &= \alpha_N + \beta_{N1}P_{1jt} + \beta_{N2}P_{2jt} + \cdots + \beta_{N,N-1}P_{N-1,jt}\\
&\quad + X_{jt}\gamma_N + \varepsilon_{Njt},
\end{aligned}$$

where X_{jt} denotes the characteristics of the travel date j and the booking date t such as, for example, travel date day of week, booking date day of week, the number of days booked in advance, and the total market traffic (approximated by the number of unique users searching for accommodations for the travel date on the booking date). Different hotels may price these attributes differently. In other words, γ_i is firm-specific. For example, a hotel catering mostly to leisure travelers may not differentiate its prices significantly based on day of the week, whereas a hotel with a mix of business and leisure customers may charge very different prices on weekdays versus weekends to segment the market. The term ϵ_{ijt} represents idiosyncratic shocks, either demand or supply shocks, that may affect hotel i's price for travel date j on booking date t. For example, price can be high if there is a local convention scheduled to happen on a particular travel date or if certain rooms are blocked because of special events. Note that these shocks, ϵ_{ijt}'s, are likely correlated across hotels as a result of macroeconomic conditions, regional events, and demand competition among similar hotels.

We use the linear specification to approximate the competitive response function for its computational efficiency in high dimensions. In Online Appendix EC.1, we show that linear best-response function can arise with profit maximizing sellers facing a linear demand function; when the demand function is nonlinear, such as in the multinomial logit model, the best-response function can be well approximated by Taylor linear expansion (approximation error ranges from 0.04% to 0.6%). We then provide empirical evidence that linear specification provides good approximation

in our data. Such specification is also consistent with the industry practice of using the weighted average of competitor prices as the benchmark price. The results are also consistent when we use logarithms of prices.

We would also like to note that, compared with other more sophisticated equilibrium models, which often require assumptions that every player has complete knowledge of everyone else's payoff functions and that all players are rational, our reduced-form modeling approach does not preimpose such assumptions on how players compete. It rather relies on data to help us understand the underlying competition patterns in such a fragmented and differentiated market. Moreover, the linear additive structure is particularly appealing to achieve the dual purpose of causal inference and high-dimensional variable selection, as we shall show later.

The ordinary least square regression in this context is subject to classic simultaneity bias (Wooldridge 2010). A typical solution is to identify factors that enter only one equation but not others. Suppose we observe E_{ijt}, $\forall\, i$; each enters one equation only:

$$\begin{aligned}
P_{1jt} &= \alpha_1 + \beta_{12}P_{2jt} + \beta_{13}P_{3jt} + \cdots + \beta_{1N}P_{Njt} \\
&\quad + X_{jt}\gamma_1 + \delta_1 E_{1jt} + \varepsilon_{1jt}, \\
P_{2jt} &= \alpha_2 + \beta_{21}P_{1jt} + \beta_{23}P_{3jt} + \cdots + \beta_{2N}P_{Njt} \\
&\quad + X_{jt}\gamma_2 + \delta_2 E_{2jt} + \varepsilon_{2jt}, \\
&\cdots \\
P_{Njt} &= \alpha_N + \beta_{N1}P_{1jt} + \beta_{N2}P_{2jt} + \cdots + \beta_{N,N-1}P_{N-1,jt} \\
&\quad + X_{jt}\gamma_N + \delta_N E_{Njt} + \varepsilon_{Njt}.
\end{aligned}$$

One can then use the firm-specific shock E_{ijt} as an instrument for the corresponding price P_{ijt} in all price equations other than that of the focal hotel i. All N equations would thus be identified. Unfortunately, firm-specific shocks are usually only observable to firms themselves but rarely to researchers—and in particular, rarely *at a large scale*. For instance, one could identify a handful of location-specific events that affect only a selected group of hotels (Lederman et al. 2014) but not all of them.

The novel idea that we propose in this paper exploits demand signals observed through online search and clickstream data. With the increasing availability of online search and clickstream data, particularly data that come from third-party platforms hosting exhaustive lists of hotels competing in a market, we can observe demand variations at a more granular level. For instance, we could observe how many travelers searched for three-star hotels around Times Square in New York City one week before the desired travel date. As a consequence, we could find how many customers considered or were exposed to a specific hotel in their search results. It is not necessary for a hotel manager to observe the clickstream data in order for the instrument to be valid. Technically, the instrument only needs to be correlated with the overall demand signal that hotel managers observe.

Note that instruments E_{ijt} are likely correlated across hotels. However, such correlation does not undermine the validity of the instruments. The key requirement for E_{ijt} to be a valid instrument for P_{ijt} is that it does not affect the other firms' prices $P_{kjt}, k \neq i$ *directly*. For example, if E_{ijt} affects hotel k's price P_{kjt} indirectly through correlation with hotel k's shock E_{kjt}, the instrument is still valid, because E_{kjt} is observable and hence can be controlled. In mathematical terms, we require *conditional* independence: $\mathrm{Cov}(E_{ijt}, \epsilon_{kjt} | E_{kjt}) = 0$, $\forall k \neq i$, $\forall i$. That is, E_{ijt} is uncorrelated with ϵ_{kjt}, $k \neq i$ once E_{kjt} is partialed out (see chap. 4 in Wooldridge 2010 for proof). We discuss in detail in Online Appendix EC.3 what conditional independence requires in our context.

To estimate the model, one can use a two-stage least square approach. In the *first* stage, we regress the endogenous variable P_{ijt} on instruments E_{ijt} and exogenous variables X, and then we can compute the predicted values $\hat{P}_{ijt}(X_{jt}, E_{ijt})$ and the predicted residual $\hat{\varepsilon}_{ijt}(X_{jt}, E_{ijt})$:

$$\text{Stage 1: } P_{ijt} = \alpha_i + X_{jt}\gamma_i + \delta_i E_{ijt} + \varepsilon_{ijt}, \quad \forall\, i. \tag{1}$$

In the *second* stage, there are two options to estimate the price coefficients. One is to replace the original prices $P_{-i,jt}$ in the regressors, which denotes all prices other than firm i's price, with the predicted prices $\hat{P}_{-i,jt}(X_{jt}, E_{-i,jt})$ from the first stage. This method is called 2SPS, for two-stage predictor substitution. Alternatively, one could include both the original prices $P_{-i,jt}$ and the predicted residuals $\hat{\varepsilon}_{-i,jt}(X_{jt}, E_{-i,jt})$ from the first stage. This method is called 2SRI, for two-stage residual insertion. Either approach will give us unbiased estimates of β_{ijt} (Hausman 1978):

$$\begin{aligned}
\text{Stage 2 (2SPS): } P_{ijt} &= \alpha_i + \hat{P}_{-i,jt}\beta_{i,-i} + X_{jt}\gamma_i \\
&\quad + \delta_i E_{ijt} + \epsilon_{ijt}, \quad \forall\, i, \\
\text{Stage 2 (2SRI): } P_{ijt} &= \alpha_i + P_{-i,jt}\beta_{i,-i} + X_{jt}\gamma_i + \delta_i E_{ijt} \\
&\quad + \hat{\varepsilon}_{-i,jt}\theta_{i,-i} + \epsilon_{ijt}, \quad \forall\, i,
\end{aligned}$$

where

$$\begin{aligned}
P_{-i,jt} &= [P_{1jt}, P_{2jt}, \ldots, P_{i-1,jt}, P_{i+1,jt}, \ldots, P_{Njt}], \\
\hat{P}_{-i,jt} &= [\hat{P}_{1jt}, \hat{P}_{2jt}, \ldots, \hat{P}_{i-1,jt}, \hat{P}_{i+1,jt}, \ldots, \hat{P}_{Njt}], \\
\hat{\varepsilon}_{-i,jt} &= [\hat{\varepsilon}_{1jt}, \hat{\varepsilon}_{2jt}, \ldots, \hat{\varepsilon}_{i-1,jt}, \hat{\varepsilon}_{i+1,jt}, \ldots, \hat{\varepsilon}_{Njt}], \\
\beta_{i,-i} &= [\beta_{i1}, \beta_{i2}, \ldots, \beta_{i,i-1}, \beta_{i,i+1}, \ldots, \beta_{iN}], \quad \text{and} \\
\theta_{i,-i} &= [\theta_{i1}, \theta_{i2}, \ldots, \theta_{i,i-1}, \theta_{i,i+1}, \ldots, \theta_{iN}].
\end{aligned}$$

3.2. Variable Selection in High Dimensions

We rewrite the system of equations for the second stage in matrix form, depending on whether we use 2SPS or 2SRI,

$$P = A + \hat{P}B^T + X\Gamma + E \circ \Delta + \epsilon$$

or

$$P = A + PB^T + X\Gamma + E \circ \Delta + \hat{\epsilon}\Theta + \epsilon,$$

where $\circ$ denotes the entrywise product. Here, $P = [P_1, P_2, \ldots, P_N]$, $\hat{P} = [\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_N]$, $E = [E_1, E_2, \ldots, E_N]$, $\hat{\epsilon} = [\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_N]$, $\Gamma = [\gamma_1, \gamma_2, \ldots, \gamma_N]$, and

$$A = \begin{bmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_N \\ \alpha_1 & \alpha_2 & \ldots & \alpha_N \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_1 & \alpha_2 & \ldots & \alpha_N \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & \beta_{12} & \beta_{13} & \ldots & \beta_{1N} \\ \beta_{21} & 0 & \beta_{23} & \ldots & \beta_{2N} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \beta_{N1} & \beta_{N2} & \beta_{N3} & \ldots & 0 \end{bmatrix},$$

$$\Delta = \begin{bmatrix} \delta_1 & \delta_2 & \ldots & \delta_N \\ \delta_1 & \delta_2 & \ldots & \delta_N \\ \ldots & \ldots & \ldots & \ldots \\ \delta_1 & \delta_2 & \ldots & \delta_N \end{bmatrix},$$

$$\Theta = \begin{bmatrix} 0 & \theta_{12} & \theta_{13} & \ldots & \theta_{1N} \\ \theta_{21} & 0 & \theta_{23} & \ldots & \theta_{2N} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \theta_{N1} & \theta_{N2} & \theta_{N3} & \ldots & 0 \end{bmatrix}.$$

Here, X is the matrix that denotes the characteristics of the travel date, the booking date, and the total market traffic. Let K denote the number of variables in X, let J denote the number of travel dates, let T denote the number of booking dates, and let N denote the number of firms. We note that P, $\hat{P}$, E, ϵ, A, and Δ are all JT by N matrices, Γ is a K by N matrix, and B and Θ are N by N matrices.

We are particularly interested in estimating the competition matrix B. When the number of competitors, N, is large, the number of parameters to be estimated increases quadratically. For example, if our universe contains 200 competitors, then there will be $200 \times (200 - 1) = 39{,}800$ parameters to be estimated, which is computationally infeasible even without other controls.

It is reasonable to expect, however, that the competition matrix B will likely be sparse for two reasons. First, we notice sparse demand substitution patterns using online hotel search and clickstream data (see Online Appendix EC.2 for empirical evidence). A sparse demand model indicates that one hotel's price change will likely affect a handful of competitors who compete closely. Therefore, there is limited need to respond to those hotels whose prices have almost zero impact on one's demand. Second, the costs of tracking—and more importantly, optimizing and responding to a large number of market players—can be prohibitively high. As a result, hotel managers focus their responses on a few main competitors: we commonly observe that hotels select 5–10 other hotels to be their main competition for price benchmarking, as shown by competition set surveys.[3] Despite the likely sparse competitive response patterns, prices can still be correlated with the overall market demand. Therefore, we include the total market traffic in the covariate matrix X.

To select the relevant competitors for each hotel, we will use the $l - 1$ norm regularization LASSO, developed by Tibshirani (1996). LASSO is a widely used method for high-dimensional estimation problems because of its statistical accuracy for prediction and variable selection coupled with its computational feasibility (Bühlmann and van de Geer 2011).[4] For each one of the N equations, we solve one of the following optimization problems, depending on whether we use 2SPS or 2SRI:

$$\min_{\alpha_i, \gamma i, \beta_i, \delta_j} \left(\frac{1}{2JT} \sum_{j,t=1}^{J,T} \left(P_{ijt} - \alpha_i - \hat{P}_{-i,jt}\beta_{i,-i}^T - X_{jt}\gamma_i - \delta_j E_{jt}\right)^2 + \lambda \sum_{n=1, n\neq i}^{N} |\beta_{in}| \right), \quad \forall i, \quad (2)$$

or

$$\min_{\alpha_i, \gamma i, \beta_i, \delta_j} \left(\frac{1}{2JT} \sum_{j,t=1}^{J,T} (P_{ijt} - \alpha_i - P_{-i,jt}\beta_{i,-i}^T - X_{jt}\gamma_i - \delta_j E_{jt} - \hat{\varepsilon}_{-i,jt}\theta_{i,-i}^T)^2 + \lambda \sum_{n=1, n\neq i}^{N} (|\beta_{in}| + |\theta_{in}|) \right), \quad \forall i, \quad (3)$$

where λ is a penalty or tuning parameter. The optimization is convex, which enables efficient estimation of parameters. By varying λ, one can control the level of sparsity of the model. It is, in general, a difficult task to choose the appropriate amount of regularization to select true relevant covariates. Various approaches have been proposed. In Section 4.1, we will discuss in detail how to best choose the tuning parameter for our purpose.

4. Fine-Tuning Estimation Using Simulation Data

The objective of this section is to demonstrate the properties of the proposed estimator and to fine-tune the estimation procedure using simulated data. Specifically, to achieve the dual purpose of causal inference and high-dimensional variable selection, we address the following key points in the estimation: (1) selection of the tuning parameter, (2) method to implement instrumental variable estimation in high dimensions (2SPS or 2SRI), and (3) impact of the strength of the instrument and how to adjust for potentially weak instruments. We address these questions using simulated data where we know the true model parameters. In the simulation, we adopt the same empirical specification as introduced in the previous section to ensure consistency and generalizability of the results. These results will then guide our estimation strategy on the real data.[5]

4.1. Selection of the Tuning Parameter

As we can see from the LASSO formulation, the tuning parameter controls the accuracy and sparsity of the model. Various approaches have been proposed for selecting the tuning parameter. In practice, a commonly used approach is cross-validation (CV) to select a reasonable tuning parameter λ that minimizes the cross-validated mean squared error. Cross-validation is effective when used for the purpose of prediction. However, it may select too many variables when the primary interest is to select variables that are economically relevant. One approach to correct this overestimation tendency is to use Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Both AIC and BIC balance the goodness of fit of the model and its complexity, while BIC penalizes complexity more strongly than AIC. It is shown that BIC can identify the true model consistently, while methods based on cross-validation cannot (Wang et al. 2007).

Another approach is the two-step adaptive LASSO, where adaptive weights are used for penalizing different coefficients in the $l-1$ penalty (Zou 2006). Specifically, we first conduct the regular LASSO as in Equations (2) and (3), and we obtain an initial estimate of β_i^{init} and θ_i^{init}. We then weight the penalization component using the absolute value of the initial estimates $\hat{\beta}_i^{\text{init}}$ and $\hat{\theta}_i^{\text{init}}$. Adaptive LASSO yields a sparser solution and can be used to reduce the number of false positives; thus, it exhibits less bias than LASSO:

$$\min_{\alpha_i,\gamma i,\beta_i}\left(\frac{1}{2JT}\sum_{j,t=1}^{J,T}\left(P_{ijt}-\alpha_i-\hat{P}_{-i,jt}\beta_i^T-X_{jt}\gamma_i-E_{jt}\delta_j\right)^2+\lambda\sum_{n=1,n\neq i}^{N}\frac{|\beta_{in}|}{|\hat{\beta}_{in}^{\text{init}}|}\right),\quad \forall i,$$

or

$$\min_{\alpha_i,\gamma i,\beta_i}\left(\frac{1}{2JT}\sum_{j,t=1}^{J,T}\left(P_{ijt}-\alpha_i-P_{-i,jt}\beta_i^T-X_{jt}\gamma_i-E_{jt}\delta_j-\hat{\varepsilon}_{-i,jt}\theta_i^T\right)^2+\lambda\sum_{n=1,n\neq i}^{N}\left(\frac{|\beta_{in}|}{|\hat{\beta}_{in}^{\text{init}}|}+\frac{|\theta_{in}|}{|\hat{\theta}_{in}^{\text{init}}|}\right)\right),\quad \forall i.$$

We hereby compare the performances of the four commonly used criteria for tuning parameter selection: cross-validation, AIC, BIC, and adaptive LASSO (ALASSO). Under cross-validation, we choose the tuning parameter, which gives the minimum mean squared prediction error (Min MSE) and the tuning parameter, which gives the minimum mean squared prediction error plus one standard deviation (1SE). To see how each criterion performs for the purpose of variable selection, we demonstrate the results through simulating a single price equation as well as simultaneous price equations.

Table 1 shows the number of variables selected under each criterion, model precision, and recall with and without simultaneity. Precision is defined as the fraction of selected variables that are true positives. Recall is defined as the fraction of true positives that are selected. A good model should have both high precision and high recall; however, there is usually a trade-off between precision and recall. Typically, the more variables selected, the higher the recall—that is, the fewer true negatives will be left out. Meanwhile, the more variables selected, the lower the precision, because many selected variables are false positives.

We first focus on the results obtained without simultaneity (i.e., rows (1)–(3) in Table 1). Cross-validation-based methods (CV Min and CV 1SE) tend to select a large number of variables (row (1)). As a result, they are able to achieve a high recall probability. However, the precision of the model is considerably low (29.9% and 69.5% in row (2)). That is, cross-validation is more appropriate for prediction purposes but not for variable selection purposes. On the other hand, BIC and ALASSO achieve significantly better precision than cross-validation (83.7% and 87.7% versus 29.9% and 69.5% in row (2)) , without much sacrifice in recall (91.6% and 91.6% versus 99.5% and 96.7% in row (3)). Similar patterns can be observed for the results based on simultaneous price equations.

4.2. Two-Stage Least Squares vs. Two-Stage Residual Insertion

We now compare the two tactics previously mentioned to implement the instrumental variable approach in high dimensions: 2SPS versus 2SRI.

The lower panel (rows (7)–(12)) of Table 1 illustrates the comparison of these two methods, as well as the results of not using instruments at all. Rows (4)–(6) show the model selection accuracy is significantly lower when simultaneity is present compared with the case without simultaneity (rows (1)–(3)) under all tuning parameter selection criteria. In particular, precision is much lower because more variables tend to be selected. This is because even though some price variables are not direct influencers, they can still be correlated with the price variable of interest if they indirectly affect the price variable through other price variables.

Incorporation of instruments through 2SPS (rows (7)–(9)) does not actually improve performance. In particular, the recall probability is even worse than not using the instrument (rows (4)–(6)). On the other hand, 2SRI (rows (10)–(12)) provides sizable improvement over both the no instrument case and 2SPS, achieving higher precision (59.8% compared with 33.6% and 37.1%) and higher recall (97.6% compared with 88.6% and 61.0%) when adaptive LASSO is used as the tuning parameter selection criterion. To illustrate the differences graphically, we plot the precision-recall curve in Figure 1. Note that a more desirable model will generate both higher precision and higher recall. Therefore,

Table 1. Variable Selection Accuracy on Simulated Data

	CV Min	CV 1SE	AIC	BIC	ALASSO
Without simultaneity					
(1) No. of variables selected (median)	50	19	23	17	15
(2) Precision (average) (%)	29.9	69.5	60.5	83.7	87.7
(3) Recall (average) (%)	99.5	96.7	98.0	91.6	91.6
With simultaneity					
No IV					
(4) No. of variables selected (median)	101.5	58	52.5	16	39
(5) Precision (average) (%)	14.4	24.2	25.9	42.9	33.6
(6) Recall (average) (%)	97.6	94.6	92.8	50.1	88.6
IV—2SPS					
(7) No. of variables selected (median)	89	31	14	1	22.75
(8) Precision (average) (%)	15.8	30.5	45.5	84.7	37.1
(9) Recall (average) (%)	92.5	69.0	50.8	11.4	61.0
IV—2SRI					
(10) No. of variables selected (median)	65	37	32	18	24
(11) Precision (average) (%)	23.0	39.5	45.3	70.2	59.8
(12) Recall (average) (%)	99.5	99.2	98.9	87.3	97.6

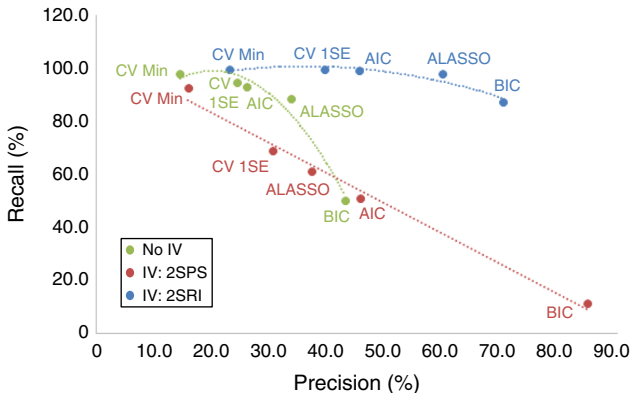
Notes. *Without simultaneity* means that we generated Gaussian data with $N = 800$ observations and $p = 199$ predictors based on the following equation: $p_{1t} = \beta p_{-1,t} + \sigma\epsilon_{1t}$. Prices are simulated under multivariate normal distributions with zero means and an identity variance–covariance matrix. The term ϵ is simulated from the standard normal distribution. The coefficient vector β has 15 nonzero entries, and $\beta = \{3, 2, 2.5, 3, 2, 2.5, 3, 2, 2.5, 3, 2, 2.5, 3, 2, 2.5, 0, 0, \ldots, 0\}$. The number of nonzero coefficients is chosen to be on par with the average number of competitors identified in our empirical setting; σ is chosen so that the signal-to-noise ratio is approximately 0.43 (a corresponding R-square of 30%), again on par with the empirical data. *With simultaneity* means that we generated $N = 800$ observations of 200 price variables based on the following simultaneous equations, $p = \delta Wp + \epsilon$, where $\epsilon = \alpha E + \nu$. The term W represents the adjacency matrix of a random directional graph with 200 nodes and an expected outdegree of 15 (on par with empirical data), without self-links; ϵ is the vector of shocks and can be decomposed into two orthogonal vectors E (observed) and ν (unobserved); E_i is used as the instrument for its corresponding price variables p_i; α controls the relative portion of observed shocks and is set to 1 in the simulation (i.e., the instrument explains 50% price variation). Both E and ν are simulated from the standard normal distribution. Scalar δ is chosen so that the average R-square is approximately 30%, again on par with the empirical data. All results are based on 100 replicas.

the closer the precision-recall curve is to the upper right corner in the plane, the better the model. As one can see from the figure, it is striking how much better 2SRI performs relative to 2SPS and to the case without using any instrument.

The reason for the performance gap between 2SRI and 2SPS is the following. A shortcoming of replacing the original variable with the predicted value from the first stage (2SPS) is that only partial variation from the original variable is passed along to the second stage, likely causing insignificant estimates. When we select variables in high dimensions, such variables are likely to be dropped by LASSO not because they are irrelevant but because the predicted value does not capture sufficient variations from the original variable. In 2SRI, on the contrary, both the original variable and the predicted residual are included. When LASSO conducts variable selection, if both the original variable and the residual are selected or if neither is selected, the equation is the same as including only the predicted value. When the original value is not selected but the residual is selected, it is again practically the same as when including the predicted value only but dropping it in the end. In other words, if the variable is not selected with full variation from the original value, it will most likely not be selected with limited variation from the first-stage predicted value.[6]

However, when the original variable is selected but not the residual, it indicates that endogeneity is not a concern (or is only a mild concern) for that variable, and in this case, we have used the full amount of variation present in the original variable. Therefore, 2SPS will likely falsely claim some price variables as insignificant, simply because insufficient variation is passed along by the predicted value, while falsely claim other variables as significant. By contrast, 2SRI will allow us to identify more relevant competitors' prices by using full variation in the original price variable, meanwhile

Figure 1. (Color online) Recall-Precision Curve



addressing endogeneity whenever *necessary*. We note that our results are in some way consistent with Terza et al. (2008), who also find 2SRI outperforms 2SPS though under a different class of nonlinear models.

4.3. Strength of the Instruments and the Estimation Performance

As in low-dimensional instrumental variable estimation, the performance of the instrumental variable is likely affected by its strength—in other words, the amount of variations that an instrument captures from the original endogenous variable. The weaker the instrument, the noisier the estimates will be (larger standard errors). In high-dimensional analysis with variable selection as the objective, the strength of the instrument is also likely to affect the accuracy of variable selection. In this section, we analyze the impact of the strength of the instrumental variable on the accuracy of the selected variable and of the estimates, and we evaluate methods to alleviate concerns of potentially weak instruments.

As we can see from Table 2, column (1), as the instrument becomes weaker (i.e., explains less variation from the original endogenous variable), the accuracy of variable selection, both in terms of precision and recall, decreases. Precision decreases from 59.8% to 55.4% to 47.9% as the amount of the explained variation goes down from 50% to 20% to 5%. Similarly, recall probability decreases from 97.6% to 78.0% to 63.4%. The estimates also become noisier as the instrument becomes weaker as shown in the left panel of Table 3. The average estimated standard error goes up from 0.034 to 0.067 to 0.134.

Note that since we observe exposures from only one distribution channel, the amount of variation that we can capture in the price variable might be limited. To alleviate the impact of potentially weak instruments, we propose to use residual bootstrapping to achieve higher accuracy in estimation. Bootstrap has been used as a method to adjust the estimated standard errors when instruments are estimated in a two-stage fashion. Moreover, bootstrap has also been proposed as a method to correct for biases in the estimation of confidence intervals following LASSO. It is known that standard bootstrap is inconsistent, so Chatterjee and Lahiri (2011) propose a modified bootstrap method that produces consistent estimates of variances of the LASSO estimator under mild conditions. In particular, they show that residual bootstrap is consistent for adaptive LASSO.

Thanks to these properties of the bootstrapping method, we will use residual bootstrap in our context as well to increase the model selection and estimation accuracy in the presence of potentially weak instruments in high dimensions. We first conduct the regular LASSO (first stage of adaptive LASSO), which allows us to estimate the residual for each observation based on the estimated model. We then resample with replacement from the estimated residuals and use these bootstrapped residuals to regenerate the outcome variables. We bootstrap 100 replicas and then perform the proposed analysis on each replica. Finally, we examine which variables are selected more frequently from the 100 bootstrapped samples and calculate the 90% confidence intervals of the estimates. Columns (2)–(4) in Table 2 show that residual bootstrap is able to significantly improve the model selection accuracy, particularly for the case involving weaker instruments. When the instrument explains only 5% original variation, both precision and recall go up with the use of residual bootstrap. The precision increases

Table 2. Model Selection Accuracy of Bootstrap Adaptive LASSO Estimator

		Residual bootstrap		
	No bootstrap (1)	Selected by ≥20% resample (2)	Selected by ≥40% resample (3)	Selected by ≥60% resample (4)
IV explains 50% variation				
No. of variables selected (median)	24	16	15	13.5
Precision (average) (%)	59.8	89.1	94.6	97.0
Recall (average) (%)	97.6	95.6	92.3	88.0
IV explains 20% variation				
No. of variables selected (median)	21	15	11	8
Precision (average) (%)	55.4	73.8	84.5	88.7
Recall (average) (%)	78.0	74.6	63.3	50.6
IV explains 5% variation				
No. of variables selected (median)	20	17	10	5
Precision (average) (%)	47.9	63.3	72.0	77.6
Recall (average) (%)	63.4	70.0	49.0	31.7

Notes. Data are simulated using the same system of equations as in Table 1. We adjust the value of α to allow the instrument to explain varying levels of variations in prices. Results are based on 100 bootstraps for each of the 200 price equations.

Table 3. Parameter Estimation Accuracy of Bootstrap Adaptive LASSO Estimator

	No bootstrap			Residual bootstrap		
IV explains	Empirical coverage of 90% CI (%)	Coeff. mean (std. dev.)	Std. err. mean (std. dev.)	Empirical coverage of 90% CI (%)	Coeff. mean (std. dev.)	Std. err. mean (std. dev.)
50% variation	95.3	0.068 (0.079)	0.034 (0.002)	96.9	0.107 (0.029)	0.096 (0.083)
20% variation	96.4	0.065 (0.108)	0.067 (0.006)	97.3	0.119 (0.030)	0.047 (0.012)
5% variation	95.7	0.078 (0.167)	0.134 (0.025)	97.2	0.148 (0.048)	0.070 (0.035)

Notes. Data are simulated using the same system of equations as in Table 1. True parameter $\delta = 0.15$, chosen to be on par with the real data. Empirical coverage is defined as the fraction of the estimated 90% confidence intervals (CIs), which contain the true parameter value. Mean and standard deviations are calculated for variables selected by at least 20% bootstrapped samples. The table shows that residual bootstrap generates confidence intervals with better coverage of the true parameter. It also shows that residual bootstrap generates more consistent estimates and smaller standard errors, especially in the presence of weaker instruments. We repeat the same procedure with 100 bootstrapped samples for each of the 200 price equations.

from 47.9% to 63.3% and the recall increases from 63.4% to 70.0%, when we focus on, for example, those variables that are selected by at least 20% of the bootstrapped samples. Similarly, residual bootstrap produces confidence intervals with better empirical coverage, as shown in Table 3, which is defined as the fraction of time that the 90% confidence interval contains the true parameter value.

5. Data

We obtain hotel demand and price information using a search and clickstream data set from an online travel aggregator.[7] A distinguishing feature of our data is the availability of very detailed, and high-frequency, data on consumer search actions. These data are automatically generated as consumers compare options on the travel platform and filter those they are interested in. The frequency at which a hotel is observed during travel searches is indicative of current demand conditions for that particular hotel. In other words, clickstream data are a stream of signals that indicate shocks to consumer preferences for staying at a particular hotel on a particular date. Online platforms also generate large volumes of price observations, at high frequency: prices are recorded as soon as they are observed by consumers on the screen.

The data set contains the entire search history of every user who searched for hotels in New York City on the travel aggregator's website during the observation period from May 1, 2007 to May 31, 2007. It includes information regarding destination and dates a customer searched for, which filtering and/or sorting criteria were applied, which hotels were returned in response, and which hotels were clicked, if any. In our data, there are 418 hotels/hostels registered in New York City and its vicinity; 363 are within Manhattan, among which 68 are nonhotels (hostels, bed and breakfast, lodge, guest house, etc.). We focus on the 229 hotels that are located in Manhattan and are exposed at least once during the period of study, because this is a set of choices that generates most interest among travelers to New York City and because it generates far more exposures and clicks than the rest. Among them, 60.0% are independent properties, representing 43.0% of room capacity. The remaining 40.0% are branded properties. Among the branded properties, there are a few major chains including Hilton Worldwide, Choice, Marriott, IHG, and Starwood operating 53 properties under 27 different brand names and representing 36.1% of room capacity in the market.

5.1. Customer Search Patterns

A typical customer arrives at the travel aggregator's website and initiates a search by submitting a request, which includes the destination (the city), check-in and checkout dates, number of guests, and number of rooms. An average search leads to 350–400 results within New York City and its vicinity, which are displayed in pages with 15 hotels per page. Once the search results are loaded onto the browser, the customer will decide whether or not to engage in additional search actions, including filtering, sorting, and paging. If a customer becomes interested in a particular hotel in the search result list, she will click to see more details. In total, we observe 53,357 unique search histories for New York City during May 2007.

Consumers who search on the website adopt various search strategies to narrow down the set of hotels that fit their specific preferences. A detailed breakdown can be seen in Table 4. Each sample path could contain multiple search actions. For example, a sample path could be that a customer, after the initial search result was loaded, applied a landmark filter of Times Square, then a neighborhood filter of Midtown West, and finally a star filter of four stars and above; then the customer starts paging through the listed results. Every time a search action is taken, the results will be refreshed.

Table 4. Percentage of Customers Engaged in Each Category of Search Action

	Location	Price	Quality	Brand	Amenities/Services
No action (42.5%)					
Filter (44.0%)	Landmark (18.81%), neighborhood (11.70%), distance (10.69%)	Price range (15.82%)	Stars (6.98%)	Brand and hotel name (3.94%)	Amenities and services (1.92%)
Sort (33.5%)	Distance ascending (3.53%)	Price ascending (22.97%) Price descending (1.78%)	Rating descending (3.22%)	Hotel name (1.27%)	

Through this sequence of search actions, we learn that there is one potential demand for hotels in Midtown West—in particular, those close to Times Square with four stars or above for certain dates of travel. All hotels exposed to the customer had the potential chance to capture this specific demand. The average number of actions taken by a customer is 1.8 (paging actions excluded). The median number of unique hotels a customer is exposed to is 15 (repetitive and transient exposures excluded). The average number of hotels clicked per customer is 0.73—that is, a clickthrough rate of 1.17%. An average hotel is exposed to 7,900 customers (with possible repetition).

5.2. Hotel Pricing Strategy

The hotel industry exhibits significant price variation across properties, room types, customer types, days of travel, and days of booking. Our sample of 229 Manhattan hotels offers an average room rate of $279.9 with a standard deviation of $170.9. Price variations are observed across hotels and across dates: both travel dates and booking dates. Price variation along travel date is mostly driven by market segmentation. For instance, business travelers tend to travel on weekdays, while leisure travelers tend to travel on weekends or holidays. Variation among booking dates is driven by market segmentation as well. For example, business travelers tend to book hotel rooms close, but usually not too close, to the travel date. Leisure customers tend to book far in advance of the travel date. Sometimes, leisure travelers also make last-minute reservations to enjoy deep discounts when rooms are still available. Inventory controls also drive price variations among booking dates. A typical hotel offers 10–12 rate bands and manages the opening or closing of these rates dynamically over the booking horizon.[8]

Another factor that drives price variations in the hotel industry is competition. Monitoring competitors' prices is part of everyday operations in the industry. "Call-around" is a common practice whereby "hotels engage in regular communications, two or three times daily, to exchange standard [room] rates . . . . In determining whether to manually override the reservation system, the Regional Revenue Manager may periodically consult various sources of information concerning competitor rates, including publicly listed rates through Internet sites or other market information."[9] In addition to the call-around practice, hoteliers subscribe to automated tools, such as PriceTrack, RateVIEW, and HotelCompete Rate Analyzer, to monitor competitors' rates closely. These tools help hoteliers track rates on a daily basis or even more frequently for competitors in their prespecified competition set (Cross et al. 2009). Finally, the development of online search engines exposes a wealth of readily accessible price information.

For our purpose, a product in this market is defined as a one-night stay for a specified travel date in the future and on a given booking date. However, we do not directly observe prices for each night but only the quoted average price during the length of stay. In Online Appendix EC.4, we describe how we construct nightly prices using quoted prices.

6. Estimation Results

In this section, we first discuss how we construct and evaluate potential instrumental variables based on the two critical criteria: the exclusion and inclusion criteria. We then apply the instrumental variable approach to high-dimensional variable selection. Finally, we show examples of selected competitors and contrast them with those selected based purely on price correlations.

6.1. Choice of Instrumental Variable

A good instrument has to satisfy two criteria: (1) the exclusion criterion that the instrumental variable is uncorrelated with the unobserved error term in the main equation and (2) the inclusion criterion that the instrumental variable should be (strongly) correlated with the endogenous variable to avoid being a weak instrument. Both conditions are critical for the instrument to be theoretically valid as well as practically efficient.

6.1.1. Clicks and Exposures. We observe two types of demand signals: clicks and exposures. Since clicks happen after prices are observed by customers, they are likely endogenous with prices and hence are not valid instruments. Exposures, on the other hand, are primarily driven by exogenous travel needs—that is, on which dates to travel and preferences in quality and locations (e.g., four-star hotels close to the Financial District). These exposures are not correlated with

prices charged by specific hotels and hence are valid instruments. Some exposures, though, can be results of filtering based on a certain price range or sorting based on ascending or descending prices. We exclude these types of exposures because they are endogenous with prices. In other words, we consider exposures resulting from initial search without filtering or re-sorting and exposures resulting from filtering and re-sorting based on non-price criteria (brand, location, quality, amenities, services, etc.) as valid for instrument construction.

We further distinguish two types of exposures: transient and actual. We discover from the data that in many cases when a customer knows the type of hotel that she is looking for (say, hotels in the Financial District), she applies the location filter right away, once the initial search results are shown. In this case, she spends minimal time on the initial results given by the search engine because the generic search results do not satisfy her specific travel needs. In other situations, a consumer with complex preferences applies multiple filters in a row. Each time a filter is applied, a new set of results will be shown; however, the consumer may not pay attention to the search results until she arrives at the final result once all filters are applied. We call these intermediate exposures transient. For our purposes, transient exposures are not included in our construction of instrumental variables because they do not reflect real demand. Transient exposures are determined using the time duration (in seconds) that a customer spends on a page. Specifically, results that are displayed for less than 60 seconds will be considered transient. We choose this cutoff using the amount of time that customers spend on pages that led to clicks as a benchmark. The rationale is that if any result of a page has been clicked, it will not be a transient session because the customer was making active decisions. We observe that conditional on a page of search results being clicked, 90% of customers spend more than 60 seconds on the page before making the first click. More precisely, we define a page of search results as transient if the duration is below 60 seconds *and* if no results shown on the page are clicked. We also experimented with other cutoffs (45 and 75 seconds), and the results are consistent. When a hotel is exposed to a customer multiple times, we count it as one exposure because there can be at most one demand. We also weight the exposures by the number of rooms requested. Thus, we obtain the number of customers that a hotel is exposed to on each booking date for a given travel date.

6.1.2. Level of Aggregation. To construct the instruments, we consider different levels of aggregation over the booking date horizon, and we define the following three measures.

1. *Same-day exposures* (E_{ijt}; i.e., exposures received by hotel i on travel date j for a booking date t): "Same-day" refers to the same booking date as the booking date when the price is observed.

2. *Recent-week exposures* ($\sum_{s=t-7}^{t-1} E_{ijs}$): It is likely that when hotel managers respond to demand signals, they respond not to signals on that particular day of booking but rather to demand signals observed over the recent booking period. We construct another demand measure using the total exposures in the past seven days of booking (the current day of booking excluded) prior to the booking date when the price is observed for each travel date. Aggregation at seven days also removes the day-of-week effect.[10] As demands over adjacent travel dates are likely to be correlated, we also test exposures aggregated at ±1 day, ±2 days, and ±3 days around the travel date, with and without aggregation over the recent week. We find the performances of these measures similar to, but not better than, the measures constructed without aggregation over travel dates.

3. *Recent-week clicks* ($\sum_{t-7}^{t-1} C_{ijs}$, where C_{ijs} denotes clicks received by hotel i for travel date j on booking date s): In a spirit similar to what we do for exposures, we also test realized demand shocks (i.e., clicks), aggregated over the past seven days, again excluding the current booking date because the number of clicks received on the current booking date is directly affected by the price charged on that day. It is likely that aggregate clicks in the past seven days are still correlated with the current-day price if there is some degree of price stickiness over time. Nevertheless, we test its inclusion validity as an instrument.

One practical concern with the instrumental variable approach is the strength of the instrument. We thus test the first-stage explanatory power of each previously proposed instrument using Equation (1) to select the set of instruments that explain most variations from the original price variables. All coefficients are hotel specific, as the model is estimated separately for each hotel.

We compare the additional explanatory power on top of all exogenous variables in X_{jt}, which include travel date and observation date characteristics and advance purchases. We also compare the statistical significance of the estimated coefficients. Table 5 demonstrates that the second instrument, recent-week exposures, strongly dominates the other two instruments, same-day exposure and recent-week clicks, with significantly higher explanatory power and the number of cases being statistically significant. Recent-week exposures explain on average 5.60% of additional variation in hotel prices, relative to 2.25% and 4.93% explained by the other two measures. In terms of statistical significance, the recent-week exposures variable outperforms the other two instruments by a large margin.

Table 5. Choice of Instruments: Results from the First-Stage Regression

	Incremental explanatory power (increase in R-squares) (%)				Statistical significance (counts based on T-stats)			
Instrument	Mean	$p25$	$p50$	$p75$	$t > 1.64$	$t > 1.96$	$t > 2.56$	Total no. of hotels
Same-day exposures	2.25	0.23	0.82	2.57	42	40	32	177
Recent-week exposures[a]	5.60[b]	1.87	4.43	8.01	84	79	69	177
Recent-week clicks[a]	4.93	1.29	4.07	6.44	39	35	24	177

[a]Total exposures (or clicks) counted from seven days to one day prior to the current booking date. In other words, current booking-day exposures (or clicks) are excluded from this measure.
[b]Note that 5.60% represents a 5.60-percentage-point increase in first-stage R-square.

6.1.3. Hotel Pricing Strategy—Results from the First-Stage Regression. A significant portion of price variation in the hotel industry can be explained by characteristics of the booking date, the travel date, advance purchases, the exposure, and the overall market demand (defined as the total number of users who searched for travels in the destination given a travel date and a booking date). We demonstrate such pricing patterns using a few hotels with different star levels as examples: Comfort Inn Chelsea Hotel (two stars by Choice Hotels), Amsterdam Court Hotel (three stars, independently owned), DoubleTree Suites Times Square (four stars by Hilton), and The Peninsula New York (five stars by The Peninsula Hotels). As shown in Table 6, travel date exhibits a significant day-of-week effect, which differs across hotels. While premium hotels DoubleTree and The Peninsula charge significantly higher prices on Monday to Thursday compared with Friday and Saturday, budget and economy hotels Comfort Inn Chelsea and Amsterdam Court Hotel charge slightly lower (or no higher) prices on Monday to Thursday relative to Friday and Saturday. This observation makes intuitive sense because the former hotels attract mostly business customers

Table 6. Hotel-Specific Pricing Strategies—Examples

	Comfort Inn Chelsea (2 stars; Choice Hotels)		Amsterdam Court (3 stars; Independent)		DoubleTree Times Square (4 stars; Hilton)		Peninsula New York (5 stars; Peninsula Hotels)	
	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.
Recent-week exposures	0.331***	0.087	0.458***	0.116	0.887***	0.184	2.331***	0.560
Booking date day of week								
Monday	−0.808	4.116	−4.632	4.123	−11.141	6.827	−1.213	24.882
Tuesday	−12.819***	4.496	−0.707	4.549	−30.921***	6.653	−22.361	28.904
Wednesday	−6.352	4.367	1.166	4.828	−28.035***	6.604	−43.769	28.518
Thursday	−5.671	4.123	3.228	4.343	−18.338***	6.650	−13.797	25.515
Friday	−7.007*	4.243	−4.178	4.240	−5.881	6.678	−1.137	23.601
Saturday	0.866	4.370	4.908	4.860	−2.155	7.256	−20.477	24.576
Sunday	Baseline		Baseline		Baseline		Baseline	
Travel date day of week								
Monday	−8.083**	3.421	10.190**	4.191	38.128***	7.214	121.892***	23.581
Tuesday	−9.900**	4.072	20.475***	4.982	55.589***	6.685	180.181***	25.527
Wednesday	−11.428***	3.844	13.315***	4.051	51.708***	6.716	191.873***	23.321
Thursday	0.541	3.739	4.798	3.435	36.073***	5.936	141.666***	22.280
Friday	2.838	3.575	10.357***	3.968	−6.818	5.751	−57.068**	22.445
Saturday	−4.253	4.151	12.164***	4.140	−18.101***	5.893	−111.419***	26.199
Sunday	Baseline		Baseline		Baseline		Baseline	
Advance purchase								
Advance < Day	−30.014***	10.252	−66.632***	5.385	−26.546	18.122	−63.097	57.942
1 day ≤ Advance < 3 days	−32.219***	6.082	−72.633***	4.828	−56.974***	10.628	47.913	42.850
3 days ≤ Advance < 7 days	−20.731***	4.246	−53.825***	5.029	−49.025***	7.043	73.654**	30.260
7 days ≤ Advance < 14 days	−4.876	3.088	−5.373	4.131	−25.140***	4.638	77.911***	17.954
14 days ≤ Advance < 21 days	−13.258***	2.181	0.672	2.653	−3.942	4.206	89.991***	14.320
21 days ≤ Advance < 30 days	Baseline		Baseline		Baseline		Baseline	
Total traffic	0.196***	0.058	0.246***	0.073	0.491***	0.071	−0.028	0.389
Constant	236.515***	4.253	232.302***	5.636	417.506***	8.017	782.816***	29.319
No. of obs.	668		660		588		484	
R-square (adjusted)	0.3134		0.3972		0.3261		0.4576	

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Table 7. Types of Hotels Selected

Stars	Avg. no. of competitors	Within 0.2 miles (%)	Within 0.5 miles (%)	Same stars (%)	Same and adjacent star levels (%)
1	3.2	0.0	8.3	51.7	85.0
2	8.9	4.1	13.0	35.8	81.0
3	10.7	7.5	22.8	34.1	89.9
4	8.8	7.8	23.4	33.4	84.8
5	8.8	5.2	17.9	14.9	46.9
Total	9.1	6.8	21.0	33.2	83.2

who travel on weekdays, while the latter hotels attract more leisure customers who travel mostly on weekends. For all hotels shown in this example, prices are somewhat nonlinear over the booking horizon. Total traffic exhibits a positive and significant effect on prices in all but the five-star hotel.

6.2. Variable Selection

Once we obtain the predicted values and predicted residuals for all prices using instruments, we now apply the LASSO method to identify key competitors. As we did with simulated data, we compare different criteria used to select the tuning parameter as well as the performance of 2SPS and 2SRI. We obtained consistent evidence as with simulated data, as shown in Online Appendix EC.5. For the final estimation, we use adaptive LASSO together with 2SRI and 100 bootstrapped samples to select the set of competitors for each hotel (selected by at least 20% bootstrapped samples).[11]

The selection results are summarized in Table 7. Overall, our method selects most hotels with the same or adjacent star levels as competitors (83.2%) and a significant portion from hotels within 0.5 miles as competitors (21.0%), though these fractions vary by star levels.[12]

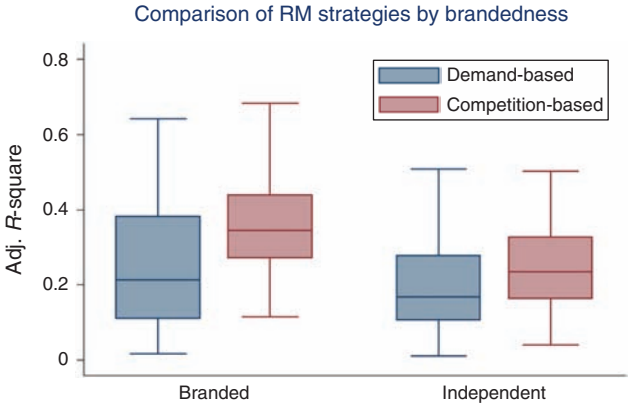
7. Hotel Competition in New York City

Now that we have identified key competitors of each hotel, we study the patterns of price competition in New York City by examining the extent to which prices are determined by competition, the physical and quality boundaries of price competition, and the mismatch of the price-based competition network and the demand-based competition network.

7.1. Level of Engagement in Competition-Based Revenue Management

As we discovered previously, demand signals and exogenous factors such as characteristics of travel and booking dates and advance purchases explain a significant portion of price variation in this market, consistent with practices discussed in Talluri and van Ryzin (2005). We refer to such practice as demand-based revenue management as it broadly reflects the practice

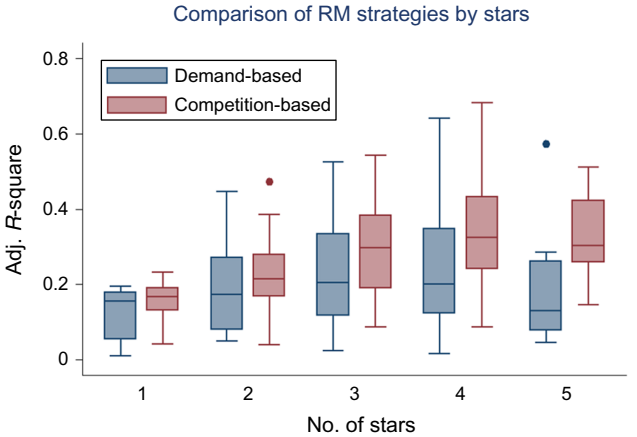
Figure 2. (Color online) Boxplots of Price Variations Explained by Brandedness



of segmenting markets based on customers' willingness to pay. We compare the amount of price variation explained by demand-based revenue management and by competition-based revenue management. Figures 2 and 3 show that engagement in either type of revenue management practice is prevalent across hotels of all star levels, branded or not. Percentage of variation explained in both cases is calculated as an additional adjusted R-square when including either set of variables (demand variables or competitor price variables) in addition to residuals selected by LASSO. Demand-based factors alone explain on average 22.3% of price variation, while competitor prices explain on average 30.2% of the variation—an additional 7.9% of variation.

The engagement in revenue management differs by the brandedness and the star level of a hotel, as shown in Figures 2 and 3, respectively. Branded hotels, which are nearly half of the total number of hotels in the sample, are more engaged in revenue management practices in general, including both demand-based and competition-based revenue management.

Figure 3. (Color online) Boxplots of Price Variations Explained by Stars



Note. The outside values (two dots) are plotted using the standard boxplot definitions.

Table 8. Quality and Geographical Boundaries of Price Competition—Marginal Effects from Logit Model of Network Links

	Branded		Indep		Two stars or below		Three stars		Four stars		Five stars	
	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.
Brand–Brand	0.016***	0.004	—		0.007	0.007	−0.018***	0.006	0.005	0.005	0.016	0.015
Brand–Indep	Baseline		—		−0.003	0.006	−0.019***	0.006	−0.017***	0.004	−0.004	0.013
Indep–Brand	—		−0.029***	0.003	−0.033***	0.005	−0.032***	0.005	−0.022***	0.004	0.000	0.014
Indep–Indep	—		Baseline		Baseline		Baseline		Baseline		Baseline	
Same quality (stars)	0.034***	0.005	0.023***	0.004	0.047***	0.009	0.031***	0.006	0.019***	0.004	0.045**	0.022
Same neighborhood	0.002	0.004	0.007*	0.004	0.004	0.006	0.012*	0.006	0.001	0.004	0.006	0.012
No. of hotels	87		90		30		51		84		12	
No. of obs.	15,312		15,840		5,280		8,976		14,784		2,112	
Log likelihood	−3,229.40		−3,035.08		−904.11		−2,022.82		−2,893.41		−413.30	

Note. All marginal effects evaluated at mean.
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Similarly, hotels with higher star levels are also more engaged in the general practice of revenue management: both demand-based and competition-based.

7.2. Boundaries of Price Competition

We examine two common boundaries of competition, horizontal (i.e., geographical distance) and vertical (i.e., quality), and how these boundaries are defined differently for branded versus independent hotels and hotels at different star levels.

Table 8 shows how much being in the same quality tier or being in the same neighborhood affects the likelihood that two hotels compete directly in price. We find that branded hotels are more constrained by quality boundaries while less by geographical boundaries compared with independent hotels. In other words, when choosing whose price to follow, branded hotels are more concerned about whether the hotel is within the same quality tier but less whether the hotel is close by compared with independent hotels. We also find that price competition for budget hotels (one and two stars) and luxury hotels (five stars) is more constrained by quality boundaries while less by geographical boundaries compared with midlevel quality hotels (three and four stars). That is, when choosing whose prices to follow, budget and luxury hotels are more concerned about quality than distance, while economy and upscale hotels are more concerned about distance than quality.

7.3. Price-Based vs. Demand-Based Competition Networks

In this section, we contrast the price-based competition network with the demand-based competition network. We estimate a consumer choice model using the clickstream data. Specifically, we estimate a random coefficient multinomial logit model, which allows for heterogeneous tastes among consumers over a broad set of hotel and trip characteristics including price, stars, independent versus branded, neighborhood (9 neighborhoods), amenities (6 amenities), chains (14 largest chains in New York City), travel day of week, and days in advance. The details and the estimates of the model are presented in Online Appendix EC.6.

From the model estimates, we obtain self-price and cross-price elasticities between all pairs of hotels. The cross-price elasticities, in particular, measure the intensity of competition between any pair of hotels—that is, how much will hotel i's demand change, measured in percentage, if hotel j's price changes by 1%. We therefore construct a demand-based competition network using cross-price elasticities and compare it with the previously obtained competition network based on price reaction functions. For comparability, we adopt two approaches to construct the demand-based network. In one approach, competitive relationships are formed using a universal cutoff on the estimated cross-price elasticities, such that the demand-based and the price-based networks have the same density. In the second approach, competitive relationships are formed based on node-specific cutoffs on the estimated cross-price elasticities, such that the same node has the same outdegree in the two networks. We find there is an 89.8%–90.1% overlap between the price-based and the demand-based competition networks. We then estimate when the mismatch is likely to occur. As shown in Table 9, when hoteliers react to competitors' prices, they tend to miss those potential competitors who are geographically farther away or are of different quality tiers (i.e., different star levels), and tend to react to those who are closer in distance and who are at the same quality tier.[13]

8. Managerial Implications for Proactive vs. Reactive Hoteliers

In this section, we demonstrate how a proactive hotelier can leverage the knowledge of competitive responses generated by our algorithm. We will compare the performances of a proactive hotelier, who anticipates competitive responses and strategically takes them into

Table 9. Price-Based and Demand-Based Competition Network Mismatch

	Relationships missed				Nonrelationships included			
	(1)		(2)		(3)		(4)	
	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.
Brand–Brand	0.074	0.206	0.082	0.187	0.010	0.067	−0.001	0.067
Brand–Indep	0.907***	0.302	0.781***	0.248	−0.265***	0.070	−0.264***	0.071
Indep–Brand	1.052***	0.291	0.961***	0.256	−0.607***	0.077	−0.606***	0.078
Indep–Indep	Baseline		Baseline		Baseline		Baseline	
Same quality (stars)	−0.718***	0.181	−0.711***	0.160	0.474***	0.056	0.459***	0.057
ln(*Distance*)	0.208**	0.092	0.074	0.085	−0.079***	0.029	−0.085***	0.030
No. of obs.	1,589		1,769		29,557		29,377	
Pseudo R^2 square	0.046		0.037		0.013		0.013	
Log likelihood	−456.6		−573.0		−6216.2		−6,085.2	

Notes. All columns measure the mismatch of price-based network against demand-based network. Columns (1) and (2) estimate the likelihood of missing a competitive relationship conditional on it is identified in the demand-based network. Columns (3) and (4) estimate the likelihood of including a nonrelationship from the demand-based network. For comparability, in columns (1) and (3), demand-based competitive relationships are formed using a universal cutoff on estimated cross-price elasticity such that the demand-based and the price-based networks have the same density. In columns (2) and (4), demand-based competitive relationships are formed based on node-specific cutoffs on estimated cross-price elasticity such that the same node has the same outdegree in the two networks. All columns are estimated using the logit model.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

account when making decisions (setting prices or offering promotions), and a reactive hotelier, who does not take into account competitive responses but merely reacts to what competitors do myopically when making decisions. In particular, we want to know whether a hotel or a chain of hotels decides to offer web promotions to boost online traffic, how a proactive hotelier would take into account the potential competitive responses, and what the consequences are if he does not.

Consider a rather general demand model, in which the demand for hotelier i for travel day j on booking date t can be written as a function of own price and competitors prices: $d_{ijt}(p_{ijt}; p_{-ijt})$. Suppose the hotelier has a growth target in mind, which is to increase online demand by $\Delta d = x\%$ by offering web promotions. He will choose a discount level of ϕ such that

$$\frac{d_{ijt}(p'_{ijt}; p'_{-ijt}) - d_{ijt}(p_{ijt}; p_{-ijt})}{d_{ijt}(p_{ijt}; p_{-ijt})} = \Delta d, \quad \text{where } p'_{ijt} = (1-\phi)p_{ijt}.$$

If hotel i charges p'_{ijt}, other hotels that closely follow hotel i's price will update their prices as a response. Given the competitive price response functions estimated based on our algorithm (i.e., $\partial p_{kjt}/\partial p_{ijt} = \beta_{ki}$, $k \neq i$), each of the other hotel's new prices can be approximated by

$$p'_{kjt} \simeq p_{kjt} + \beta_{ki}(p'_{ijt} - p_{ijt}), \quad \forall k \neq i, \tag{4}$$

where β_{kj}, $k \neq j$ measures the sensitivity of hotel k's price to hotel j's price. Note that we approximate competitor hotels' new prices using only the direct price responses but do not further include those price changes more than one step away. Such simplification gives us a reasonable approximation of competitors' new prices in a short time interval when the web promotion is offered. It is often difficult to expect even those smart hoteliers to know the intensity of competitive responses among other hotels, nor is it their primary concern as the intensity of response decays quadratically beyond the primary link.

A reactive hotelier i will choose p'_{ijt} while ignoring the potential competitive responses; that is, he chooses a discount level ϕ_{re} conditional on the currently offered competitive prices to solve the following problem:

$$\frac{d_{ijt}(p'_{ijt}; p_{-ijt}) - d_{ijt}}{d_{ijt}(p_{ijt}; p_{-ijt})} = \Delta d, \quad \text{where } p'_{ijt} = (1-\phi_{re})p_{ijt}. \tag{5}$$

A proactive hotelier, on the other hand, will instead choose ϕ_{pro} to solve

$$\frac{d_{ijt}(p'_{ijt}; p_{-ijt}(p'_{ijt})) - d_{ijt}(p_{ijt}; p_{-ijt})}{d_{ijt}(p_{ijt}; p_{-ijt})} = \Delta d, \quad \text{where } p'_{ijt} = (1-\phi_{pro})p_{ijt}, \tag{6}$$
$$p'_{-ijt}(p'_{ijt}) = p_{-ijt} + \beta_{-i,i}(p'_{ijt} - p_{ijt}).$$

We compare the pricing differences between the two types of hoteliers, and the growth target missed when ignoring competitive responses, at three levels of growth target: 5%, 10%, and 20%. To obtain such estimates, we first estimate the demand model as presented in Online Appendix EC.6 and then use the estimated demand model to simulate consumer choices under different pricing strategies. Note that the choice model is estimated conditional on the observed consideration set of each consumer. A typical consumer

Table 10. Pricing Errors and Growth Target Missed When Ignoring Competitive Responses

	Growth target (%)		
	5	10	20
Chainwide promotions			
Pricing errors (%)	4.3 ± 3.9 (0.1–11.1)	4.3 ± 3.8 (0.1–11.2)	4.2 ± 3.8 (0.1–11.2)
Growth target missed (%)	4.4 ± 4.0 (0.1–11.4)	4.5 ± 4.0 (0.1–11.6)	4.6 ± 4.1 (0.1–12.1)
Individual hotel promotions			
Pricing errors (%)	2.1 ± 5.3 (0.0–19.9)	2.2 ± 5.8 (0.0–20.5)	2.2 ± 5.8 (0.0–21.7)
Growth target missed (%)	2.6 ± 9.0 (0.0–23.1)	2.7 ± 9.2 (0.0–23.6)	2.9 ± 9.5 (0.0–24.7)

Notes. Mean ± standard deviation (95% confidence interval). The results are simulated using the random coefficient multinomial logit model estimated from consumer clicks, where the last clicked option is used as the proxy for the final choice. The results are similar when using a randomly selected clicked option as the proxy for the final choice. The results are obtained from 100 rounds of simulations and 2,000 customers for each round.

chooses from a sparse consideration set (with a median of 15 hotels in our data) rather than the whole universe of hotels. To simulate a demand pattern that is as close as possible to what is observed in the data, we first simulate the sparse consideration sets. To do so, we estimate coexposure probabilities for all pairs of hotels and then simulate consideration sets using a multivariate Bernoulli distribution whose variance-covariance matrix is based on the coexposure probabilities estimated from the data. We perform 100 rounds of simulations. In each simulation, we simulate 2,000 customers (2,000 consideration sets) and their choices using the estimated demand model. We analyze the reactive hotelier and the proactive hotelier's decisions by solving Equations (5) and (6).

Table 10 shows that for chainwide promotions, the pricing error made by a reactive hotel chain ranges from 0.1% (2.5th percentile) to 11.1% (97.5th percentile) with an average of 4.3%. It leads to a similar range of growth target missed, 0.1% (2.5th percentile) to 11.4% (97.5th percentile) with an average of 4.4%, under the 5% growth target. For individual hotel promotions, the average scale of the pricing error (2.1%) and missed target (2.6%) is smaller; however, some hotels may suffer more significantly than others (pricing errors up to 19.9% and missed target up to 23.1%). We observe similar results under 10% and 20% growth targets. We would like to note that the price elasticity estimates obtained from our clickstream data are conservative as we observe only clicks, but not the final choice and due to potential price endogeneity. The real price elasticity in the hotel industry can be two to eight times larger (see, e.g., Newman et al. 2014, Anderson and Xie 2011, Lederman et al. 2014), in which case the scale of pricing errors and missed target will be significantly higher as well.

9. Conclusions

To understand how firms compete in markets with a large number of players is challenging in two ways. First, a large number of instruments are required to demonstrate pairwise causality. Second, dimensionality makes it hard to estimate a global competition model without imposing any structure or sparsity on the problem. We propose a methodology that combines the use of instrumental variables with high dimensional variable selection to resolve these issues. The proposed approach also requires minimal assumptions on market structure and the equilibrium at play. It allows us not only to uncover competition patterns in large fragmented markets and offer direct managerial implications to many small business players but also to inform future theoretical modeling efforts that attempt to better characterize competition in these nuanced markets.

Applying the methodology to the New York City hotel market, we made interesting observations about competition. First, engagement in competition-based revenue management is prevalent across brands (and non-brands) and across all quality tiers. Second, price competition for branded hotels is more constrained by quality rather than geographical boundaries, compared with independent hotels. Price competition of budget and luxury hotels is also more constrained by quality rather than geographical boundaries, as compared to economy and upscale hotels. Last, we find that when hoteliers react to competitors' prices, they tend to miss those potential competitors who are geographically farther away or are of different quality tiers (i.e., different star levels).

Practically, our methodology can be readily applied by online travel agencies or search aggregators to provide value-added services to hotel owners driven by large-scale data analytics. It would allow hotel owners to become more aware of competitors' moves and manage competition proactively. With conservative estimates of (self and cross)-price elasticities obtained from a random coefficient multinomial logit model allowing for rich demand substitution patterns, we find that accounting for competitive responses will lead to more accurate price/promotion decisions (avoiding up to 11%–20% pricing errors and 11%–23%

missed growth targets for certain hotels and chains). Given our findings, incorporating competition into revenue management and pricing models will be useful to address the rising need for intelligent reactions to competition in various industries.

Our analysis is, of course, not without limitations. First, the lack of hotel inventory data means that we are not able to directly control for the impact of inventory on price. Because of the practice of revenue management in the hotel industry, inventory is an important determinant of the dynamic changes in price (Talluri and van Ryzin 2005). Accounting for inventory levels in the price equations will better allow researchers to control for unobserved correlated shocks that may affect the performance of the instrument. Second, we only have access to lowest available room rates rather than rates of all different room types that a hotel may offer. It would be interesting to see how competition patterns may vary for different room types (i.e., different market segments).

Acknowledgments

The authors thank the corporate sponsor for making the data available. They also thank department editor Vishal Gaur, the anonymous associate editor, and three anonymous referees as well as seminar and conference participants for their valuable comments that helped improve this paper.

Endnotes

[1] A hotel may specify multiple room rates based on different room types. One could model each rate separately if data are available. We only have access to the lowest available rate through the data sponsor. We find, however, that a hotel's rates of different room types are highly correlated—the piecewise correlation ranges from 83.4% to 99.8% between rates of two-double-bed, queen-bed, and king-bed rooms and suites, based on the data set published by Bodea et al. (2009), which contains bookings and available rates for five different hotels.

[2] We model price competition using simultaneous price equations. We do not use sequential price equations because estimation of such a model requires even more frequent price data to accurately characterize which firm moves first.

[3] Hotel Comp Set Analysis—Untapped Opportunity. Hotel Compete. May 16, 2012.

[4] Another method that reduces dimensionality is the principal components approach (PCA). We choose our method over PCA because (1) PCA extracts orthogonal variations from a collection of correlated variables, but our main objective is to identify who follows whom in price competition. (2) To establish causality, the setup of the LASSO (linear, minimization of sum of squared errors) makes it very appealing to be combined with the instrumental variables. However, it is unclear how to establish causality in the PCA framework.

[5] We make several interesting conclusions based on our estimation on the simulated data. Although out of the scope of this paper, it would be an interesting future research direction to derive asymptotic properties of the proposed estimator and provide theoretical justifications to the fine-tuning procedure.

[6] Also note that when the residual is selected but not the original variable, it means that the two prices are correlated through other channels such as correlated demand, but the focal hotel does not actively follow the other hotel's prices.

[7] The online travel aggregator who sponsors the data for this research operates as an information aggregator rather than a merchant. In other words, it retrieves price information from multiple central reservations systems but does not set prices itself.

[8] Given that hotels' dynamic inventory positions are not publicly available, we cannot account for changes in inventory when explaining price variations. Therefore, inventory changes will be in the error terms. If these inventory changes are idiosyncratic in nature, it will not affect the performance of the instrument. Or if they are correlated across hotels as a result of correlated demand, such correlation will not affect the performance of the instrument either once the exposure (underlying demand) of each hotel is controlled. If, however, they are correlated as a result of other supply-side reasons, which are then correlated with demand, it may affect the performance of the instrument. See Online Appendix EC.3 for detailed discussions. We would also like to note that the inventory level is private information to each individual hotel. Therefore, our analysis is practical in that it uses exact same data a hotel manager (or a travel agency) is likely to have.

[9] An agreement by and among the Attorney General of the State of Connecticut, LQ Management L.L.C., and La Quinta Franchising, L.L.C., March 2010.

[10] There can also be a concern that the search engine's ranking algorithm will take into account the prices charged by different hotels and hence introduces an indirect link between prices and exposures. We alleviate the concern through the following steps. First, we focus on exposures, which are less affected by rankings rather than clicks. Especially since many consumers engage in filtering (44.0%) and re-sorting (33.5%), the resulting exposures are more driven by consumer preferences and less by the original search engine ranking. Second, we exclude all exposures that can be directly affected by prices—that is, re-sorting or filtering based on prices. Finally, instead of using the same-day exposures, we chose to use lagged exposures (last-week exposures *excluding* the booking date exposures), such that even if the search engine generated rankings based on the prices offered on the given booking date, it will not affect our measure because the same-day exposures are excluded.

[11] For the complete set of 177 hotels, the estimation takes 30.3 hours on an Intel Core i7 computer with 32 GB memory, with the use of glmnet package developed by Friedman et al. (2010) in R.

[12] We note that the percentage of the same or adjacent star levels is low for five-star hotels, which is partly caused by the few number of five-star hotels to begin with. When we estimate the likelihood to competing with hotels from the same stars using a logit model in Table 8, five-star hotels are actually more likely to compete with same quality tier hotels than hotels with three or four stars. We also tried to refine the model by imposing sparsity structures based on similarity in hotel characteristics. For example, if we impose a large penalty on a hotel two or more stars apart, the percentage of same or adjacent star level hotels increases to 61.9% for five-star hotels. However, we choose to present the original unstructured results to demonstrate the relevancy of the least constrained model.

[13] We would like to note a few caveats when interpreting the results. First, we do not observe the actual choices of consumers in our clickstream data, because our data sponsor is a travel aggregator and does not allow direct bookings on its website. Therefore, our demand model is estimated using clicks as proxies for final choices (see details in Online Appendix EC.6). Second, demand estimated using these data best represents the behaviors of consumers who book hotels through online channels—in other words, the leisure and transient business segment. Therefore, it may not represent all types of demand. Finally, the clicks used to estimate the demand model are low-frequency rare events; this makes it difficult to estimate a structure-free parsimonious demand model. Rather, the demand model, and hence the cross-price elasticities, are estimated with preimposed structures from the random-coefficient multinomial logit model. For these reasons, the observed mismatches should not

necessarily be interpreted as hoteliers' mistakes. However, it is an interesting observation and one that requires future research for further validation.

References

Adida E, Perakis G (2010) Dynamic pricing and inventory control: Uncertainty and competition. *Oper. Res.* 58(2):289–302.

Anderson CK, Xie XQ (2011) A choice-based dynamic programming approach for setting opaque prices. *Production Oper. Management* 21(3):590–605.

Anderson SP, De Palma A, Thisse J-F (1989) Demand for differentiated products, discrete choice models, and the characteristics approach. *Rev. Econom. Stud.* 56(1):21–35.

Ang E, Kwasnick S, Bayati M, Plambeck EL, Aratow M (2016) Accurate emergency department wait time prediction. *Manufacturing Service Oper. Management* 18(1):141–156.

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

Belobaba PP (1989) Application of a probabilitics decision model to airline seat inventory control. *Oper. Res.* 37(2):183–197.

Bensinger G (2015) Amazon's third-party merchants are a growing piece of the sales pie. *Digits* (blog) (January 5), https://blogs.wsj.com/digits/2015/01/05/amazons-third-party-merchants-a-growing-piece-of-the-sales-pie/.

Bodea T, Ferguson M, Garrow L (2009) Choice-based revenue management: Data from a major hotel chain. *Manufacturing Service Oper. Management* 11(2):356–361.

Boyd EA (2007) *The Future of Pricing: How Airline Ticket Pricing Has Inspired a Revolution* (Palgrave-Macmillan, New York).

Bühlmann P, van de Geer S (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications* (Springer, New York).

Chatterjee A, Lahiri SN (2011) Bootstrapping Lasso estimators. *J. Amer. Statist. Assoc.* 106(494):608–625.

Cross RG, Higbie JA, Cross DQ (2009) Revenue management's renaissance: A rebirth of the art and science of profitable revenue generation. *Cornell Hospitality Quart.* 50(1):56–81.

Feenstra RC, Levinsohn JA (1995) Estimating markups and market conduct with multidimensional product attributes. *Rev. Econom. Stud.* 62(1):19–52.

Fisher M, Gallino S, Li J (2018) Competition-based dynamic pricing in online retailing: A methodology validated with field experiments. *Management Sci.* 64(6):2496–2514.

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* 33(1):1–22.

Gabszewicz JJ, Thisse JF (1979) Price competition, quality, and income disparities. *J. Econom. Theory* 20(3):340–359.

Gallego G, Hu M (2014) Dynamic pricing of perishable assets under competition. *Management Sci.* 60(5):1241–1259.

Gallego GR, van Ryzin G (1994) Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Marketing Sci.* 40(8):999–1020.

Hausman JA (1978) Specification tests in econometrics. *Econometrica* 46(6):1251–1271.

Hotelling H (1929) Stability in competition. *Econom. J.* 39(153):41–57.

Koulayev S (2014) Search for differentiated products: Identification and estimation. *RAND J. Econom.* 45(3):553–575.

Lederman R, Olivares M, van Ryzin G (2014) Identifying competitors in markets with fixed product offerings. Working paper, Columbia Business School, New York.

Li J, Granados N, Netessine S (2014) Are consumers strategic? Structural estimation from the air-travel industry. *Management Sci.* 60(9):2114–2137.

Martínez-de-Albéniz V, Talluri K (2011) Dynamic price competition with fixed capacities. *Management Sci.* 57(6):1078–1093.

Netessine S, Shumsky RA (2005) Revenue management games: Horizontal and vertical competition. *Management Sci.* 51(5):813–831.

Newman JP, Ferguson ME, Garrow LA, Jacobs TL (2014) Estimation of choice-based models using sales data from a single firm. *Manufacturing Service Oper. Management* 16(2):184–197.

Olivares M, Cachon GP (2009) Competing retailers and inventory: An empirical investigation of General Motors' dealerships in isolated U.S. markets. *Management Sci.* 55(9):1586–1604.

Pinkse J, Slade ME, Brett C (2002) Spatial price competition: A semiparametric approach. *Econometrica* 70(3):1111–1153.

Rudin C, Vahn G-Y (2016) The big data newsvendor: Practical insights from machine learning. Working paper, London Business School, London.

Ryzhov IO, Han B, Bradic J (2015) Cultivating disaster donors using data analytics. *Management Sci.* 62(3):849–866.

Salop S (1979) Monopolistic competition with outside goods. *Bell J. Econom.* 10(1):141–156.

Talluri KT, van Ryzin GJ (2005) *The Theory and Practice of Revenue Management* (Springer, New York).

Tereyağoğlu N, Fader PS, Veeraraghavan SK (2018) Multiattribute loss aversion and reference dependence: Evidence from the performing arts industry. *Management Sci.* 64(1):421–436.

Terza JV, Basu A, Rathouz PJ (2008) Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *J. Health Econom.* 27(3):531–543.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc.* 58(1):267–288.

Tirole J (1988) *The Theory of Industrial Organization* (MIT Press, Cambridge, MA).

Vulcano G, van Ryzin G, Chaar W (2010) Choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing Service Oper. Management* 12(3):371–392.

Wang H, Li R, Tsai C (2007) Tuning parameter selector for the smoothly clipped absolute deviation method. *Biometrika* 94(3):553–568.

Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. (MIT Press, Cambridge, MA).

Zou H (2006) The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* 101(476):1418–1429.