The Value of Field Experiments in Estimating Demand Elasticities and Maximizing Profit

by

Jimmy Qiuyuan Li

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Signature redacted

Author ....
Department of Electrical Engineering and Computer Science
May 16, 2014

Signature redacted

Certified by ....
Professor John N. Tsitsiklis
Clarence J. Lebel Professor of Electrical Engineering
Thesis Supervisor

Signature redacted

Certified by ....
Professor Duncan Simester
NTU Professor of Marketing
Thesis Supervisor

Signature redacted

Accepted by ....
Professor Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

The Value of Field Experiments in Estimating Demand Elasticities and Maximizing Profit

by

Jimmy Qiuyuan Li

Submitted to the Department of Electrical Engineering and Computer Science
on May 16, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

In many situations, the capabilities of firms are better suited to conducting and analyzing field experiments than to analyzing sophisticated demand models. However, the practical value of using field experiments to optimize marketing decisions remains relatively unstudied. We investigate category pricing decisions that require estimating a large matrix of cross-product demand elasticities and ask: how many experiments are required as the number of products in the category grows?

Our main result demonstrates that if the categories have a favorable structure, then we can learn faster and reduce the number of experiments that are required: the number of experiments required may grow just logarithmically with the number of products. These findings potentially have important implications for the application of field experiments. Firms may be able to obtain meaningful estimates using a practically feasible number of experiments, even in categories with a large number of products.

We also provide a relatively simple mechanism that firms can use to evaluate whether a category has a structure that makes it feasible to use field experiments to set prices. We illustrate how to accomplish this using either a sample of historical data or a pilot set of experiments. Historical data often suffer from the problem of endogeneity bias, but we show that our estimation method is robust to the presence of endogeneity.

Besides estimating demand elasticities, firms are also interested in using these elasticities to choose an optimal set of prices in order to maximize profits. We formulate the profit maximization problem and demonstrate that substantial profit gains can also be achieved using a relatively small number of experiments.

In addition, we discuss how to evaluate whether field experiments can help optimize other marketing decisions, such as selecting which products to advertise or promote. We adapt our models and methodologies to this setting and show that the main result that relatively few experiments are needed to estimate elasticities and to increase profits continues to hold.

Thesis Supervisor: Professor John N. Tsitsiklis
Title: Clarence J. Lebel Professor of Electrical Engineering

Thesis Supervisor: Professor Duncan Simester
Title: NTU Professor of Marketing

Acknowledgments

I have been incredibly fortunate to have Prof. John Tsitsiklis and Prof. Duncan Simester as my advisors. Over the past five years, they have provided me with tremendous support, guiding me not only through research and academics but also through graduate school life. Their invaluable suggestions and encouragement inspired me to consider new ideas and perspectives and ultimately led to the successful completion of this thesis. I am so very grateful to have had the opportunity to work with and learn from them.

I would like to thank Prof. Patrick Jaillet, who has served on both my thesis committee and my RQE committee, for his helpful feedback and suggestions. I would also like to thank Prof. Paat Rusmevichientong and Spyros Zoumpoulis, with whom I have collaborated and had many insightful discussions over the years. My doctoral studies and this thesis were partially supported by NSF grants CMMI-0856063 and CMMI-1158658, for which I am very grateful.

My time in graduate school has been unforgettable because of all the people I've met and befriended and the experiences we've shared. Thanks to everyone in the SyNDeG group for the intellectual inspiration and the good times, especially during our ski trips. I am also grateful for everyone in the LIDS, RLE, ORC, and Sloan Marketing communities for welcoming me and helping me through this journey.

It feels like a long time ago now, but I did in fact have a life before graduate school. Thank you to all of my friends for supporting me through the years, making sure I didn't become a hermit, and reminding me when times where tough that there is a light at the end of the tunnel.

Finally, my deepest gratitude goes to my parents and family for their unconditional love and support. Thank you for your unwavering belief in me and for all that you've done to give me the opportunities that I've had. This thesis is dedicated to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The increased availability of demand data has been widely reported and many firms have been investigating how best to use "Big Data" to improve their marketing decisions. One option is to conduct analyses on historical data. However, historical data are not always available, and it can be difficult to determine causation from historical data. An alternative approach is to use field experiments, which can provide an exogenous source of variation that establishes causation. Yet conducting field experiments is often costly, and optimizing marketing decisions may require a lot of experiments if there are many parameters to optimize and/or if the parameters can take a wide range of values. The feasibility of using field experiments to improve marketing decisions in practice remains relatively unstudied. We investigate this issue by considering settings in which firms must estimate the elasticity of demand in response to price changes. We ask how many experiments are required to estimate these elasticities as the number of products grows.

Using experiments to optimize marketing decisions may be relatively straightforward when there are few products. Experimentally manipulating variables can allow retailers to optimize their decisions using just a handful of experiments. However, in large categories containing many products with interdependent demands, the problem is more challenging.[1] The number of parameters to estimate grows quickly with the number of

[1] Interdependencies between products are now well-documented. For example, Anderson and Simester (2001) report that placing "sale" signs on products can increase demand for those prod-

products, and so the number of field experiments required may be impractically large.

We consider a large set of n products and assume that there may be complementary or substitute relationships between them. As a result, varying the price of one product may affect the demand not just of that item but also of other products sold by the firm. As the number of products, n, increases, the number of parameters to estimate grows at the rate of n^2 (and may grow even faster for nonlinear models). On the other hand, if an experiment reveals the demand for each item, we learn n pieces of information from each experiment. This suggests that the number of experiments required to learn all of the parameters will grow at least linearly with the number of products.

Our main result shows that if the problem has a favorable structure, we can learn faster and reduce the number of experiments that are required. In particular, we will show that if the number of complementary or substitute relationships affecting any one product is bounded, then the number of required experiments instead grows logarithmically with the number of products. This result holds even if the firm is not sure which particular pairs of products have complementary or substitute relationships, as long as there is a bound on the number of cross-product relationships that each product has. We also obtain a similar result if the joint impact of own- and cross-product effects on any single product is bounded.

Assuming that such a favorable structure exists, we show that we can learn the set of elasticities quickly. But how do we know if a favorable structure exists? To answer this question, we provide a practical method for evaluating whether a product category has a favorable structure that makes it feasible to use field experiments to set category prices. Although the method is probably too technical to be used directly by most managers, the techniques should be accessible to analysts tasked to provide advice to managers on this issue. The method does not provide an estimate of how many experiments

ucts by up to 60%, but can decrease sales of other products by similar amounts. Manchanda et al. (1999) report own-price elasticities for laundry detergent and fabric softener of -0.40 and -0.70, respectively. The cross-price elasticities are -0.06 (the price of softener on demand for detergent) and -0.12. For cake mix and frosting, the own-price elasticities are -0.17 and -0.21, respectively, while the cross-price elasticities are -0.11 (frosting price on cake mix demand) and -0.15.

are required. Instead, it provides a means of estimating whether the product category exhibits structural characteristics that make it possible to obtain accurate results within a realistic number of experiments. In particular, we propose a method for estimating bounds on the number and size of interdependencies between products. The method can be implemented using a pilot set of experiments or using historical data. Using synthetic data, we verify that this method can recover the correct structural bounds via simulations. We also apply this method to a sample of real sales data from the "Cold remedies" category. Our empirical results suggest that the "Cold remedies" category does exhibit a favorable structure, and therefore the elasticity parameters can indeed be feasibly estimated using relatively few experiments.

Historical data are often readily available and a more convenient alternative to running costly field experiments. However, care must be taken when using historical price and demand data to estimate elasticities because, unlike with randomized field experiments, historical prices may not have been set exogenously. For example, some event could have led to both store managers' adjusting prices and a simultaneous demand shock, not necessarily completely due to the price change. The resulting endogeneity can lead to biased elasticity estimates by misattributing some of the change in demand to the change in price. We account for potential endogeneity in our historical data by taking an instrumental variables approach and show that our estimation methodology is in fact robust to endogeneity.

Finally, we move from estimating demand elasticities to making optimal pricing decisions and maximizing profit. Taking the resulting estimates of elasticities from our estimation procedure, we show that a relatively straightforward profit maximization algorithm can lead to substantial gains in profit, even with a relatively small number of experiments.

These findings potentially have important implications for the application of field experiments in settings where there is a large number of parameters to estimate. Because the number of required experiments may grow logarithmically rather than linearly with the number of products, firms may be able to obtain meaningful estimates and make profitable decisions using a realistically small number of experiments, even in categories with a large number of products.

Although we focus on pricing decisions, the range of marketing decisions on which firms can experiment is broad. Experiments may be used to choose which products to promote, as well as to optimize the length of product lines and to choose creative copy and media plans. We discuss how to extend our results to making promotional decisions, and in the Conclusions we discuss possible extensions to other types of marketing decisions.

1.1 Related work

The feasibility of learning a large number of parameters through experimentation is relatively unstudied, particularly in social science settings. However, the topic does relate to at least two literatures.

1.1.1 Optimal experimental design

First, there is the line of research on optimal experimental design. In the marketing literature, there is work focusing on efficient experimental design for conjoint studies (see Louviere et al. 2000, Chapter 5; and Louviere et al. 2004 for reviews of this literature). Recent contributions to this literature have focused on adaptively designing experiments (Toubia et al. 2003) or on optimal designs when customers' utility functions depart from a standard compensatory specification (see, for example, Hauser et al. 2010, Liu and Arora 2011). An often used measure of the efficiency of an experimental design is the D-error: $\det[I(\theta \mid X)]^{-1/m}$, where I is the information matrix, θ are the unobserved parameters, X is the experimental design matrix, and m is the dimension of I. The information matrix is calculated from the variance of the first-order derivatives of the log-likelihood with respect to θ (Huber and Zwerina 1996). Optimizing this criterion with respect to X yields locally optimized designs for any θ. Because θ is not known when designing the experiments, Bayesian approaches can be used to minimize the D-error over the prior distribution of the parameter values (Sandor and Wedel 2001).

When each experiment generates an explicit reward or cost, an alternative formulation of the experimental design problem is as a multi-armed bandit problem, where the objective is to choose a sequence of experiments to maximize the total reward over some time horizon. In this context, each experiment can be thought of as choosing and pulling an arm of the multi-armed bandit, and the reward could be sales, advertising click-through rates, or some other measure. Because we learn the reward distribution of each arm of the bandit only after pulling it, there exists a trade-off between *exploiting* the best arm currently known by pulling it every time and *exploring* new arms in search of something even better. In the classic bandit model, the reward distributions of each arm are assumed to be independent, and so anything learned from pulling one arm does not reveal anything about a different arm. As a result, when there is a large number of parameters (and therefore a large number of arms), many pulls, or experiments, are required to learn the reward distributions of all the arms. Recent work has proposed an alternative model in which the arms have statistically dependent reward distributions, and therefore pulling one arm also gives information about other arms. In this setting, the correlation between payoffs of different arms allows for faster learning, even when the number of arms is very large (Dani et al. 2008, Mersereau et al. 2009).

The focus on the information learned from experiments is a common feature of both this literature and the research in this thesis. However, we do not focus on identifying optimal experimental designs. Instead we use random experimental designs, which ensure independence across experiments and allow us to apply a series of results that rely on this independence. Because it will generally be possible to improve upon these designs, our guarantees on the information learned will continue to hold when optimal designs are used.

We investigate the practical value of field experiments by studying the number of experiments required. Other studies have also investigated the required size of field experiments. For example, Lewis and Rao (2012) conducted a set of 25 field experiments involving large display advertising campaigns, each one including over 500,000 unique users and totaling over $2.8M worth of impressions. Even with such large experiments, the data generated little meaningful information about the ROI of the campaigns, demonstrating that in settings where the effect sizes are small and the response measures are highly stochastic, very large field experiments may be required to generate useful information.

1.1.2 Estimation under sparsity

The second related literature is that on estimation and learning under assumptions of sparsity. Beginning with variable selection in regressions, research has focused on determining which subset of potential predictors should be included in the "best" model. This can equivalently be thought of as setting the coefficients associated with a subset of predictors to zero, thereby giving rise to a sparse model. Various approaches have been proposed, including the use of regularization, such as the "Lasso" of Tibshirani (1996) and the Stochastic Search Variable Selection procedure developed in George and McCulloch (1993).

More recently, the assumption of sparse structures has been used to show that if an unknown vector $\mathbf{x} \in \mathbb{R}^N$ is sparse, then it can be recovered using measurements of the form $\mathbf{y} = \Phi\mathbf{x}$, even with much fewer than N measurements. Results in the field, which is often referred to as "compressive sensing", generally provide conditions on (i) the sparsity index (i.e., the number of nonzero entries of $\mathbf{x}$), (ii) the number of measurements, and (iii) the ambient dimension N, in order to guarantee recovery of $\mathbf{x}$. We refer the reader to Candès (2006) for a short survey and to Candès et al. (2006), Candès and Tao (2005) for a deeper treatment.

More directly relevant to our work are the results on information-theoretic limits of sparsity recovery in Wainwright (2009). For a noisy linear observation model based on sensing matrices drawn from the standard Gaussian ensemble, a set of both sufficient and necessary conditions for asymptotically perfect recovery is derived. Our theoretical findings are best thought of as an application of the results in Wainwright (2009). An exception is the estimation of the sparsity parameters in Chapter 4 and the investigation of how these parameters vary with the size of the problem (i.e., the number of products). This is the first work of which we know that addresses these issues.

Originating from and motivated by applications in signal processing, coding theory, and statistics, compressive sensing results have a variety of other relevant applications. Previous applications related to marketing include Farias et al. (2013), which introduces a paradigm for choice modeling where the problem of selecting an appropriate choice model (either explicitly, or implicitly within a decision-making context) is itself automated and data-driven. For this purpose, the sparsest choice model consistent with observed data is identified.

In this work, we leverage sparsity to obtain a dramatic improvement in the rate of learning. If each product is substitutable by or complementary with a limited number of other products, we show that the number of required experiments grows logarithmically with the number of products.

1.2 Overview

We consider pricing decisions for a firm with a large assortment of products. The firm would like to know how price changes will affect demand. We propose a model for the demand function, which tells us the quantities demanded under any pricing decision. In order to learn the parameters of this function, we perform experiments by varying the prices of certain products and observing the quantities demanded. Because each experiment is costly to run, the firm would like to learn the parameters using as few experiments as possible.

The experiments that we contemplate include both a treatment group and a control group. The construction of these groups will vary depending on the nature of the firm. For a direct marketing firm, the groups may be constructed by randomly assigning individual customers to the two groups. For a brick-and-mortar retailer, the groups might be constructed by randomly assigning stores. In a business-to-business setting, the firm might randomly assign regions, distributors, or resellers. We assume that the results of the experiment are analyzed by aggregating the customers in each group and comparing the mean response between the two groups. Essentially all firms are capable of performing this aggregate analysis (as long as they can vary prices and measure the response).[2] This aggregation also ensures that the error terms can be modeled as Gaussian.

Our findings can also apply to settings where the firms vary prices across different time periods. Demand in the different time periods could in principle be adjusted to account for seasonality or day-of-week differences (before submitting the data to our model), perhaps using demand for a sample of unrelated products or demand in different stores. We caution that we will assume that errors are independent across experiments (though not across products within the same experiment), and this independence assumption may be threatened when a common set of measures is used to adjust for seasonality. The independence assumption is more likely to hold when randomization occurs separately for each experiment, and when the control group provides an accurate control for any intervening events (such as seasonality).

We also caution that our results are not well-suited to experiments where firms randomly assign products to treatment and control groups if the demands for those products are possibly related. For example, a firm may vary prices on half of the items in a product category and leave the other half of the prices unchanged. Recall that the goal of this thesis is to investigate how a firm can estimate the entire matrix of cross-price elasticities, and so the second half of the products cannot function as controls. There is another reason to be concerned about this experimental design: unless the cross-price elasticities are zero between products in the two groups, the experimental manipulation of prices in the treatment group of products will confound the demands for products in the control group.

[2] Even though direct marketing firms can often analyze experimental results at the individual customer level, in our experience most firms simply aggregate the results and compare the mean response between treatment and control groups.

We recognize that it is possible to augment experimental data with more complex econometric analysis (e.g., as in Manchanda et al. 1999). This raises an interesting but distinct question: what is the value of sophisticated analyses in evaluating experimental data? This question is beyond the scope of the present work. Instead, our results can be interpreted as describing the "information" that is revealed by experimental data. Conditions under which experimental data are more informative are likely to yield better estimates both when using simple comparisons and when augmenting the data with sophisticated econometric analysis.

The rest of this thesis is structured as follows.

In Chapter 2, we propose a model for demand that captures the effects of cross-product demand elasticities.

In Chapter 3, we develop a method for estimating these elasticities and provide bounds on the number of experiments required to achieve accurate estimates under various structural assumptions on the demand model.

In Chapter 4, we investigate whether the structural assumptions we make are valid using real-world sales data. Our methodology also provides a practical way for managers to evaluate whether it is feasible to set prices using field experiments.

In Chapter 5, we present simulation results that support our theoretical bounds on the speed of learning.

In Chapter 6, we examine the presence of endogeneity in historical data and modify our estimation methodology to use an approach based on instrumental variables in order to account for endogeneity. Our results suggest that our estimation method is robust to endogenous data.

In Chapter 7, we extend the problem from estimating elasticities to maximizing profit and propose an algorithm that achieves substantial profit gains with relatively few experiments.

In Chapter 8, we consider the alternative setting of choosing which products to promote or advertise. We adapt our model to this promotional setting and show that our results hold in this setting as well.

Finally, in Chapter 9, we conclude and describe directions for extensions and future research.

Chapter 2

Demand model

In this chapter, we introduce our model of demand. Throughout this thesis, we consider each experiment as a comparison between two conditions. The first condition is a control under which the firm takes "standard" actions; in the second, treatment condition, the firm varies prices. For ease of exposition (and without loss of generality), we assume that prices are set at a "baseline" level in the control condition.

2.1 Modeling own- and cross-price elasticities

The response in demand to a firm's action is difficult to predict because there are multiple effects at play due to cross-product substitute and complementary relationships. In the following sections, we present a model that captures these effects.

2.1.1 Individual and pairwise effects

Changing the price of product i may have two effects:

(i) It may affect the demand for the product itself.

(ii) It may also affect the demand for other products through substitution away from the focal product or complementarity with the focal product.

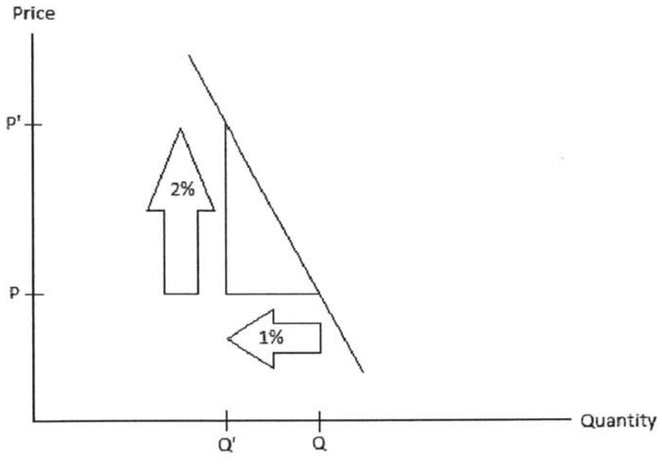


Figure 2-1: An illustration of a demand curve and price elasticity. Increasing price by 2% from P to P' results in quantity (or demand) decreasing by 1% from Q to Q'. The associated price elasticity is $-1\%/2\% = -1/2$.

For the first effect, we introduce a quantity a_{ii} to represent the percentage change in demand for product i per marginal percentage change in the price of product i itself. Figure 2-1 illustrates a demand curve and this definition of price elasticity. These percentage changes in demand and in price are measured with respect to the baseline levels under the control condition.

For the second effect, we first consider a pair of products in isolation. Intuitively, there are three possible scenarios:

1. If products i and j are substitutes, decreasing the price of j may decrease the demand for i if customers substitute purchases of j for purchases of i.

2. If i and j are complements, decreasing the price of j may increase the demand for i as more demand for j leads to more demand for i.

3. Varying the price of j may also have no effect on the demand for i.

For each pair of products i and j, we introduce a quantity a_{ij} to represent the percentage change in demand for product i per marginal percentage change in the price of product

j. The quantity a_{ij} would be positive, negative, and zero, in cases 1, 2, and 3 above, respectively. This definition of the a_{ij}'s matches the usual definition of price elasticity of demand (e.g., Mas-Colell et al. 1995).

2.1.2 Cumulative effects

We are interested in settings in which there are scores of products with hundreds of interactions at play. If multiple prices are varied simultaneously, how do these changes combine and interact to produce an overall effect on demand?

To capture the cumulative effects, we propose a linear additive model of overall substitution and complementarity effects. Specifically, to calculate the *overall* percentage change in demand for product i, we take all of the products j whose prices are varied and sum together each one's individual effect on the demand for i.

Let Δq_i be the overall percentage change in the demand for i, and let us express the percentage change in the price of product j from the baseline as

$$x_j = \frac{x_j^t - x_j^b}{x_j^b},$$

where x_j^t and x_j^b are the treatment and baseline (i.e., control) prices, respectively, of product j. We denote the number of products by n. Then, by our model, we can write the overall percentage change in demand for i as

$$\Delta q_i = \sum_{j=1}^{n} a_{ij} x_j.$$

We can further simplify notation by collecting all of the pairwise effects as elements of a matrix $\mathbf{A}$, where (as suggested by the notation) the entry in the i^{th} row and j^{th} column, a_{ij}, gives the percentage change in demand for product i per marginal percentage change in the price of product j. Similarly, we can collect price variation decisions into a vector $\mathbf{x}$ whose j^{th} element x_j is equal to the percentage change from the baseline in the price of product j, and we can also collect the overall percentage change in demand for each product into a vector $\Delta \mathbf{q}$. The overall percentage change in each product's demand due to price changes $\mathbf{x}$ is therefore given by the product

$$\Delta \mathbf{q} = \mathbf{A}\mathbf{x}.$$

We do not impose symmetry (i.e., $a_{ij} = a_{ji}$) or transitivity (i.e., $a_{ij} > 0, a_{jk} > 0 \Rightarrow a_{ik} > 0$) on the $\mathbf{A}$ matrix for two reasons. First, there are examples where these constraints are intuitively unlikely to hold: e.g., price decreases on cameras may increase battery sales but not vice versa, violating symmetry; price decreases on milk may increase sales of cereal, and price decreases on cereal may increase sales of soymilk, but price decreases on milk may not increase sales of soymilk, violating transitivity. Second, neither symmetry nor transitivity is a necessary assumption for our analysis, and imposing these constraints would only make our results weaker and less applicable. Instead, we want the space of "allowable" $\mathbf{A}$ matrices to be as large as possible. Furthermore, if the true $\mathbf{A}$ matrix is indeed symmetric or transitive, then because our method gives accurate estimates, the estimated matrix would also be close to symmetric or transitive with high probability.

We also assume that the matrix $\mathbf{A}$ is constant. It is possible that there may be time dependencies or seasonal effects that could lead to changes in the $\mathbf{A}$ matrix. The model could accommodate these possibilities as long as these dynamics are known so that we can continue to estimate a static set of parameters. If the parameters themselves change in a manner that is not known, then the results of an experiment performed at some time t may not provide much information about the value of the parameters in future periods. Note that this limitation is obviously not specific to our model.

We emphasize that the matrix $\mathbf{A}$ captures *percentage changes* in demand. To calculate actual demand quantities, we also need a baseline level of demand for each product. Recall that we assume there is a fixed set of firm actions, corresponding to the control condition, which achieves a certain level of demand. We let this be the baseline level of demand and denote it by the vector $\mathbf{q}^b$. The overall change in demand for a product in response to the price changes is then given by the product of the baseline demand and the percentage change in demand.

2.2 Noiseless model

Let $\mathbf{q}^t$ be the vector of actual demand levels in response to a decision $\mathbf{x}$, which we refer to as the *treatment* demand level. We then have the following equation for our model:

$$\mathbf{q}^t = \mathbf{q}^b + \mathbf{q}^b \circ (\Delta\mathbf{q}) = \mathbf{q}^b \circ (\mathbf{e} + \mathbf{A}\mathbf{x}), \tag{2.1}$$

where $\circ$ denotes component-wise multiplication, and $\mathbf{e}$ is the vector of all 1's. We can also rewrite Equation (2.1) as

$$\Delta\mathbf{q} = \frac{\mathbf{q}^t - \mathbf{q}^b}{\mathbf{q}^b} = \mathbf{A}\mathbf{x}, \tag{2.2}$$

where the division is performed component-wise. The left-hand-side gives the percentage change in demand for each product, and the right-hand-side gives the model of how that change is caused by the decision vector. This form suggests a way of learning $\mathbf{A}$: For each experiment, choose a decision vector $\mathbf{x}$, observe the resulting $\mathbf{q}^b$ and $\mathbf{q}^t$, and calculate $\Delta\mathbf{q}$. This gives a system of linear equations from which we can recover $\mathbf{A}$, ideally using as few experiments as possible.

2.3 Noisy model

In reality, the demand function is not captured perfectly by Equation (2.1), and the demand that we observe will also be subject to measurement noise. We model this error with an additive term $\mathbf{w}$, which is a vector of random variables $(w_1, w_2, \ldots, w_n)$. Our complete model is then given by

$$\mathbf{q}^t = \mathbf{q}^b \circ (\mathbf{e} + \mathbf{A}\mathbf{x} + \mathbf{w}), \tag{2.3}$$

Term	Description
$\mathbf{A}$	A matrix capturing the substitution and complementarity effects: the element a_{ij} represents the percentage change in demand for product i per marginal percentage change in the price of product j
$\mathbf{x}^t$	A vector of treatment prices
$\mathbf{x}^b$	A vector of baseline prices
$\mathbf{x}$	A decision vector, whose entries are percentage changes in price from the baseline
$\mathbf{w}$	The random error or noise vector
$\mathbf{q}^t$	The treatment demand
$\mathbf{q}^b$	The baseline demand, which is assumed to be known from the control condition
$\hat{\mathbf{A}}$	An estimate of the true matrix $\mathbf{A}$
n	The number of products
s	The number of experiments

Table 2.1: Summary of notation

which can also be written as

$$\Delta\mathbf{q} = \frac{\mathbf{q}^t - \mathbf{q}^b}{\mathbf{q}^b} = \mathbf{A}\mathbf{x} + \mathbf{w}. \tag{2.4}$$

The functional form $\Delta\mathbf{q} = \mathbf{A}\mathbf{x} + \mathbf{w}$ is convenient for analytical tractability. However, our analysis does not place any limitations on how $\Delta\mathbf{q}$ is defined. Indeed, we could use different variations, including alternatives that ensure symmetry in the measures of demand increases and decreases. Table 2.1 summarizes the relevant notation used in our model.

2.3.1 Statistics of the noise terms

For our analysis, we make the following assumptions on the noise terms.

Assumption 1 (Zero-mean, sub-Gaussian noise, i.i.d. across experiments). *For any experiment, each w_i has zero mean and is sub-Gaussian with parameter c for some constant $c \geq 0$. Furthermore, the random vector $\mathbf{w}$ is independent and identically distributed across different experiments.*

We assume that the noise terms have zero mean, and therefore that our model has no systematic bias. We also assume that the noise terms across different experiments are independent and identically distributed. However, we do not assume that the noise terms are independent across different products within the same experiment. In other words, each experiment gets an independent draw of $\mathbf{w} = (w_1, \ldots, w_n)$ from a single joint distribution in which the w_i's can be dependent. Indeed, the noise terms within the same experiment may be correlated across products (e.g., between products within the same category). Fortunately, our analysis does not require independence at this level.

Sub-Gaussian random variables are a generalization of Gaussian random variables, in the sense that their distributions are at least as concentrated around their means as Gaussian distributions.

Definition 1. A random variable X is *sub-Gaussian* with parameter $\sigma > 0$ if

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(\sigma^2\lambda^2/2), \quad \forall\lambda \in \mathbb{R}.$$

A sub-Gaussian random variable X with parameter σ satisfies the following concentration bound:

$$\mathbf{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right), \quad \forall\epsilon \geq 0.$$

As suggested by the notation, the parameter σ plays a role similar to that of the standard deviation for Gaussian random variables. Examples of sub-Gaussian random variables with parameter σ include Gaussian random variables with standard deviation σ and bounded random variables supported on an interval of width 2σ. Hence, by using sub-Gaussian noise terms, we encompass many possible distributions. In all cases, sub-Gaussianity allows us to bound the concentration of the noise around its mean.

2.4 Limitations and extensions of the model

Before continuing, let us briefly discuss some of the limitations and possible extensions of our model.

By assuming a linear model, we are implicitly assuming that the elasticities are the same at all points along the demand curve. Although this may be appropriate for small price changes, it is unlikely to be true when price changes are relatively large. However, we can ensure that price changes are small by bounding the magnitude of permissible price changes in the treatment conditions. However, we caution that this is not without cost: greater variation in the size of price changes can increase the rate of learning. More generally, we can interpret our linear model as an approximation to the true model in the neighborhood around the baseline levels of price and demand, in the spirit of a first-order Taylor approximation.

The model also assumes additive separability in the impact of multiple price changes on the demand for any single product i. This is convenient for analytical tractability. In Appendix A, we show that it is relatively straightforward to extend our findings to a log-linear (multiplicative) demand model. Log-linear demand models have been widely used in practice, in both academia and the marketing analytics industry.

In some cases, a firm may want to focus on improving the prices of only a subset of products within a category. This could occur if some items sell relatively low volumes and optimizing these prices is not a priority (or if their retail prices are set by manufacturers). This may also arise if too many experiments are required to optimize the prices of all products in the category, and so the firm would like to focus on only those products that it considers most important. We can easily accommodate this possibility by identifying the products that the firm does not want to experiment with and collapsing these products into a single "other" product. Sales of this "other" product is simply the total sales of the products within it. We can also construct a price index for the "other" product by averaging the prices of the corresponding items. (Because the firm does not want to experiment with these prices, the value of the corresponding x_j's will always equal zero.) This allows the firm to focus on a subset of products in the category, while continuing to take into account the impact on sales across the entire category.

2.5 High-dimensional problem

Having presented our model, we emphasize the high-dimensional nature of the problem in more specific terms. In our model, with n products, $\mathbf{A}$ would be an $n \times n$ square matrix, and hence there would be n^2 unknown parameters to be estimated. Even with 50 products, a reasonable number for many product categories, there would be 2,500 parameters. In order to estimate all of these parameters accurately, we expect to need to perform many experiments.

Unfortunately, each experiment is costly to the firm in terms of not only time and resources needed to run it, but also opportunity costs. Therefore, our goal is to estimate the parameters accurately and to make good decisions using as few experiments as possible.

Although we are faced with a difficult problem, our main insight is that even though there are many products, each one is likely to interact with only a small subset of the remaining products. In terms of our model, this means that the $\mathbf{A}$ matrix is likely to have many entries equal to zero. Our main result shows that if $\mathbf{A}$ exhibits this sparse structure, we can greatly reduce the number of experiments needed to learn $\mathbf{A}$ and to find a good decision vector $\mathbf{x}$, even if the locations of the nonzero terms are not *a priori* known.

Chapter 3

Estimating the A matrix

In order to find an optimal set of firm actions, we will first estimate the substitute and complementary relationships between products, which are modeled by the matrix $\mathbf{A}$. In this chapter, we describe a general methodology for estimating $\mathbf{A}$, introduce our structural assumptions, present bounds on the number of experiments needed to learn $\mathbf{A}$ accurately, and discuss our results.

3.1 Random experimental design

Our goal is to learn $\mathbf{A}$ as quickly as possible, and so we would like to design experiments (i.e., $\mathbf{x}$ vectors) that give as much information as possible. One approach is to design decision vectors deterministically in order to maximize some orthogonality measure between decision vectors. However, because we do not make any assumptions about how the locations or values of the entries of $\mathbf{A}$ are distributed, for any deterministic design, there will be classes of $\mathbf{A}$ matrices for which the design is poor.

As an alternative, we use random experiments: the decision of how much to change the price of a particular product for a given experiment will be a random variable. Moreover, if we make these decisions independently across products and across experiments, we achieve approximate orthogonality between all of our experiments. By using randomization, we are also able to take advantage of the extensive body of probability theory and prove that we can learn every element of **A** to high accuracy with high probability, for *any* **A** matrix. Next, we describe our estimation procedure in more detail.

3.2 Unbiased estimators, convergence, and concentration bounds

For each parameter a_{ij}, we define a statistic y_{ij} that is a function of the random decision vector and the resulting (random) observed demands. This statistic is therefore also a random variable, and we design it so that its mean is equal to a_{ij}. In other words, we find an unbiased estimator for each parameter.

If we perform many independent experiments and record the statistic y_{ij} for each one, the laws of large numbers tell us that the sample mean of these statistics converges to the true mean, which is exactly the parameter a_{ij} that we are trying to estimate. This sample mean is a random variable, and its probability distribution will become more and more concentrated around a_{ij} as we collect more samples (i.e., perform more experiments). To get a sense of the speed of convergence, we calculate a bound on the concentration of the distribution around a_{ij} after each additional sample. This bound will in turn allow us to prove results on the number of experiments needed to achieve accurate estimates with high confidence.

3.3 Uniformly ϵ-accurate estimates

Our goal is to learn the **A** matrix accurately to within a certain bound with high probability. To be precise, let $\hat{a}_{ij}$ be our estimator of a_{ij}, an arbitrary element in the matrix **A**. We adopt a conservative criterion, which requires

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon\right) \leq \delta,$$

where $\epsilon > 0$ is the tolerance in our estimates and $1 - \delta \in (0, 1)$ is our confidence. In other words, we would like the probability that our estimates deviate substantially from their true values to be low, no matter what the true $\mathbf{A}$ matrix is. Because of the maximization over all entries in the matrix, we require that every single entry meets this criterion. Hence, we refer to this as the *uniform ϵ-accuracy* criterion. This notion of error is known as "probably approximately correct" in the machine learning field, which also aims to learn accurately with high probability (see Valiant 1984).

Ideally we would like both ϵ and δ to be small so that we have accurate estimates with high probability. But in order to achieve smaller ϵ and δ, intuitively we would need to run more experiments to gather more data. Our first objective is to determine, for a given number of products n and fixed accuracy and confidence parameters ϵ and δ, how many experiments are needed to achieve those levels uniformly. This answer in turn tells us how the number of experiments needed scales with the number of products.

3.3.1 Interpretation and discussion

As has been described, uniform ϵ-accuracy is an intuitive measure of accuracy. It is also a conservative measure because it requires every entry of $\mathbf{A}$ to be estimated accurately. Alternatively, we can consider other criteria, such as bounding the root-mean-square error:

$$\mathbf{P}\left(\sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n}(\hat{a}_{ij} - a_{ij})^2 \geq \epsilon}\right) \leq \delta.$$

This is a relaxation of the uniform ϵ-accuracy criterion: if an estimator $\hat{a}_{ij}$ satisfies uniform ϵ-accuracy, then it also satisfies the RMSE criterion. Therefore, any positive results on the speed of learning under uniform ϵ-accuracy also hold under weaker criteria, such as the RMSE criterion. Our results then give a worst-case upper bound, in the sense that the number of experiments required to achieve a weaker criterion would be no more than the number of experiments required to achieve the stricter uniform ϵ-accuracy criterion.

A similar point can be made about the method used to design the experiments and estimate the parameters. Improvements on our random experimental design and our relatively simple comparisons of the treatment and control outcomes should lead to further improvements in the amount of information learned and therefore decrease the number of experiments required to achieve uniform ϵ-accuracy.

3.4 Asymptotic notation

In order to judge different learning models, we compare how many experiments are needed to achieve uniform ϵ-accuracy. Because our goal is to investigate the informational value of experiments and because we are interested in the regime where the number of products is large, we focus on how quickly the number of experiments needed increases as the number of products increases. To capture the scale of this relationship, we use standard asymptotic notation (see Appendix B for a detailed description).

3.5 Estimation of general A matrices

We first consider the problem of estimating general $\mathbf{A}$ matrices, without any assumptions of additional structure. Based on the technique outlined in Section 3.2, our precise estimation procedure is the following:

1. Perform independent experiments. For each experiment, use a random, independent decision vector $\mathbf{x}$, where for each product, x_j is distributed uniformly on $[-\rho, \rho]$, where $0 < \rho < 1$. Observe the resulting vector of changes in demand $\Delta\mathbf{q}$.

2. For the t^{th} experiment and for each a_{ij}, compute the statistic

$$y_{ij}(t) \triangleq \beta \cdot \Delta q_i(t) \cdot x_j(t),$$

where $\beta \triangleq 3/\rho^2$.

3. After s experiments, for each a_{ij} compute the sample mean

$$\hat{a}_{ij} = \frac{1}{s}\sum_{t=1}^{s} y_{ij}(t),$$

which is an unbiased estimator of a_{ij}.

The following theorem gives a bound on the accuracy of this estimation procedure after s experiments.

Theorem 1 (Estimation accuracy with sub-Gaussian noise for general $\mathbf{A}$ matrices). *Under Assumption 1, for any $n \times n$ matrix $\mathbf{A}$ and any $\epsilon \geq 0$,*

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon\right) \leq 2n^2 \exp\left\{-\frac{s\epsilon^2}{\max\limits_i 36\left(\sum\limits_{\ell=1}^{n} a_{i\ell}^2 + \dfrac{c^2}{\rho^2}\right)}\right\}. \qquad (3.1)$$

See Appendix C for the proof.

To ensure uniformly ϵ-accurate estimates with probability $1-\delta$, it suffices for the right-hand-side of (3.1) to be less than or equal to δ. Therefore, with a simple rearrangement of terms, we find that s experiments are sufficient if s satisfies

$$s \geq \frac{\max_i 36\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}{\epsilon^2} \log\left(\frac{2n^2}{\delta}\right).$$

The above bound tells us that if there is more noise (larger c) or if we desire more accurate estimates (smaller ϵ and δ), then more experiments may be required, which agrees with intuition. However, the term $\sum_{\ell=1}^{n} a_{i\ell}^2$ may be quite large and, as it is a sum of n quantities, may also scale proportionately with n. In that case, our estimation procedure may in fact require $O(n \log n)$ experiments in order to achieve uniform ϵ-accuracy, which can be prohibitively large.

3.6 Introducing structure

The previous result allows for the possibility that with general $\mathbf{A}$ matrices, many experiments may be required to estimate the underlying parameters. Fortunately, we recognize that our problem may have an important inherent structure that allows us to learn the $\mathbf{A}$ matrix much faster than we would otherwise expect.

We consider three different types of structure on the matrix $\mathbf{A}$. In the following sections, we motivate these assumptions, state the number of experiments needed to learn $\mathbf{A}$ in each case, and interpret our results.

3.6.1 Bounded pairwise effects

Motivation: Our first assumption is based on the idea that a product can affect the demand for itself or for any other product only by some bounded amount. In other words, varying the price of a product cannot cause the demand for itself or any other product to grow or diminish without limit. In terms of our model, we can state the assumption precisely as follows.

Assumption 2 (Bounded pairwise effects). *There exists a constant b such that for any n, any $n \times n$ matrix $\mathbf{A}$, and any pair (i, j), $|a_{ij}| \leq b$.*

This is our weakest assumption as we do not place any other restrictions on $\mathbf{A}$. In particular, we allow every product to have an effect on every other product. By not imposing any additional assumptions, we can use this variation of the model as a benchmark to which we can compare our two subsequent variations. Since all elements of $\mathbf{A}$ may be nonzero, we refer to this as the case of "dense" $\mathbf{A}$ matrices.

Result: With this additional assumption, we show that our estimation procedure as described in Section 3.5 can learn all elements of $\mathbf{A}$ to uniform ϵ-accuracy with $O(n \log n)$ experiments.

Corollary 1.1 (Sufficient condition for uniformly ϵ-accurate estimation of dense $\mathbf{A}$). *Under Assumptions 1 and 2, for any $n \times n$ matrix $\mathbf{A}$ and any $\epsilon \geq 0$,*

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon\right) \leq 2n^2 \exp\left\{-\frac{s\epsilon^2}{36\left(nb^2 + c^2/\rho^2\right)}\right\}.$$

Therefore, to ensure uniformly ϵ-accurate estimates with probability $1-\delta$, it suffices for the number of experiments to be $O(n \log n)$.

Discussion: This result gives an upper bound on the number of experiments needed to learn the entries of $\mathbf{A}$, in the sense that with the best estimation method, the asymptotic scaling of the number of experiments needed to achieve uniform ϵ-accuracy will be no worse than $O(n \log n)$. However, this upper bound is again not practical as it suggests that in the worst case, the number of experiments needed may scale linearly with the number of products. Because we would like to keep the number of experiments small, we hope to achieve a sublinear rate of growth with respect to the number of products. Fortunately, this is possible if the $\mathbf{A}$ matrix is "sparse", as we discuss in the next section.

3.6.2 Sparsity

Motivation: Although a category may include many items, not all items will have relationships with one another. For example, varying the price of a nighttime cold remedy may not affect the demand for a daytime cold remedy.

Under our model of demand and cross-product elasticities, a pair of items having no interaction means that the corresponding element in the $\mathbf{A}$ matrix is zero. If many pairs of items have no relationship, then our $\mathbf{A}$ matrix will have many zero elements, which is referred to as a "sparse" matrix. In terms of our model, we express the assumption of sparsity as follows.

Assumption 3 (Sparsity). *For any n, there exists an integer k such that for any $n \times n$ matrix $\mathbf{A}$ and any i, $|\{j : a_{ij} \neq 0\}| \leq k$.*

For each row of $\mathbf{A}$, we bound the number of entries that are nonzero to be no more than k. Interpreting this in terms of products, for each product, we assume that there are at most k products (including itself) that can affect its demand. Note that we do not assume any knowledge of how these nonzero entries are distributed within the matrix. This is important as it means we do not need to know *a priori* which products have a demand relationship with one another and which do not.

Result: As long as the underlying matrix $\mathbf{A}$ exhibits this sparse structure, we have the following result on the number of experiments needed to estimate $\mathbf{A}$ with uniform ϵ-accuracy using our estimation method.

Corollary 1.2 (Sufficient condition for uniformly ϵ-accurate estimation of sparse $\mathbf{A}$). *Under Assumptions 1, 2, and 3, for any $n \times n$ matrix $\mathbf{A}$ and any $\epsilon \geq 0$,*

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon\right) \leq 2n^2 \exp\left\{-\frac{s\epsilon^2}{36\,(kb^2 + c^2/\rho^2)}\right\}.$$

Therefore, to ensure uniformly ϵ-accurate estimates with probability $1-\delta$, it suffices for the number of experiments to be $O(k \log n)$.

Discussion: This result shows that if the $\mathbf{A}$ matrix is sparse, the number of experiments needed scales on the order of $O(k \log n)$, instead of $O(n \log n)$ as for the case of dense $\mathbf{A}$ matrices. Thus, the number of experiments needed grows logarithmically (hence, sublinearly) in the number of products, n, and linearly in the sparsity index, k. As long as k does not increase too quickly with n, this may be a significant improvement over $O(n \log n)$. As anticipated in the introduction, sparsity can yield much faster learning. The gap between a theoretical requirement of $O(k \log n)$ and a theoretical requirement of $O(n \log n)$ experiments could be dramatic for practical purposes in settings with a large number of products, and therefore in estimation problems with a large number of parameters. Of course this requires that k does not grow too quickly with n. We will investigate this possibility in Chapter 4.

By thinking about the amount of abstract "information" contained in a sparse matrix as opposed to in a dense matrix, we can gain some intuition as to why a sparse matrix is easier to estimate. When trying to learn a model, if we know that the true model lies in a restricted class of possible models, then we expect to be able to learn the true model faster than if no such restrictions were known. Our assumptions of sparsity effectively reduce the universe of possible $\mathbf{A}$ matrices in this manner. If $\mathbf{A}$ could be any $n \times n$ matrix, then for each row of $\mathbf{A}$, there would be on the order of n bits of unknown information (i.e., a constant number of bits for the value of each entry in the row). On the other hand, if we knew that the row has only k nonzero entries, there would instead be on the order of k bits of unknown information (i.e., a constant number of bits for the value of each *nonzero* entry in the row). There would also be uncertainty in the location of the nonzero entries. There are $\binom{n}{k}$ ways of choosing k entries out of n to be the nonzero ones, and therefore there are $\binom{n}{k}$ possible locations of the nonzero entries within the row, which can be encoded as an additional $\log_2 \binom{n}{k}$ bits of unknown information, which is approximately of order $O(k \log n)$ bits. Based on these rough calculations, we can see that knowing that a matrix is sparse with only k nonzero entries reduces the degrees of freedom and amount of uncertainty and therefore allows for faster estimation.

3.6.3 Bounded influence (weak sparsity)

Motivation: Assumptions 2 and 3 are both based on the intuition that the substitution and complementarity effects between products are bounded. This was done through placing hard bounds on the magnitude of *each* pairwise effect (i.e., the magnitude of each element of $\mathbf{A}$) and by limiting the number of possible relationships a product can have (i.e., the number of nonzero elements in each row of $\mathbf{A}$).

An alternative approach, in the same spirit, is instead to bound the aggregate effect on a product's demand due to all price variations. The intuition here is that although there may be many products, the demand for any individual product cannot be swayed too much, no matter how other products' prices are varied. This can be thought of as a "weak" sparsity assumption: we do not assume that many elements of $\mathbf{A}$ are zero; instead we assume that the overall sum across any row of $\mathbf{A}$ stays bounded. We express this assumption in terms of our model as follows.

Assumption 4 (Bounded influence). *For any n, there exists a constant d such that for any $n \times n$ $\mathbf{A}$ matrix, the following inequality is satisfied for every i:*

$$\sum_{j=1}^{n} |a_{ij}| \leq d.$$

As another interpretation, Assumption 3 can be thought of as bounding the ℓ_0 "norm" of the rows of $\mathbf{A}$: $\|\mathbf{a}_i\|_0 \leq k$. Assumption 4 above can be thought of as a relaxation that instead bounds the ℓ_1 norm of the rows of $\mathbf{A}$: $\|\mathbf{a}_i\|_1 \leq d$.

Result: Using similar analysis, we show that the number of experiments needed to achieve uniform ϵ-accurate estimation under the assumption of bounded influence is on the order of $O(d^2 \log n)$.

Corollary 1.3 (Sufficient condition for uniformly ϵ-accurate estimation under bounded influence). *Under Assumptions 1 and 4, for any $n \times n$ matrix $\mathbf{A}$ and any $\epsilon \geq 0$,*

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon\right) \leq 2n^2 \exp\left\{-\frac{s\epsilon^2}{36\,(d^2 + c^2/\rho^2)}\right\}.$$

Therefore, to ensure uniformly ϵ-accurate estimates with probability $1-\delta$, it suffices for the number of experiments to be $O(d^2 \log n)$.

Discussion: The above result shows that even with a weaker sparsity condition, where we allow all parameters to be nonzero, we are still able to achieve an order of growth that is logarithmic in the number of products. Note that if Assumptions 2 and 3 are satisfied with constants k and b, respectively, then Assumption 4 will also be satisfied with $d \triangleq kb$, and so the bounded influence assumption can subsume the combination of bounded pairwise effects and sparsity assumptions. However, using the more general bounded influence assumption to capture sparsity leads to a weaker result because it does not leverage all of the structural details of the sparsity assumption. Specifically, with $d = kb$, Corollary 1.3 would give a scaling of $O(k^2 \log n)$ for learning a k-sparse $\mathbf{A}$ matrix (where the dependence on b has been suppressed), which is slower than the scaling of $O(k \log n)$ given by invoking Corollary 1.2.

As was the case under the sparsity assumption, the nature of the logarithmic scaling $O(d^2 \log n)$ under bounded influence depends on how quickly d changes with n. We will investigate this relationship in Chapter 4.

3.7 Standard errors and confidence intervals

Besides providing a result on the speed of learning, Theorem 1 also allows us to construct confidence intervals for the elasticity estimates by rearranging (3.1). Specifically, for

$$\epsilon = \sqrt{\frac{\max_i 36(\sum_{\ell=1}^n a_{i\ell}^2 + c^2/\rho^2)}{s} \log\left(\frac{2n^2}{\delta}\right)},$$

we have that $\mathbf{P}\left(|\hat{a}_{ij} - a_{ij}| \leq \epsilon\right) \geq 1 - \delta$. Under each structural assumption, we can also replace the (unknown) sum $\sum_{\ell=1}^n a_{i\ell}$ with the appropriate bound.

Although this confidence interval has an analytical form given by our theory, it will be loose because we have used upper bounds of quantities in the derivation of (3.1). It also depends on parameters that we do not know, namely the a_{ij}'s and c. An alternative is to use the jackknife or bootstrap to estimate standard errors and use these to construct confidence intervals. For each experiment t we obtain a measurement $y_{ij}(t)$ for a particular unknown elasticity parameter a_{ij}, and our estimator $\hat{a}_{ij}$ is the sample mean of these y_{ij}'s. Therefore, to estimate the standard error of our estimator after s experiments, we can resample from our s measurements of y_{ij}'s and calculate the sample mean of this resample. By resampling many times, we obtain a distribution of sample means, from which we can estimate the standard deviation of our sample mean estimator.

3.8 Lower bound

The previous results provide upper bounds on the number of experiments needed for accurate estimates. For example, in the case of sparsity, using our estimation method, no more than $O(k \log n)$ experiments are needed to achieve uniform ϵ-accuracy. However, these results do not tell us whether or not there exists another estimation method that requires even fewer experiments.

Given our demand model, the bounds on the allowable price variations, and the noise in the data, information theory tells us the maximum amount of information about the a_{ij}'s that can be learned from a single experiment. This fundamental limit in the "value" of each experiment in estimating the $\mathbf{A}$ matrix then allows us to calculate a lower bound on the number of experiments required. We do not actually need to develop a specific estimator that achieves this lower bound, but we know that no estimator can do better than this lower bound.

For the special case of i.i.d. Gaussian noise, we now present such a lower bound on the number of experiments needed, which shows that no matter what estimation procedure we use, there is a minimum number of experiments needed to achieve uniform ϵ-accuracy. The only requirement we impose on the estimation procedure is that it relies on experiments with bounded percentage price changes. The bounds we impose on the percentage price changes can be justified by practical considerations: the natural lower bound on price changes comes from the fact that prices cannot be negative, while the upper bound on the percentage changes captures the fact that the manager of a store is likely to be opposed to dramatic price increases for the purposes of experimentation.

Theorem 2 (Necessary condition for uniform ϵ-accurate estimation under sparsity with Gaussian noise). *For $\lambda > 0$, let*

$$\mathcal{A}_{n,k}(\lambda) \triangleq \left\{ \mathbf{A} \in \mathbb{R}^{n \times n} : |\{j : a_{ij} \neq 0\}| = k, \forall i = 1, \ldots, n; \min_{i,j: a_{ij} \neq 0} |a_{ij}| \geq \lambda \right\}$$

be the class of $n \times n$ $\mathbf{A}$ matrices whose rows are k-sparse and whose nonzero entries are at least λ in magnitude. Let the noise terms be i.i.d. $\mathcal{N}(0, c^2)$ for some $c > 0$. Suppose that for some $\epsilon \in (0, \lambda/2)$ and $\delta \in (0, 1/2)$, we have an estimator that

(a) experiments with percentage price changes $x \in [-1, 1]$ (i.e., the price of each product cannot fall below 0 and cannot increase by more than 100%), and

(b) for any $\mathbf{A}$ matrix in $\mathcal{A}_{n,k}(\lambda)$ achieves uniformly ϵ-accurate estimates with probability $1 - \delta$.

Then, the number of experiments used by the estimator must be at least

$$s \geq \frac{k \log(n/k) - 2}{\log(1 + k^2\lambda^2/c^2)}.$$

The proof is given in Appendix D.

As the number of products grows, the asymptotically dominant scaling terms are

$$s \geq \Omega\left(\frac{k \log(n/k)}{\log k}\right).$$

Since $\log k$ is small compared to k and $\log n$, we have an essentially matching lower bound to the $O(k \log n)$ upper bound given in Corollary 1.2, which shows that our estimation procedure achieves close to the best possible asymptotic performance.

3.9 Discussion

The previous results demonstrate the power of sparsity in multiple flavors. Without any assumptions on the structure of the problem, the number of experiments needed may grow linearly with the number of products. For our target regime of large numbers of products, this leads to a solution that appears to be practically infeasible. However, by recognizing the inherent properties of the problem, we show that even with randomly designed experiments we are able to learn $\mathbf{A}$ using a number of experiments that scales only logarithmically with the number of products. With a large number of products, the difference between linear and logarithmic growth is tremendous: e.g., for $n = 100$, $\log(100) \approx 4.6$. This gives hope that we can indeed learn the $\mathbf{A}$ matrix in a practically feasible number of experiments. In Chapter 5, we present simulations that support these results.

While our findings help reveal how many experiments are required, it is also helpful to ask how many experiments are feasible. When firms are using field experiments to set policy (rather than academics using them to test theories), we have found that they are often willing to run a rather large number of experiments.

The answer will clearly depend upon the nature of the firm's actions and the particular setting. Varying advertising or pricing decisions in online or direct marketing settings can often be implemented at low cost, making it feasible to implement hundreds or even thousands of experiments. For example, Capital One reportedly implements tens of thousands of randomized field experiments each year.

In traditional retail settings, the cost of making in-store changes is generally higher, and randomization must often occur at the store level rather than at the individual customer level (introducing an additional source of measurement error). However, even in this setting, firms with multiple locations can implement a large number of experiments in different samples of stores to test pricing, product placement, and other merchandising decisions. For example, a large brick-and-mortar retailer was quickly able to run 200 between-store pricing experiments to decide how to price private label items when national brands are promoted. Documented examples of high-volume experimentation in traditional retail settings include Bank of America varying actions between bank branches and Harrah's varying a wide range of practices across its casinos.

In other settings, implementing field experiments is more challenging. For example, when deciding how to manage a distribution network, a firm may be limited to only a handful of experiments every few years, as these experiments will tend to disrupt existing relationships and require extended periods to observe the outcome.

Chapter 4

Estimating the sparsity parameters

In Chapter 3, we showed that under sparsity assumptions, the number of experiments needed to estimate the $\mathbf{A}$ matrix scales as $O(k \log n)$ or $O(d^2 \log n)$. However, the sparsity parameters k and d are not known to the researchers or the store managers and must also be estimated from data. In addition, the rate at which these parameters grow with n will also impact the nature of the scalings $O(k \log n)$ and $O(d^2 \log n)$. If k and d grow quickly with n, then the $O(k \log n)$ and $O(d^2 \log n)$ growth rates will again mean that it may be infeasible to use experiments to set prices in large categories. In this chapter, we present a methodology for estimating these sparsity parameters from data and subsequently apply the methodology to historical sales data to provide evidence that k and d grow sublinearly with n.

4.1 A model selection approach

There are two potential ways of obtaining data to estimate these sparsity parameters: (1) from a "pilot" set of experiments and (2) from historical data. Using either source, we use what is essentially a model selection approach. We divide the data into calibration and validation sub-samples. We then repeatedly estimate the $\mathbf{A}$ matrix using the calibration sub-sample for different values of the sparsity parameter, and we choose the value of the sparsity parameter for which the estimated $\mathbf{A}$ matrix has the best fit with the validation sub-sample.

Different variants of this general approach are available, including different measures of "goodness-of-fit" of the validation sub-sample. We can also use different approaches to cross-validate, including m-fold cross validation where we randomly split the data into m buckets and rotate which of the buckets we treat as the validation sample. In the discussion below, we describe our methodology more formally and present results of both simulations and empirical analyses to illustrate its performance.

4.2 Methodology

Let $\mathbf{a}_i$ be the (unknown) $1 \times n$ row vector of elasticities for the i^{th} product. Suppose we have s data points from either s experiments or s periods of historical data: $\Delta\mathbf{q}_i$ is a $1 \times s$ vector of changes in demand for the i^{th} product, and $\mathbf{X}$ is an $n \times s$ matrix of pricing decisions. For some value τ, we solve the following optimization problem (the "Lasso"; see Tibshirani 1996), which looks for the $\mathbf{a}_i$ that best fits the data but is still constrained to be "sparse":

$$\begin{aligned} \min_{\mathbf{a}_i} \quad & \|\Delta\mathbf{q}_i - \mathbf{a}_i^T\mathbf{X}\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{a}_i\|_1 \le \tau. \end{aligned}$$

Alternatively, we can express the problem in the following form:

$$\min_{\mathbf{a}_i} \; \|\Delta\mathbf{q}_i - \mathbf{a}_i^T\mathbf{X}\|_2^2 + \lambda\|\mathbf{a}_i\|_1. \tag{4.1}$$

Here, τ and λ are tuning parameters that control the level of sparsity of the resulting solution. For each choice of the tuning parameters, we obtain one solution, $\hat{\mathbf{a}}_i$, to the optimization problem. To assess the quality of each solution, we cross-validate it using the given data and select the one that gives the lowest cross-validation error as the best solution. From this best solution, we recover its sparsity parameters and propose these measures as estimates of the true sparsity parameters. As we obtain additional data, we can repeat this procedure to update our estimates of the sparsity parameters.

Although this methodology focuses on a single product/row i, the same procedure can be performed on each row independently, using the same set of data, to obtain estimates of k and d for each row. Our model calls for parameters that uniformly bound the sparsity of the entire matrix. Therefore, to arrive at estimates of the overall sparsity parameters for the entire matrix, we take the maximum over the estimates of each individual row's sparsity parameters. Note that this approach is valid for either hard sparsity (k) or bounded influence (d). We will test the methodology on both cases.

4.3 Pilot experiments

In order to perform the procedure described in the previous section, we first require some data. One possible source of data is a set of "pilot" experiments: a relatively small sequence of pricing experiments and corresponding observed demand quantities.

4.3.1 Simulation

In practice, managers can conduct actual pilot experiments and collect the necessary data. In this subsection, we simulate pilot experiments by generating synthetic experimental data. To ensure that our simulations use realistic parameters, we initialize them using data from a large-scale pricing experiment that was conducted for another purpose (Anderson et al. 2010). The experiment was implemented at a large chain of stores that sells products in the grocery, health and beauty, and general merchandise categories. Eighteen of the chain's stores participated in the study, in which prices were experimentally manipulated on 192 products for seventeen weeks, with the treatments randomly rotated across the eighteen stores (see Anderson et al. 2010 for additional details). From this study, we obtained distributions for the diagonal and off-diagonal entries of the $\mathbf{A}$ matrix.

Our simulation proceeds as follows:

1. Choose fixed values of n and d (or k) and generate the true $\mathbf{A}$ matrix randomly

from the seed distributions. Choose a fixed value of c, the standard deviation of the normal error term $\mathbf{w}$. These parameters are not used by the estimation algorithm.

2. For any given s:

    (a) Randomly generate $\mathbf{x}$ and $\mathbf{w}$ for s experiments, and calculate $\Delta\mathbf{q}$.

    (b) For a range of λ's, find the optimal solutions to (4.1).

    (c) Perform five-fold cross-validation[1] on the solutions to identify the one with the lowest cross-validation error; call this $\hat{\mathbf{a}}_i^*$. (Figure 4-1 illustrates the cross-validation process.)

    (d) Calculate $\|\hat{\mathbf{a}}_i^*\|_1$ and $\|\hat{\mathbf{a}}_i^*\|_0$. For the latter, we count only those entries that are above a certain threshold (set at 0.01) in magnitude.

    (e) For each s, replicate this 10 times and average the results. Propose the averaged values of $\|\hat{\mathbf{a}}_i^*\|_1$ and $\|\hat{\mathbf{a}}_i^*\|_0$ as estimates of d and k, respectively.

3. Plot the estimates of d and k versus a range of values of s, giving a sense of how many experiments are needed to obtain an accurate estimate of the level of sparsity.

As Figure 4-2 illustrates, our methodology provides reasonable estimates of k and d with relatively few experiments, and these results hold for different values of the true underlying sparsity parameters. These results suggest that using pilot experiments can indeed provide initial estimates of k and d. Knowing these sparsity parameters, we then have a sense of the feasibility of using our main methodology to estimate $\mathbf{A}$. In addition to providing estimates of the sparsity parameters, the data generated in these pilot experiments can also serve as additional data that can be used to estimate $\mathbf{A}$ using our main methodology. Furthermore, if the pilot experiments involve variation in

[1] Split the data set into five buckets. Estimate $\mathbf{a}_i$ on data from four buckets and cross-validate on the fifth. Rotate and do this for all five buckets and calculate the average error.

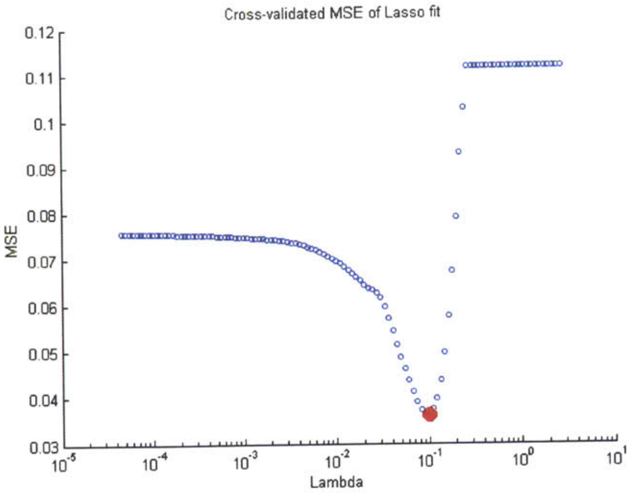


Figure 4-1: An example of the result of five-fold cross-validation. The value of λ highlighted in red gives the lowest cross-validation error. Large values of λ (to the right) heavily penalize nonzero entries, resulting in the zero vector as the solution, which does not fit the data well. As λ is lowered, we begin to get some nonzero entries in the solution, which provides a better fit of the data. However, as λ becomes even smaller, past the value marked in red, we obtain dense solutions that tend to overfit, resulting in a higher cross-validation error.

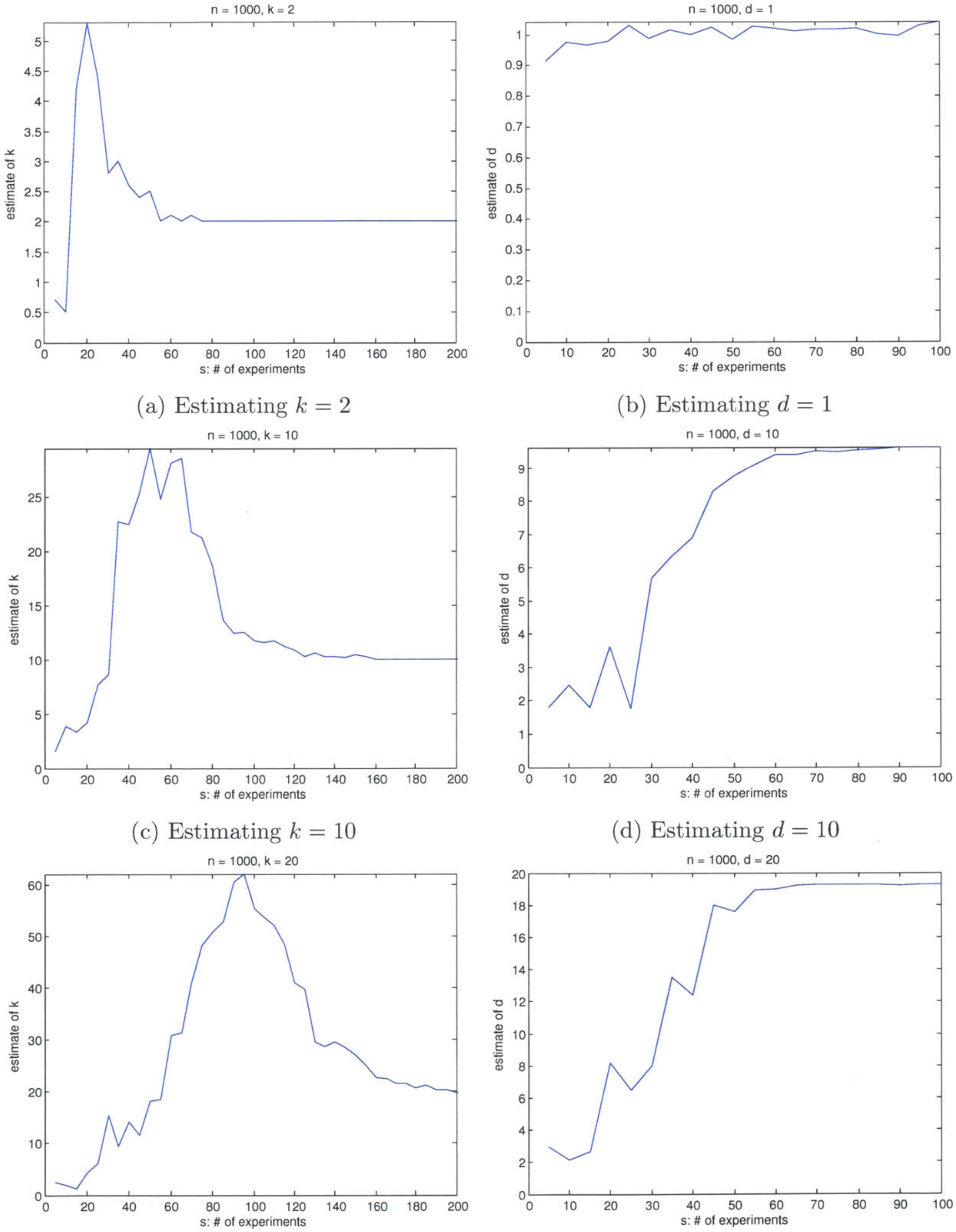
n (e.g., by experimenting on multiple stores with different category sizes), we can also investigate how the sparsity parameters grow with n.

4.4 Empirical analysis

Running 80 to 100 pilot experiments is not without cost, and so ideally a firm would like to be able to estimate k and d using its existing data. One possibility is to use historical variations in prices to estimate these parameters. Our proposed cross-validation method can be easily adapted to do so.

One limitation of using historical data is that variations in prices are often not as frequent or as large as they would be for field experiments. However, historical data often covers a long time period, and the large quantity of data may still give us accurate estimates.

Another limitation of using historical variation in control variables is that this past



(a) Estimating $k = 2$

(b) Estimating $d = 1$

(c) Estimating $k = 10$

(d) Estimating $d = 10$

(e) Estimating $k = 20$

(f) Estimating $d = 20$

Figure 4-2: Plot of the estimates of k and d versus the number of experiments, s. The estimates are near the true values of k and d, even with relatively few experiments.

variation is often not random. This has raised concerns that the resulting elasticity estimates may be biased (Villas-Boas and Winer 1999). Although these limitations are less relevant in this setting, where we seek only a preliminary estimate of k and d, we will investigate in detail the possibility of endogeneity in Chapter 6.

We use 195 weeks of historical data from a chain of 102 convenience stores. The number of products sold in each store varies, due primarily to differences in the square footage size of each store (larger stores offer wider product ranges). We will exploit this variation to illustrate how our estimates of k and d vary with the number of items in the category, n.

4.4.1 Setup

We begin with the 195 weeks of sales data from 102 stores, which we then group into 48 four-week *periods* in order to reduce the amount of noise in the data. We focus on a specific category ("Cold remedies") and perform the following procedure independently for each store:

1. We first fill in any missing data:

   (a) If a product is not sold in a given period, no data is available for that product during that period, which means that we do not know the retail price for that product during that period. We fill in this price data by linearly interpolating between the prices for that product during the two most adjacent periods for which we do have data.

   (b) However, we do know that if no data is available, the quantity sold during that period is zero, which we also fill in.

   (c) After this processing, we have a complete set of sales and price data for each product, for each of the 48 four-week periods.

2. For each product i, we compute the average quantity sold per period and the average price over the 48 periods. These will serve as the baseline demand (q_i^b)

and price levels (x_i^b), respectively.

3. To further reduce noise, we consider only those products that (i) sold over a certain threshold of units per period on average, (ii) sold at least one unit during the first four periods and last four periods (to ensure they were not introduced or discontinued during the middle of the 195 weeks), and (iii) had variations in prices above a certain threshold over the course of the 48 periods.

4. We collect all products that do not pass through the above filter and combine them into a single aggregate "product", which is included together with all other products in the analysis that follows.

5. We calculate category-level seasonality factors for each period, which are used to deseasonalize the raw demand quantities.

6. Using the price data and the (deseasonalized) sales data for each period, we then calculate their percentage change from the previously established baseline levels, which gives us $\Delta\mathbf{q}$ and $\mathbf{x}$.

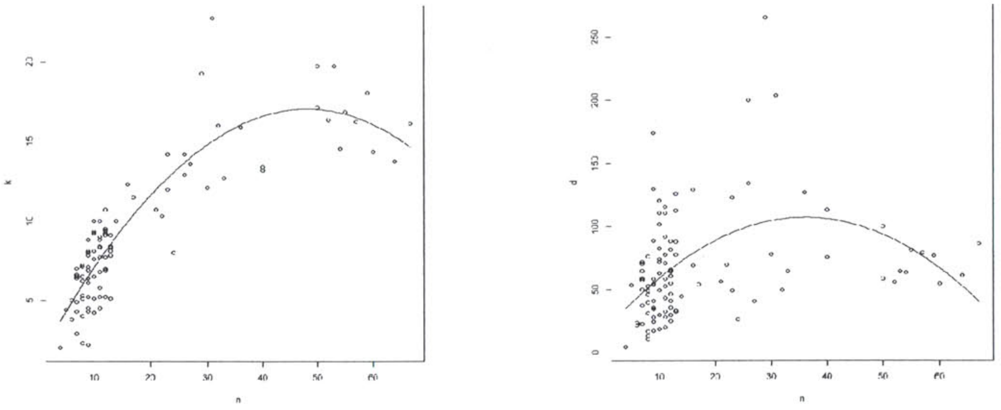
7. Equipped with $\Delta\mathbf{q}$ and $\mathbf{x}$, we then use these as input to the Lasso optimization program (4.1):

   (a) Lasso estimates vectors, so we estimate $\mathbf{A}$ row-by-row.

   (b) For each row i, we try a sequence of λ parameters and perform five-fold cross-validation in order to identify the value of λ that gives the lowest cross-validation error; call this estimate $\hat{\mathbf{a}}_i^*$. Calculate $\|\hat{\mathbf{a}}_i^*\|_1$ and $\|\hat{\mathbf{a}}_i^*\|_0$ as estimates of k_i and d_i, respectively, for row i.

   (c) Because k and d are sparsity parameters for the entire $\mathbf{A}$ matrix, we take the maximum over all of the rows' k_i's and d_i's to obtain the overall estimate of k and d.

   (d) For robustness, we repeat this entire procedure ten times and average the results.



(a) Estimating k

(b) Estimating d

Figure 4-3: Plots of n versus estimated k and d, including the quadratic fit. Sales threshold: one unit per period on average; standard deviation of price variations threshold: 0.08.

8. By performing this analysis for each store, we obtain a collection of pairs of (n, k) and (n, d) data points, which give us a relationship between the number of products and the sparsity parameters.

9. For both sets of data points, we fit a quadratic model and verify that the second-order coefficient is negative and significant, indicating that the sparsity parameters do not increase linearly with the number of products.

4.4.2 Results

Figure 4-3 presents the estimates of k and d across all of the stores (each point represents the estimates for a single store). Recall that the number of items in each category varies across the stores, which allows us to investigate the relationship between the sparsity parameters and n. The plots also show the fitted quadratic relationships between the data points, which allow us to evaluate whether the growth in the sparsity parameters is slower than linear. In Table 4.1, we report the results of these quadratic fit models.

The estimates of k reveal a relatively distinct pattern: the estimates grow with n

	Coefficient	Estimate	Std. Error	t value
Estimating k	(Intercept)	1.118	0.607	1.843
	1st-order term	0.661	0.059	11.169
	2nd-order term	-0.007	0.001	-7.461
Estimating d	(Intercept)	15.280	11.696	1.306
	1st-order term	5.068	1.142	4.438
	2nd-order term	-0.070	0.018	-3.959

Table 4.1: Summary of quadratic fit models for four-week periods with a sales threshold of one unit sold per period on average and a minimum standard deviation of price variations of 0.08. The second-order coefficients are negative and significant for both k and d.

but the growth rate is slower than linear. In the fitted quadratic equation, the quadratic term is negative and highly significant. We can speculate on the reasons for this. It is possible that customers eliminate products from their consideration sets that do not exhibit certain attributes. For example, on a specific trip, customers may focus on only nighttime cold remedies *or* daytime cold remedies. If this is the case, then introducing a new daytime product may not increase k (which is an upper bound on the number of interdependent products) because it affects demand for only the subset of items that share that attribute (i.e., daytime remedies). It was this type of behavior that Tversky (1972) anticipated when proposing that customers eliminate alternatives by aspects.

The estimates of k are relatively small (around 15) even in large categories. This suggests that in the "Cold remedies" category, the matrix of cross-price elasticities is sufficiently sparse to make estimation using field experiments feasible. This analysis also demonstrates the feasibility of using historical data to obtain initial estimates of k to evaluate when a firm can use experiments to set prices. The data that we have used is readily available to most retailers. Notably, because we obtain estimates of the sparsity parameters for each category in each store, it does not require that retailers have a large number of stores (although having many stores obviously makes experimentation easier).

Notice that for many of the stores we observe only approximately 10 items in the "Cold remedies" category. This reflects both the relatively small size of these stores as well as the screening of products based on their sales volumes and the level of price variation. To evaluate the robustness of our findings, we repeated the analysis for different minimum sales and price variation thresholds. We also replicated the findings when grouping the data into ten-week periods. In all of these variations, the results follow the pattern reported in Figure 4-3 and Table 4.1, with the quadratic coefficient from regressing k on n being negative and highly significant.

We also report estimates of d. The fitted quadratic function indicates that the growth of d with n is also sublinear.[2] However, the findings reveal a much less distinct pattern compared to the results for k. Notably, some of the estimates of d are very large (exceeding 100). Moreover, while our estimates of k are relatively robust, the estimates of d are much less robust and are sensitive to variation in the filtering parameters. One interpretation is that within the "Cold remedies" category, the weak sparsity structure is not sufficient to make it feasible to use experiments to set prices. A second interpretation is that our estimation procedure is not accurate enough to provide reliable estimates of d.

As these results show, historical data can provide useful information about the demand system – we obtain estimates of not only the sparsity parameters but also the underlying $\mathbf{A}$ matrix as well. Given that historical data are often readily available, one might forego running experiments entirely and rely solely on historical data. However, field experiments offer at least two advantages over historical data. First, historical data often suffer from endogeneity, which could lead to biased estimates. We consider this issue and propose a way to account for endogeneity in Chapter 6, but by using randomized field experiments, we ensure that our decisions are exogenous and, more generally, that we have a causal relationship between changes in prices and changes in demand. Second, many relevant conditions can change over the time horizon of the

[2] In the case of d, sublinear growth could simply reflect customer loyalty or state dependence (see, for example, Dubé et al. 2008, Erdem 1996, Keane 1997, Seetharaman et al. 1999, Anderson and Simester 2013). If even just a subset of customers is loyal to an existing product (or brand), then the introduction of additional products will have a bounded impact on sales of the existing products. The more customers who are loyal, the less growth we expect in d as n grows.

historical data. For example, some products may be discontinued and other products may be newly introduced in the middle of the time frame of the data, or macro-level shifts in consumer preferences may occur. These dynamics make it difficult not only to produce estimates using the historical data, but also to apply these estimates to the present since current conditions can be quite different from the conditions on which the estimates are based. By using field experiments, and especially when we require only a small number of them, we can quickly obtain estimates based on a static setting and immediately use the results when they are still highly applicable.

4.5 Summary

In this chapter, we described how to estimate the sparsity parameters k and d either from a pilot set of experiments or from historical data. Through simulations, we demonstrated that our estimation procedure can in fact accurately recover the true values of k and d using relatively few experiments. Using a sample of historical sales data, we also obtained actual estimates of k and d for the "Cold remedies" category. These estimates revealed that the sparsity parameters increase with n but that the growth is sublinear. Changing the price of an item within the "Cold remedies" category appears to affect the demand for no more than fifteen other items, suggesting that the $\mathbf{A}$ matrix of elasticities is in fact sparse. These findings illustrate a practical method that managers can use to evaluate whether a product category has a favorable structure to make it feasible to estimate elasticities using field experiments.

Chapter 5

Estimation performance

The theoretical results presented in Chapter 3 focused on the speed of learning, namely how many experiments are required to learn the $\mathbf{A}$ matrix accurately as the number of products grows. We provided theoretical bounds on the asymptotic growth rate. However, these bounds also depended on sparsity assumptions and the sparsity parameters, k and d. In Chapter 4, we presented a method for estimating these sparsity parameters and used historical sales data to produce actual estimates of k and d as a function of the number of products, n. With these estimates, we have a complete picture of the scaling between the number of experiments and the number of products. In this chapter, we investigate the performance of our estimation method and verify that it achieves the theoretical scaling that we derived.

5.1 Simulation setup

Because we are not able to perform actual field experiments, we will validate our methodology using simulations. To ensure that our simulations use realistic parameters, we initialize them using data from a large-scale pricing experiment that was conducted for another purpose (Anderson et al. 2010). This is the same setup that we used for the simulations in Section 4.3.1.

We also specify a collection of parameters that define the simulation: the number of products (n), structural parameters for the $\mathbf{A}$ matrix (b, k, and d), the noise distribution parameter (c), and the error criteria (ϵ and δ). We refer to these parameters together as the simulation definition. In order to compare the dense and sparse cases, we first generate a full matrix for the dense case using the two distributions for the diagonal and off-diagonal entries. We then randomly set all but k entries in each row to zero for the associated sparse case. Instead of selecting an arbitrary value for k, we use the empirical results from Section 4.4.2: for any given n, we use the quadratic fit (plus some additive noise) to calculate the associated value of k.

5.2 Simulation procedure

Given an $n \times n$ matrix $\mathbf{A}$ generated using the distributions described above, along with a definition of parameters, we can then use the procedure described in Section 3.5 to estimate $\mathbf{A}$.

To simulate one experiment, we generate a random vector $\mathbf{x}$ and random noise variables w_i. Using the true underlying $\mathbf{A}$ matrix, we then calculate the vector of percentage changes in demand $\Delta\mathbf{q} = \mathbf{A}\mathbf{x} + \mathbf{w}$ and the statistics y_{ij}, which are unbiased estimators of the a_{ij}'s. As we perform more experiments, we keep a running sum of the y_{ij}'s and compute the sample mean to obtain our estimate $\hat{a}_{ij}$. By comparing these estimates to the true $\mathbf{A}$ matrix, we can calculate the maximum absolute error across all entries: $\max_{i,j} |\hat{a}_{ij} - a_{ij}|$.

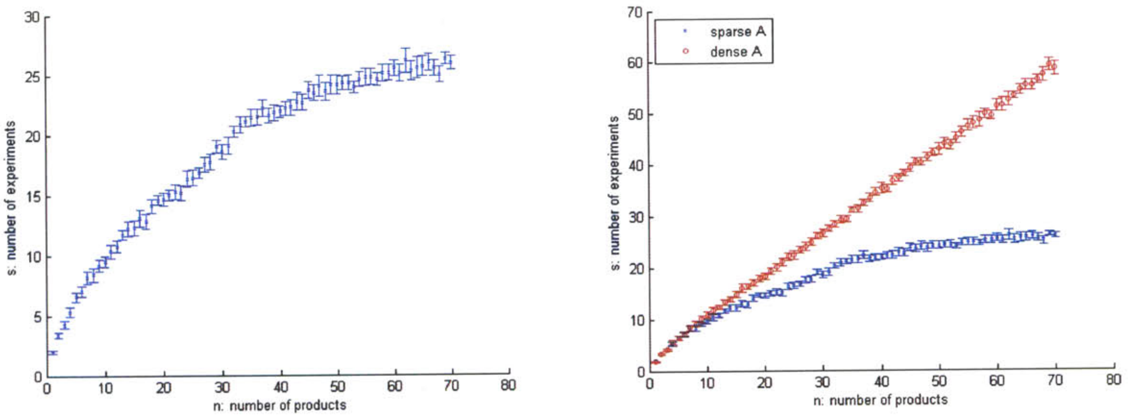
Since our criterion of uniform ϵ-accuracy requires the probability that the maximum absolute error is less than ϵ to be at least $1 - \delta$, we run 100 parallel sequences of experiments. Each sequence is essentially an independent instance of the estimation procedure. We incrementally generate more experiments for each sequence, compute updated estimates, and calculate maximum absolute errors. After any number of experiments, each sequence therefore has its own set of estimates and corresponding maximum absolute error. We say that we have achieved uniform ϵ-accuracy when at least a $1 - \delta$ fraction of the sequences have maximum absolute errors that are less than or equal to ϵ.

5.3 Simulation results

Using the preceding procedure, we can simulate the number of experiments needed to achieve uniform ϵ-accuracy for any given simulation definition. Because we are interested in how the number of experiments needed scales with the number of products, we fix a particular definition of parameters (except for n) and generate a sequence of matrices $\{\mathbf{A}_n\}$ that increase in size. For each matrix $\mathbf{A}_n$, we determine the number of experiments needed to achieve uniform ϵ-accuracy. For robustness, we replicate the entire simulation 20 times and, for each n, calculate 95% confidence intervals for the number of experiments needed.

In the case of sparse matrices, the resulting plot (Figure 5-1a) exhibits the logarithmic scaling predicted by our theoretical results. As the number of products grows, the number of experiments required grows much more slowly than the linear benchmark. Additional products require fewer and fewer additional experiments to achieve accurate estimates. On the other hand, Figure 5-1b shows that the dense case requires many more experiments than the sparse case in order to achieve the same level of estimation accuracy.[1]

[1] The results for the sparse case in Figure 5-1b are identical to the results in Figure 5-1a (the only difference is the change in the scale of the y-axis).



(a) Sparse **A** matrix

(b) Comparison of sparse and dense **A** matrices

Figure 5-1: When the **A** matrix is sparse, the number of experiments needed to achieve uniform ϵ-accuracy grows only logarithmically with the number of products. When the **A** matrix is dense, the number of experiments needed to achieve uniform ϵ-accuracy grows at least linearly with the number of products. Comparing the cases of sparse and dense **A** shows that learning is much faster in the sparse case. The bars represent 95% confidence intervals. Parameters used for this plot: $\rho = 0.5$, $c = 0.5$, $b = 5$, $\epsilon = 1.5$, $\delta = 0.1$.

Chapter 6

Addressing endogeneity in data

In Chapter 4, we described a methodology for estimating the sparsity parameters k and d using either a "pilot" set of experiments or historical data. Using pilot experiments has the advantage that prices are chosen exogenously. However, running experiments is not without cost, and so ideally a firm would like to be able to estimate k using its existing data.

A limitation of using historical variations in control variables is that this past variation is often not random. As opposed to the case of pilot experiments, with historical data, prices may be endogenous variables, which may lead to biased estimates of elasticities. For example, Villas-Boas and Winer (1999) examined the effect of endogeneity in the context of brand choice models and found significant differences in parameter estimates with and without accounting for endogeneity. In our setting, where we use historical data to estimate the sparsity parameters, biases in the elasticity estimates may carry over and bias our estimates of d. These biases may also affect whether elasticity estimates are 0 or not, which can therefore bias our estimates of k as well. Hence, we will study the robustness of our proposed methodology to endogeneity, using both simulations and actual historical data.

6.1 Modeling endogeneity

While endogeneity can manifest itself in several forms (e.g., omitted variables, measurement error, or simultaneity), the core issue is a correlation between an independent variable and the error term. In the context of our model,

$$\Delta \mathbf{q} = \mathbf{A}\mathbf{x} + \mathbf{w},$$

this amounts to the prices changes, $\mathbf{x}$, being correlated with the noise term, $\mathbf{w}$.

To see the potential for bias in the estimates, suppose that $\mathbf{x}$ and $\mathbf{w}$ are positively correlated. This would mean that increases in price for some product i (i.e., positive x_i) tend to be accompanied by simultaneous positive demand shocks for that product (i.e., positive w_i). Consider for example an unexpected rise in popularity of a certain toy accompanied by retailers' increasing the toy's price to capitalize on the fad. Assuming that the product's own-price elasticity is negative (i.e., $a_{ii} < 0$), this would lead to a positive price change x_i being associated with a demand change that is less negative than what the true elasticity would dictate. Hence, the resulting estimate of the elasticity would be positively biased (i.e., biased towards 0): one would be led to believe that the product is less price elastic than it actually is. On the other hand, if $\mathbf{x}$ and $\mathbf{w}$ are negatively correlated, the resulting elasticity estimate would be negatively biased.

One way of incorporating endogeneity into a model is to specifiy a random shock for prices and a random shock for demand and make these two shocks correlated. This is the approach taken by Villas-Boas and Winer (1999) and is also the approach we will follow. Recall that in Chapter 3, our estimation methodology chooses the price changes, $\mathbf{x}$, independently at random, and hence $\mathbf{x}$ and $\mathbf{w}$ are uncorrelated. In this chapter, we will instead specify a nonzero correlation between the price changes, $\mathbf{x}$, and noise term, $\mathbf{w}$.

Specifically, we assume that each pricing decision is composed of one part that is exogenous and i.i.d. and another part that is endogenous and correlated with the noise term. Let the pricing decision for product i be $x_i = \eta + \nu_i$, where $\nu_i \sim N(0, \sigma_\nu^2)$ and is i.i.d. across products and experiments. Let η and w_i, the noise term for product i's demand, follow a multivariate normal distribution with mean 0 and covariance matrix

$$\Sigma_{\eta w} = \begin{bmatrix} \sigma_\eta^2 & \rho\sigma_\eta\sigma_w \\ \rho\sigma_\eta\sigma_w & \sigma_w^2 \end{bmatrix},$$

where $\rho \in [-1, 1]$ captures the level of endogeneity. Under this model, the price variations of all products are correlated with the noise term and hence endogenous.

As in Chapter 4, we will use both simulations and historical data to investigate the effect of endogeneity on our estimation methodology.

6.2 Simulation

We first generate synthetic experimental data and test our methodology using simulations. Since our estimation methodology operates row-by-row, we will focus on the first row of $\mathbf{A}$ in this analysis. The simulation proceeds as follows:

1. Choose fixed values of n and k (or d) and generate the true row $\mathbf{a}_1^T$ randomly from the seed distributions.

2. Choose fixed values of σ_η, σ_ν, σ_w, and ρ.

3. For each simulated experiment, draw η and w_1 jointly from a multivariate normal distribution with mean 0 and covariance matrix $\Sigma_{\eta w}$. Generate the exogenous price variations, ν_i's, independently from a normal distribution with mean 0 and variance σ_ν^2. Combine η and the vector of ν_i's to obtain the vector of price changes, $\mathbf{x}$, for that experiment. Generate s independent samples of $\mathbf{x}$'s and w_1's in this manner to simulate s experiments. Collect these into matrix $\mathbf{X}$ and vector $\mathbf{w}_1$. Calculate $\Delta\mathbf{q}_1 = \mathbf{a}_1^T\mathbf{X} + \mathbf{w}_1$.

   Simultaneously, also generate an alternate sequence of s demand shocks from an independent $N(0, \sigma_w^2)$ distribution and collect into a vector $\tilde{\mathbf{w}}_1$. Calculate $\Delta\tilde{\mathbf{q}}_1 = \mathbf{a}_1^T\mathbf{X} + \tilde{\mathbf{w}}_1$ using the same matrix of price variations as above, but substituting this uncorrelated vector of demand shocks. We use these two models, one with endogenous prices and the other with exogenous prices, to compare the effect of endogeneity on our estimates.

4. For both models, using $\mathbf{X}$ with both $\Delta\mathbf{q}_1$ and $\Delta\tilde{\mathbf{q}}_1$, estimate the sparsity parameters:
   (a) Calculate the "Lasso" estimate of k (or d) using the same procedure as described in Section 4.3.1, including cross-validation.
   (b) Calculate also the OLS estimates as

$$\hat{\mathbf{a}}_1^{OLS} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}(\Delta\mathbf{q}_1).$$

   Compute the OLS estimate of k as $\|\hat{\mathbf{a}}_1^{OLS}\|_0$ (where we again count only those entries that are above 0.01 in magnitude) and of d as $\|\hat{\mathbf{a}}_1^{OLS}\|_1$.

5. Replicate with 100 independent draws of $\mathbf{X}$ and $\mathbf{w}_1$ and average the results.

6. Repeat for a range of ρ from -1 to 1.

6.2.1 Results

We perform the procedure described in the previous section with $n = 100$, $k = 10$, $\sigma_\eta = \sigma_\nu = 0.039$, and $\sigma_w = 0.61$.[1] We capture three dimensions of variation within our simulation setup:

1. The presence vs. absence of endogeneity (i.e., using $\mathbf{w}_1$ vs. $\tilde{\mathbf{w}}_1$),

2. Lasso vs. OLS as the estimation method, and

3. the level of endogeneity as captured through ρ.

[1] These standard deviations are taken from estimates reported in Villas-Boas and Winer (1999).

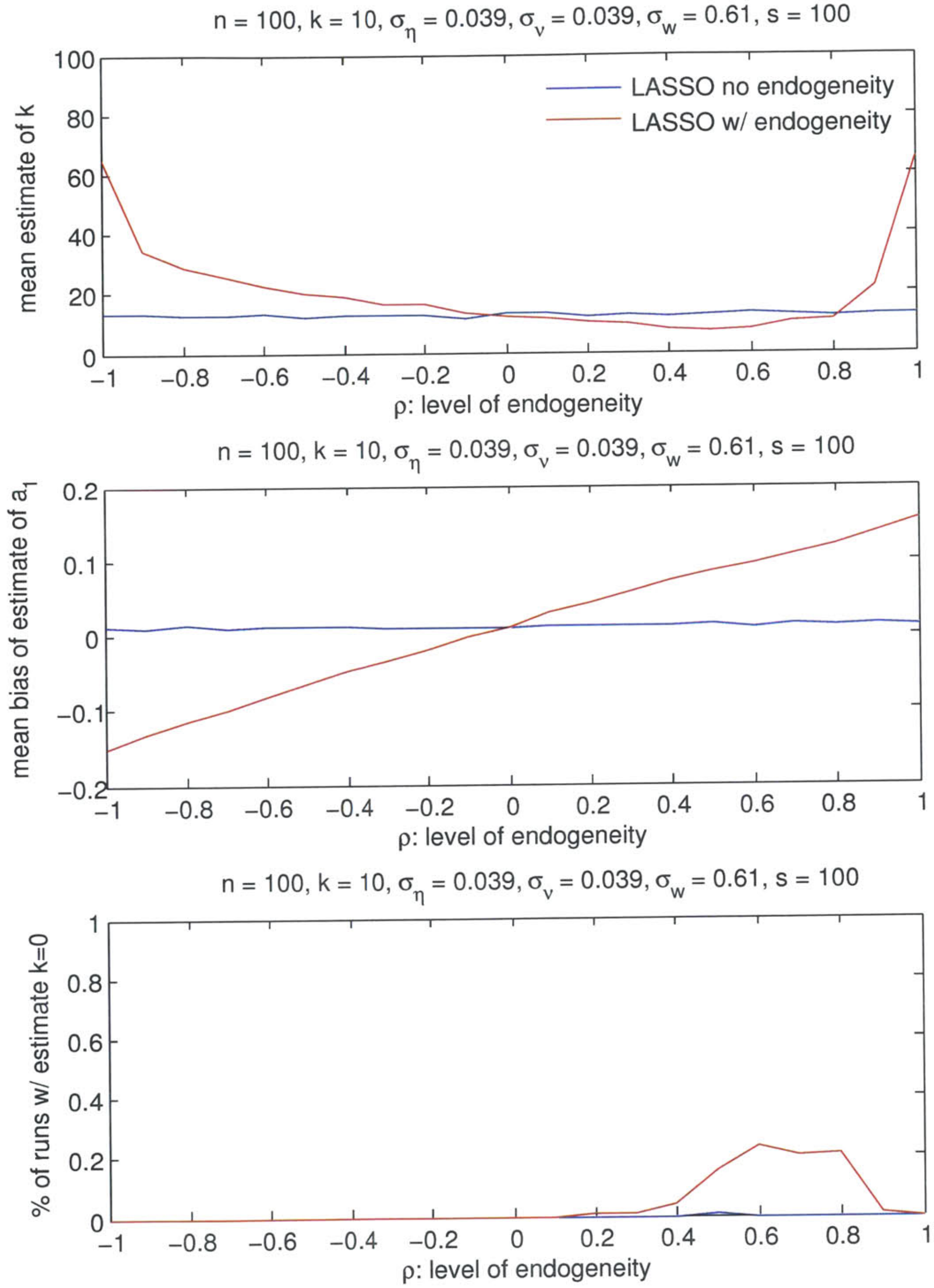


Figure 6-1: Estimates of elasticities are biased due to endogenous prices, but estimates of k are relatively robust even in the presence of endogeneity.

Comparing results across these dimensions allows us to make several observations:

1. Without the presence of endogeneity, the Lasso estimates of $\mathbf{a}_1$ are unbiased and the Lasso estimates of k are quite close to the true value of 10. In contrast, even in the absence of endogeneity, OLS estimates k to be 100, which is the same as the number of products, n. (See the blue curve in the top plot of Figure 6-1. OLS estimates of k are not plotted as they are simply equal to 100.)

2. With the presence of endogeneity, estimates of $\mathbf{a}_1$ are biased as expected. In particular, the mean bias of the estimate of $\mathbf{a}_1$ is negative for negative values of ρ and positive for positive values of ρ, which agrees with our intuition. (See the red curve in the middle plot of Figure 6-1.)

3. For Lasso, although the parameter estimates of $\mathbf{a}_1$ may be biased by endogeneity, its estimates of k are still reasonably accurate for moderate levels of endogeneity. (See the red curve in the top plot of Figure 6-1.) This performance is due to Lasso's penalty function, which rewards sparse estimates. In contrast, because it does not have a penalty function, OLS estimates of k (not shown) are 100 for all values of ρ.

4. Lasso generally overestimates k, with the bias increasing with the magnitude of ρ. This positive bias gives a conservative estimate of k, which in turns leads to a conservative estimate of the number of experiments needed to achieve accurate estimates of $\mathbf{A}$. This is arguably better than underestimating the number of experiments needed and failing to achieve the desired level of estimation accuracy.

5. Lasso can however underestimate k in some cases. Because $\mathbf{a}_1$ is negative on average and positive ρ leads to positively biased estimates, the estimate of $\mathbf{a}_1$ naturally drifts towards 0 when ρ is positive. Furthermore, because $\mathbf{a}_1$ has relatively few nonzero elements and the parameters produce a low signal-to-noise ratio (i.e., $\text{var}(x_i)/\text{var}(w_i)$), the noise term $\mathbf{w}_1$ would dominate $\Delta\mathbf{q}_1$, making it difficult to explain the data through $\mathbf{a}_1$. These two factors, combined with Lasso's penalty

function, tend to produce the zero vector as its estimate of $\mathbf{a}_1$, which would lead to underestimating k in this case. (See the red curve in the bottom plot of Figure 6-1.)

6.2.2 Discussion

The results shown in the previous section indicate that the Lasso-based method of estimating sparsity parameters is robust to endogeneity. Only under extreme levels of endogeneity do the estimates deviate substantially from the true values. In comparison, OLS performs poorly even in the absence of endogeneity: in both settings, it returns dense estimates, which are not helpful for deriving an estimate of the sparsity parameters.

Since the problem of endogeneity is present only in the context of historical data, we are likely to have access to large amounts of this data. Our estimates of the sparsity parameters will continue to improve as we have access to more data. Therefore, for the setting in which we are interested, our methodology appears to provide robust estimates of the level of sparsity, even in the face of moderate endogeneity.

6.3 Empirical analysis

In Section 4.4, we described the results of using historical data to estimate the sparsity parameters k and d, as well as to show that they increase sublinearly with n. Given the use of historical data, endogeneity is a valid concern. In this section, we modify our estimation methodology to account for endogeneity and investigate whether the sublinearity finding is robust.

6.3.1 Instrumental variables

A standard approach to dealing with endogeneity is to use instrumental variables, which are explanatory variables that are (i) correlated with the (possibly) endogenous variable but (ii) uncorrelated with the error term. If a variable satisfies these two conditions, then it can be shown that the variable can be used as a sort of surrogate for the endogenous variable to eliminate the bias in the estimates.

While econometric theory guarantees that, with a valid instrument, estimates will be unbiased, in practice the choice of instrument is critical. Can we actually identify a variable that satisfies the two requirements stated above? Namely, is there a variable that is correlated with retail prices yet uncorrelated with the error term, and is readily available? Wholesale prices have often been used as an instrument for retail prices (see, for example, Chintagunta 2002, Chintagunta et al. 2005, and Sriram et al. 2007). Intuitively, they should be correlated with retail prices because retailers often pass wholesale price changes through to their customers by changing their retail prices.

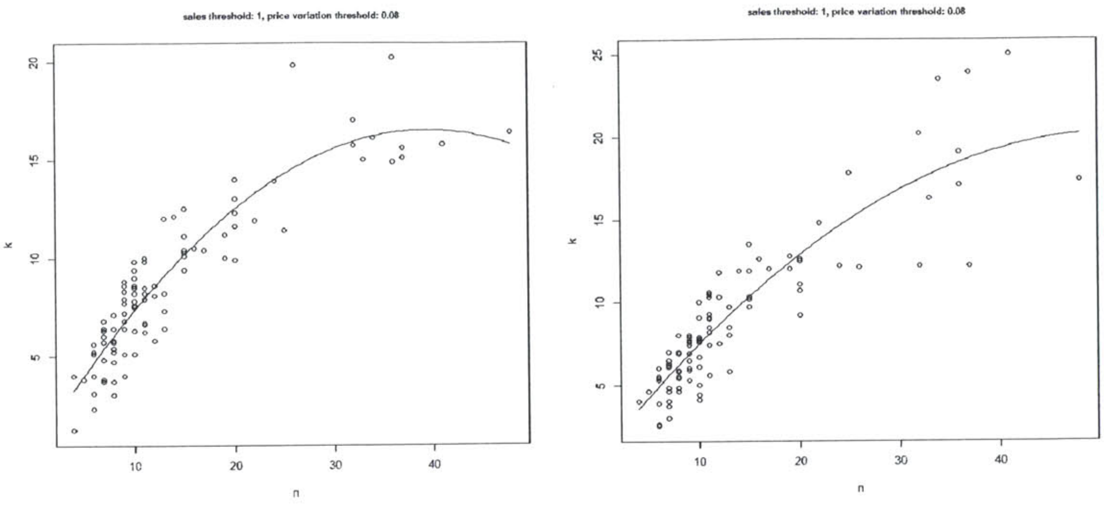
6.3.2 Two-stage instrumental variable approach

There are various methods of incorporating instrumental variables into an estimation procedure. For an OLS-based procedure, one standard method is to take a two-stage approach, which is referred to as "two-stage least squares" (2SLS):

1. In the first stage, regress the endogenous variable on the instrumental variable and obtain the predicted values of the endogenous variable.

2. In the second stage, perform the original regression, substituting in the predicted values of the endogenous variable from the first stage regression for the actual values of the endogenous variable.

As described in Chapter 4, to estimate the sparsity parameters, we use a Lasso regression instead of OLS. Hence, we modify the standard 2SLS approach by performing a Lasso regression in the second stage. More specifically, we follow the same procedure as described in Section 4.4.1, except that we first regress retail prices on wholesale prices for each SKU-store combination using OLS and save the predicted values of retail prices. We then use these predicted retail prices as the $\mathbf{X}$ in the Lasso program.



(a) Estimates of k with original methodology (b) Estimates of k with two-stage methodology

Figure 6-2: Plot of the estimates of k (a) without controlling for endogeneity and (b) using wholesale prices as an instrument for retail prices. The relationship between k and n is similar in both cases.

In order to compare the resulting estimates with and without using wholesale price as an instrument, we perform the remaining steps of the estimation procedure in parallel using both the original retail prices and the predicted retail prices. We use the same filters to ensure that both estimates are based on identical data.

6.3.3 Results

We first report on the first stage of the estimation procedure. Regressing retail prices on wholesale prices results in an R^2 of 0.69 on average for the first stage. Hence, variations in wholesale prices explains a substantial portion of the variation in retail prices.

Next, we compare the estimates of the sparsity parameters using our original methodology as described in Chapter 4 and using the two-stage approach described in this chapter. Figure 6-2 shows the estimates of k using each methodology, plotted against the number of products, n. As the plots suggest, after controlling for endogeneity using wholesale price as an instrument, we obtain estimates of k that are approximately the

	t value	Degrees of freedom	Mean of differences
$H_0 : \hat{k} = \hat{k}^{IV}; H_1 : \hat{k} \neq \hat{k}^{IV}$	-1.712	101	-0.412
$H_0 : \hat{d} = \hat{d}^{IV}; H_1 : \hat{d} \neq \hat{d}^{IV}$	-3.033	101	-22.714

Table 6.1: Paired t-tests comparing the estimates of k and d with and without using instruments to account for heterogeneity. For both k and d, not using instruments results in estimates that are negatively biased. The bias is significant for estimates of d and moderately significant for estimates of k.

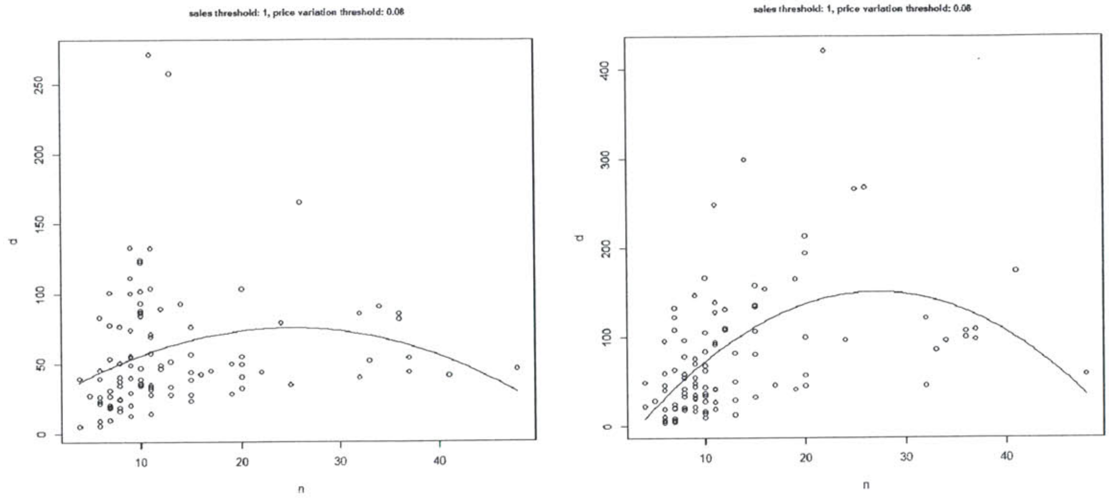
	Coefficient	Estimate	Std. Error	t value
Without instruments	(Intercept)	0.053	0.580	0.091
	1st-order term	0.832	0.071	11.799
	2nd-order term	-0.011	0.002	-6.669
With instruments	(Intercept)	0.632	0.761	0.830
	1st-order term	0.757	0.093	8.174
	2nd-order term	-0.007	0.002	-3.514

Table 6.2: Summary of quadratic fit models for estimates of k with and without controlling for endogeneity. The second-order coefficients are negative and significant in both cases.

same as the estimates of k obtained using our original methodology, without controlling for endogeneity. For each store, we obtain a pair of estimates of k, with and without controlling for endogeneity. For a more precise comparison between the two sets of estimates, we perform a paired t-test using this data with the null hypothesis that the estimates are equal and the alternative hypothesis that the estimates are not equal. As the first row of Table 6.1 shows, estimates of k without using instruments to account for endogeneity are moderately negatively biased, though the bias is not highly significant.

In Chapter 4, we also showed that the relationship between k and n is sublinear. We perform the same analysis here using the two-stage estimates of k. Table 6.2 summarizes the results of fitting a quadratic model between n and estimates of k using both sets of estimates (this quadratic fit is also plotted in Figure 6-2). The second-order coefficients are negative and significant in both cases, suggesting that k increases sublinearly with n, even after accounting for endogeneity.

We also examine our estimates of d using each methodology, which are plotted in



(a) Estimates of d with original methodology (b) Estimates of d with two-stage methodology

Figure 6-3: Plot of the estimates of d (a) without controlling for endogeneity and (b) using wholesale prices as an instrument for retail prices. The relationship between d and n is similar in both cases.

Figure 6-3. The results of fitting a quadratic model between n and estimates of d using both sets of estimates, summarized in Table 6.3, show that the second-order coefficients are negative and significant in both cases, suggesting that d increases sublinearly with n, even after accounting for endogeneity. Comparing the two sets of estimates, the estimates of d follow the same qualitative pattern with and without using instruments, as was the case for estimates of k. We again perform a paired t-test to compare the two sets of estimates of d, the results of which are presented in the second row of Table 6.1. As was the case for estimates of k, the estimates of d without using instruments are negatively biased, though the bias is more significant than it is in the case of k.

This difference is likely due to the fact that the parameter d is more difficult to estimate than the parameter k because by definition it involves the magnitude of each underlying elasticity parameter, whereas k cares only whether each parameter is zero or not. Therefore, any bias in the elasticity parameter estimates due to endogeneity would be more likely to carry over and bias estimates of d. On the other hand, because Lasso's penalty function promotes sparse estimates, the number of nonzero entries may

	Coefficient	Estimate	Std. Error	t value
Without instruments	(Intercept)	20.357	16.099	1.265
	1st-order term	4.409	1.959	2.251
	2nd-order term	-0.088	0.044	-2.008
With instruments	(Intercept)	-45.350	22.798	-1.989
	1st-order term	14.413	2.774	5.195
	2nd-order term	-0.265	0.062	-4.279

Table 6.3: Summary of quadratic fit models for estimates of d with and without controlling for endogeneity. The second-order coefficients are negative and significant in both cases.

	t value	Degrees of freedom	Mean of differences
$H_0 : \hat{a}_{ij} = \hat{a}_{ij}^{IV}; H_1 : \hat{a}_{ij} \neq \hat{a}_{ij}^{IV}$	1.399	267,869	0.016
$H_0 : \hat{a}_{ii} = \hat{a}_{ii}^{IV}; H_1 : \hat{a}_{ii} \neq \hat{a}_{ii}^{IV}$	-7.351	1462	-1.306

Table 6.4: Paired t-tests comparing the estimates of the elasticity parameters with and without using instruments to account for heterogeneity: (i) across all estimates, there is no significant difference; (ii) focusing on the diagonal entries, estimates without using instruments are significantly negatively biased.

not change significantly. In other words, not accounting for endogeneity may still produce many elasticity estimates that are zero (hence, the two sets of estimates of k are similar), but the elasticity estimates that are nonzero may be biased (hence, the two sets of estimates of d are more substantially different).

The fact that the results are not substantially different with and without accounting for endogeneity could be because our estimation method is robust to endogeneity or because the data simply do not exhibit endogeneity. Since our methodology for estimating the sparsity parameters also generates estimates of the $\mathbf{A}$ matrix itself, we now turn our attention to the estimates of the underlying elasticities to see if they are biased when not using instruments. The presence of bias would suggest that there is in fact endogeneity in the data.

We perform a paired t-test between the two sets of elasticity estimates: we match each estimate of a_{ij} without using instruments with the corresponding estimate of the same elasticity parameter using instruments. As the first row of Table 6.4 shows, the two sets of estimates are not significantly different. This is likely due to the fact that many estimates are 0. Hence we focus the comparison on only the nonzero estimates of the diagonal entries of **A**. As the second row of Table 6.4 shows, these elasticity estimates are now significantly negatively biased when instruments are not used. This suggests that endogeneity bias does exist in the data. However, the effect of this bias is relatively diminished when we move up a level and consider estimating the sparsity parameters, particularly k, instead of the raw elasticity parameters. The estimates of k do not exhibit much bias when not using instruments, and hence our estimation method appears to be robust to endogeneity.

6.4 Summary

In this chapter, we investigated the potential of endogeneity in our data, which could lead to biased estimates of elasticities in **A** and of the sparsity parameters k and d. Using a standard approach to incorporate endogeneity into our model, we first generated synthetic price and demand data that contained varying degrees of endogeneity and performed simulations that showed our estimation method still produces accurate estimates even with endogenous data. Next, we analyzed the same historical sales data as in Chapter 4, this time accounting for endogeneity by using wholesale prices as an instrument for retail prices. Using this two-stage instrumental variable approach, we show that although the underlying estimates of elasticities are biased when not accounting for endogeneity, the estimates of the sparsity parameters show less significant bias, particularly in the case of k. These results not only suggest that our method for estimating sparsity parameters is relatively robust to endogeneity, but also give us more confidence that our empirical estimates of the sparsity parameters based on real historical data, as well as our finding that they grow sublinearly with n, are indeed accurate.

Chapter 7

Profit maximization

Thus far, we have presented a complete model of demand and provided a method for estimating price elasticities quickly and accurately. A firm's goal may be simply to estimate this demand function, but firms are likely more interested in the end result, which is, for example, to use this demand function to choose an optimal set of prices in order to maximize profit. In this chapter, we study the problem of profit maximization.

7.1 Modeling the profit function

Recall our standard model of demand:

$$\Delta \mathbf{q} = \frac{\mathbf{q}^t - \mathbf{q}^b}{\mathbf{q}^b} = \mathbf{A}\mathbf{x} + \mathbf{w},$$

where $\Delta \mathbf{q}$ is the percentage change in quantities demanded and $\mathbf{x}$ is the percentage change in prices. To extend the model to profit maximization, let $\mathbf{x}^b$ be the vector of baseline prices for each product and $\mathbf{p}$ be the vector of profit margins for each product when prices are set at $\mathbf{x}^b$. We assume that $\mathbf{x}^b$ and $\mathbf{p}$ are known. We can then express the profit due to a decision $\mathbf{x}$ as

$$(\mathbf{p} + \mathbf{x}^b \circ \mathbf{x})^T \mathbf{q}^t = (\mathbf{p} + \mathbf{x}^b \circ \mathbf{x})^T [\mathbf{q}^b \circ (\mathbf{e} + \mathbf{A}\mathbf{x} + \mathbf{w})],$$

where $\circ$ denotes element-wise multiplication. The first term gives the modified profit margins given the change in prices, $\mathbf{x}$, and the second term gives the usual treatment demand level.

7.2 Maximizing expected profit gain

Because the demand levels are random variables, the resulting profit is also a random variable. A reasonable statistic to maximize is the expected profit[1]:

$$\mathbb{E}_{\mathbf{w}}\Big[(\mathbf{p}+\mathbf{x}^b \circ \mathbf{x})^T[\mathbf{q}^b \circ (\mathbf{e}+\mathbf{A}\mathbf{x}+\mathbf{w})]\Big].$$

Given our assumption of zero-mean noise (Assumption 1), the expression for expected profit simplifies to

$$\begin{aligned} & (\mathbf{p}+\mathbf{x}^b \circ \mathbf{x})^T[\mathbf{q}^b \circ (\mathbf{e}+\mathbf{A}\mathbf{x})] \\ = \; & \mathbf{p}^T\mathbf{q}^b + \mathbf{p}^T(\mathbf{q}^b \circ \mathbf{A}\mathbf{x}) + (\mathbf{x}^b \circ \mathbf{x})^T\mathbf{q}^b + (\mathbf{x}^b \circ \mathbf{x})^T(\mathbf{q}^b \circ \mathbf{A}\mathbf{x}). \end{aligned}$$

The first term gives the baseline profit level. Since it has no dependence on $\mathbf{x}$, we need consider only the remaining terms when maximizing profit: we maximize the expected *gain* in profit.

Let us define

$$P(\mathbf{x}) \triangleq \mathbf{p}^T(\mathbf{q}^b \circ \mathbf{A}\mathbf{x}) + (\mathbf{x}^b \circ \mathbf{x})^T\mathbf{q}^b + (\mathbf{x}^b \circ \mathbf{x})^T(\mathbf{q}^b \circ \mathbf{A}\mathbf{x}) \tag{7.1}$$

as the expected profit gain due to a decision $\mathbf{x}$. Note that the expected profit gain decomposes into three parts: the first term gives the gain due to the change in demand only, the second term gives the gain due to the change in profit margins only, and the third term gives the gain due to the combination of both changes.

Given the matrix $\mathbf{A}$, we maximize profit by solving the following optimization prob-

[1] Henceforth, we will use the terms "profit" and "expected profit" interchangeably.

lem:

$$\max_{-\rho \le \mathbf{x} \le \rho} P(\mathbf{x}),$$

where we constrain the price variation of each product to be no more than ρ in magnitude.

Let $\mathbf{x}^*$ be an optimal set of prices and P^* be the optimal profit gain. Because we do not know the true $\mathbf{A}$ matrix, we may not be able to discover $\mathbf{x}^*$ nor achieve this optimal profit gain. Instead, we have only an estimate $\hat{\mathbf{A}}$, based on which we choose some $\hat{\mathbf{x}}$ that achieves a profit gain of $\hat{P} \triangleq P(\hat{\mathbf{x}})$. We are interested in how close we come to achieving P^* and how many experiments that requires.

7.2.1 Knowing the true A matrix

First we consider the ideal case where we know the true $\mathbf{A}$ matrix exactly. Note that Equation (7.1) is quadratic in $\mathbf{x}$. After some algebra, we find that the profit maximization problem is equivalent to the following constrained quadratic program:

$$P^* = \max_{-\rho \le \mathbf{x} \le \rho} \mathbf{x}^T \mathbf{H} \mathbf{x} + \mathbf{f}^T \mathbf{x}, \tag{7.2}$$

where

$$\begin{aligned} h_{ij} &= x_i^b q_i^b a_{ij}, \\ f_i &= \sum_{\ell=1}^{n} p_\ell q_\ell^b a_{\ell i} + x_i^b q_i^b. \end{aligned}$$

Since $\mathbf{H}$ and $\mathbf{f}$ are known, we can solve this quadratic program using one of various algorithms and available software libraries. The quantity P^* gives the maximum possible profit gain, assuming we know $\mathbf{A}$ perfectly. Because choosing $\mathbf{x} = \mathbf{0}$ (i.e., not changing any prices) is always an option, we know that by maximizing over $\mathbf{x}$, $P^* \ge 0$. In other words, taking advantage of our knowledge of $\mathbf{A}$ can only help to increase the expected profit.

7.2.2 Knowing an estimate $\hat{\mathbf{A}}$

In reality, we do not know the true $\mathbf{A}$ matrix; instead, we have an estimate $\hat{\mathbf{A}}$ given by our experiments and estimation method described in Section 3. Hence, we do not have the true values of $\mathbf{H}$ and $\mathbf{f}$ to be used in (7.2). A simple approximation is to calculate estimates $\hat{\mathbf{H}}$ and $\hat{\mathbf{f}}$ using the estimate $\hat{\mathbf{A}}$ and to find $\hat{\mathbf{x}}$ that solves

$$\hat{\mathbf{x}} \in \arg\max_{-\rho \leq \mathbf{x} \leq \rho} \mathbf{x}^T \hat{\mathbf{H}} \mathbf{x} + \hat{\mathbf{f}}^T \mathbf{x}. \tag{7.3}$$

This approximate $\hat{\mathbf{x}}$ can also be found numerically.

Suppose we believe that the true $\mathbf{A}$ matrix is sparse, having at most k nonzero entries in each row. Then we can use this information to try to improve the preceding procedure. Specifically, we can perform a form of thresholding on our estimate $\hat{\mathbf{A}}$. In particular, we choose some κ (e.g., based on an estimate of k as obtained using the method described in Chapter 4), retain only the κ largest entries in magnitude in each row of $\hat{\mathbf{A}}$, and set all other entries to zero. In doing so, we try to take advantage of our knowledge of the sparse structure of $\mathbf{A}$. The benefit of this thresholding procedure depends on our choice of κ and how it compares to the true sparsity index k. However, we will show using simulations that overall performance is largely insensitive to incorrect values of κ.

7.2.3 Performance of profit maximization

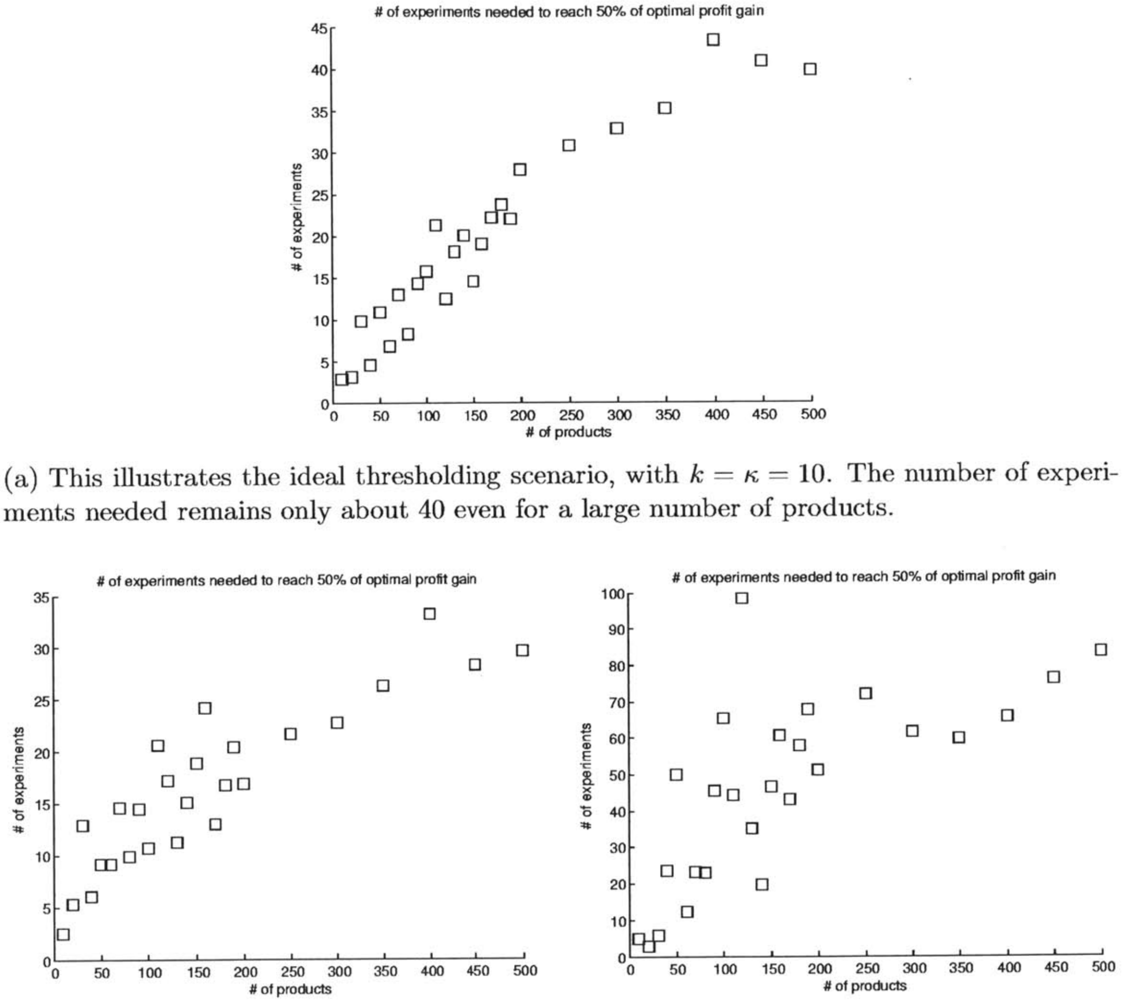
Because different retailers have different product assortments with different associated elasticities, one store or category may have more potential for profit gains on an absolute scale than another. In terms of our model, one $\mathbf{A}$ matrix may have more inherent potential profit gains to be extracted than another. As such, we judge our profit maximization procedure not by the absolute profit gain we achieve but by the fraction of the maximum profit gain that we can capture. Our quantity of interest is therefore the ratio $\hat{P}/P^*$.

We perform simulations to test the speed with which we can achieve significant profit gains in the case of sparse $\mathbf{A}$ matrices. We generate a sequence of matrices $\{\mathbf{A}_n\}$ of increasing size, with some true sparsity level, using the seed distributions. We also draw the quantities $x_i^b q_i^b$ and $p_i q_i^b$ from distributions seeded by parameters obtained from the same study that was used to generate the $\mathbf{A}$ matrices. By following the methodology described in Chapter 3, we obtain a sequence of estimates $\{\hat{\mathbf{A}}_n\}$. We select a value of κ, which is an estimate of the true sparsity parameter k, and threshold the estimates $\{\hat{\mathbf{A}}_n\}$ as described in Section 7.2.2. We then use these thresholded estimates to calculate $\hat{\mathbf{H}}$ and $\hat{\mathbf{f}}$. By solving (7.3), we obtain the decision $\hat{\mathbf{x}}$ and the corresponding profit gain $\hat{P}$. By using the true $\mathbf{A}$ matrix that was initially generated, we also find an optimal decision vector $\mathbf{x}^*$ and the optimal profit gain P^*. As a measure of the profit maximization procedure's performance, we record the number of experiments needed to achieve 50% of the maximum possible profit gain.

A key parameter in the maximization procedure is the threshold level κ: the number of elements in each row of $\hat{\mathbf{A}}$ to retain. In reality, even when the true $\mathbf{A}$ matrix is sparse, we will not know the exact sparsity constant k and will need to approximate it with an estimate. Therefore, we would like our procedure to be robust to a range of choices of κ. In the ideal scenario, we correctly select $\kappa = k$. In this case, as Figure 7-1a shows, the number of experiments needed to achieve 50% of the maximum profit gain is relatively small. Fortunately, even if we choose κ to be too small or too large, we still achieve similar results (see Figures 7-1b and 7-1c).

7.2.4 Using Lasso estimates

An alternative method of estimating the $\mathbf{A}$ matrix is to solve the Lasso program as defined in (4.1). In Chapter 4, we used this method to obtain estimates of the sparsity parameters. However, the Lasso also produces estimates of the $\mathbf{A}$ matrix itself. Recall that the estimation is performed row by row. Hence, we perform the methodology given in Section 4.2 on each row and combine the resulting estimates to construct



(a) This illustrates the ideal thresholding scenario, with $k = \kappa = 10$. The number of experiments needed remains only about 40 even for a large number of products.

(b) In this scenario, we choose $\kappa = 2$ when the true k is 10. Despite this incorrect choice, the number of experiments needed still remains relatively small even for many products.

(c) In this scenario, we choose $\kappa = 20$ when the true k is 10. Despite this incorrect choice, the number of experiments needed again remains relatively small even for many products.

Figure 7-1: In these plots, we illustrate the number of experiments needed to achieve 50% of the optimal profit gain for matrices of increasing size and under various choices of κ.

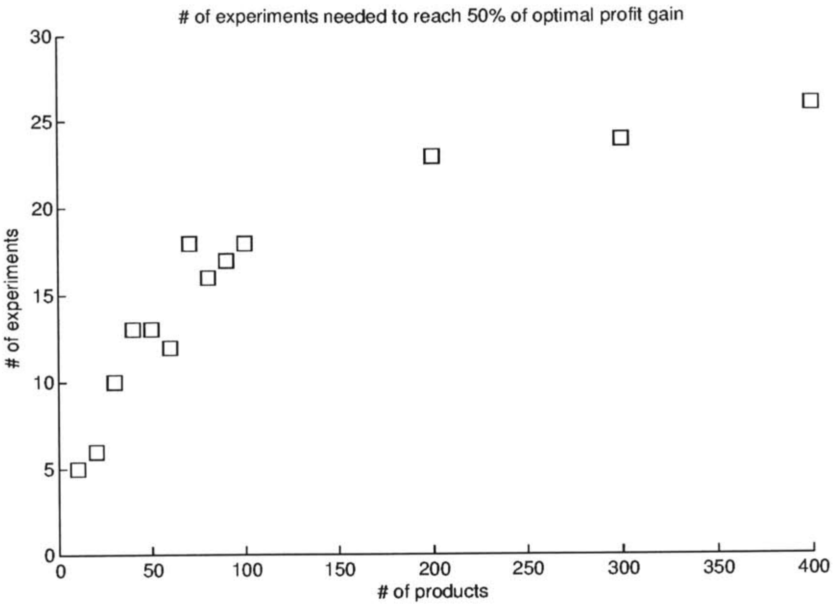


Figure 7-2: By using Lasso to estimate the **A** matrix, we also have that the number of experiments needed to achieve 50% of the optimal profit gain is relatively small. The number of experiments needed also grows sublinearly with the number of products.

the complete estimate $\hat{\mathbf{A}}$. Since the cross-validation process implicitly estimates the sparsity parameter, we do not need to pick a value of κ as in Section 7.2.2. As Figure 7-2 illustrates, by using Lasso estimates, we again require few experiments in order to achieve 50% of the optimal profit gain. Moreover, the number of experiments needed also grows sublinearly with the number of products. Hence, using a relatively small number of experiments, we can not only obtain accurate estimates of elasticities, but we can also use these estimates to obtain substantial gains in profit.

Chapter 8

Promotional decisions

Besides setting prices, firms make many other types of marketing decisions, including which products to advertise or promote. Although our analysis has been focused on pricing decisions, our model can be adapted to advertising or promotional decisions, which we shall consider in this chapter.

8.1 Modeling demand

As with setting prices, promoting a product will affect its demand, and the substitution and complementarity effects between products will also carry over to promotional decisions. Therefore, we can again use a matrix $\mathbf{A}$ to represent the own- and cross-product elasticities and a vector $\Delta\mathbf{q}$ to represent the percentage change in demand for each product, compared between treatment and control conditions. However, some modifications are required to adapt the model to the setting of promotional decisions.

If we interpret the decision to advertise or promote a product as a binary decision, then the decision variables become

$$x_j = \begin{cases} 1, & \text{if product } j \text{ is promoted,} \\ 0, & \text{if product } j \text{ is not promoted.} \end{cases}$$

For ease of exposition (and without loss of generality), we will assume that there are no promotions under the control condition. We can then model the percentage change in demand in response to the promotional decisions as

$$\Delta \mathbf{q} = \frac{\mathbf{q}^t - \mathbf{q}^b}{\mathbf{q}^b} = \mathbf{A}\mathbf{x} + \mathbf{w}.$$

In this model, we capture own- and cross-product promotional responses in the $\mathbf{A}$ matrix. This model retains the same form as our standard model, defined in Equation (2.4), with a different interpretation for the decision variables x_j and elasticities a_{ij}.

8.2 Estimating A

Given that the model under promotional decisions has the same form as the model under pricing decisions, we can apply a modified form of our estimation procedure to obtain similar results. Specifically, we now make 0/1 Bernoulli decisions for each x_j under the promotional setting. This is essentially the same setup as under the pricing setting, and we can again find estimators for each a_{ij} such that Theorem 1 holds (with slightly different constants). Therefore, we are still able to achieve uniformly ϵ-accurate estimation with $O(k \log n)$ experiments under sparsity and $O(d^2 \log n)$ experiments under bounded influence.

8.3 Maximizing profit

We can also consider the profit maximization problem under the promotional setting. Promoting a product simply calls attention to that particular product and does not change its price. Therefore, the firm's actions do not affect profit margins, and so the resulting optimization problem does not have the same form as the one under the pricing setting.

To extend our demand model to profit maximization, let $\mathbf{p}$ be the vector of profit margins for each product, which are assumed to be known. We can then express the expected profit due to a decision $\mathbf{x}$ as

$$\mathbb{E}[\mathbf{p}^T\mathbf{q}^t] = \mathbb{E}_{\mathbf{w}}\left[\mathbf{p}^T[\mathbf{q}^b \circ (\mathbf{e} + \mathbf{A}\mathbf{x} + \mathbf{w})]\right],$$

where $\circ$ denotes element-wise multiplication. Given our assumption of zero-mean noise (Assumption 1), this simplifies to

$$\begin{aligned}\mathbb{E}[\mathbf{p}^T\mathbf{q}^t] &= \mathbf{p}^T[\mathbf{q}^b \circ (\mathbf{e} + \mathbf{A}\mathbf{x})]\\ &= \mathbf{p}^T\mathbf{q}^b + \mathbf{p}^T(\mathbf{q}^b \circ \mathbf{A}\mathbf{x}).\end{aligned}$$

The first term is the baseline level of profit and the second term is the gain (or loss) in profit due to the decision vector $\mathbf{x}$. As in the case of optimizing prices, we can maximize just the expected gain in profit since the first term does not depend on $\mathbf{x}$. Following Section 7.2, we calculate the maximum profit gain assuming the true $\mathbf{A}$ matrix is known and use it as a benchmark to compare against the profit gain obtained from an estimate $\hat{\mathbf{A}}$.

8.3.1 Knowing the true A matrix

First we consider the ideal case where we know the true $\mathbf{A}$ matrix exactly. We are maximizing

$$\mathbf{p}^T(\mathbf{q}^b \circ \mathbf{A}\mathbf{x}) = (\mathbf{p} \circ \mathbf{q}^b)^T(\mathbf{A}\mathbf{x}). \tag{8.1}$$

Since this function is linear in $\mathbf{x}$, it can be maximized element-wise (i.e., independently for each product):

1. Compute $\mathbf{v} = \mathbf{p} \circ \mathbf{q}^b$, the vector of baseline profits for each product.

2. For each product i, compute $\pi_i = \mathbf{v}^T\mathbf{A}_i$, where $\mathbf{A}_i$ is the i^{th} column of $\mathbf{A}$. This gives the change in profit if product i were promoted.

3. If $\pi_i > 0$, set $x_i^* = 1$. Otherwise, set $x_i^* = 0$. The resulting vector $\mathbf{x}^*$ is an optimal solution to (8.1).

The preceding algorithm calculates the potential change in profit if a product were promoted and chooses to promote it if and only if this change is positive. Although products are coupled by their substitution and complementarity effects through the matrix $\mathbf{A}$, once we know this matrix, the profit maximization problem is very simple: we can decide whether or not to promote each item independently by calculating its potential contribution to the gain in profit.

8.3.2 Knowing an estimate $\hat{\mathbf{A}}$

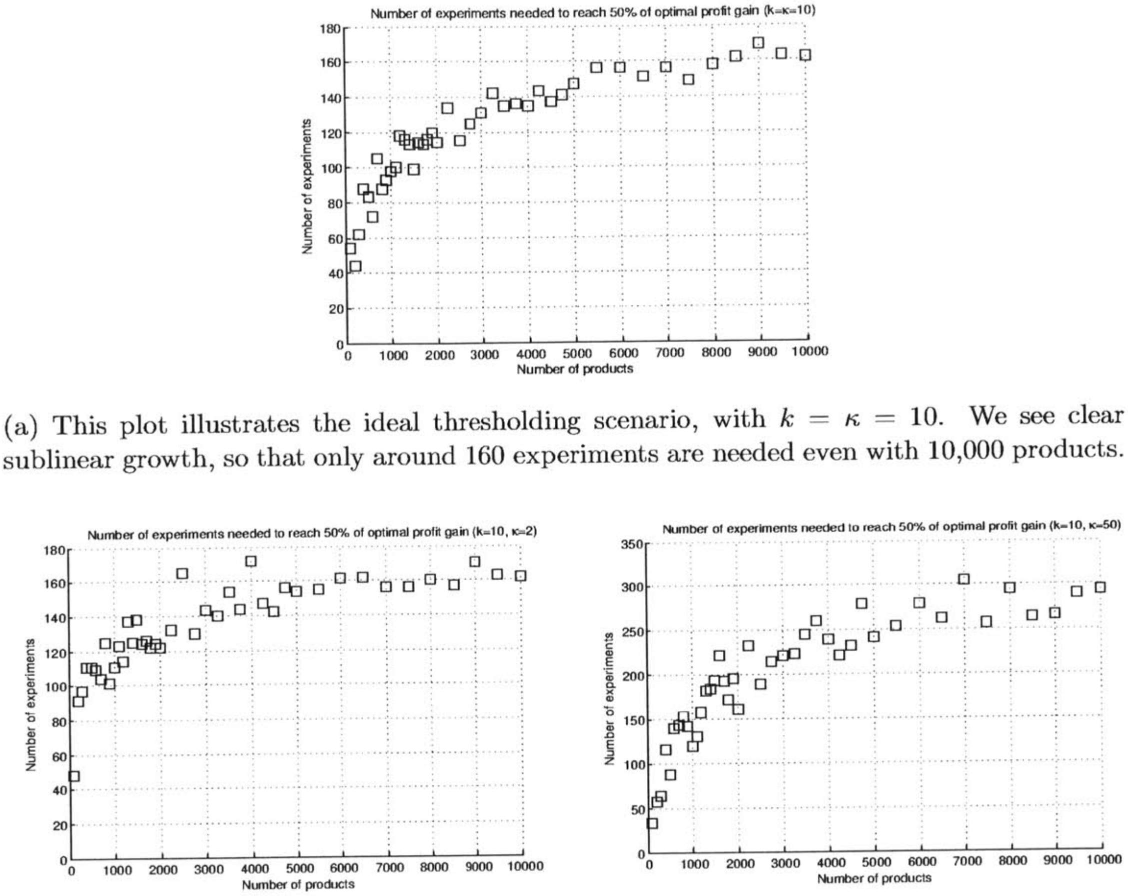
Since we do not know the true $\mathbf{A}$ matrix, we instead use our estimate $\hat{\mathbf{A}}$. As described in Section 7.2.2, we employ thresholding on $\hat{\mathbf{A}}$ and keep only the κ largest entries in magnitude in each row. Given $\hat{\mathbf{A}}$, a simple approximation is to find a vector $\hat{\mathbf{x}}$ that maximizes the expected profit gain assuming that $\hat{\mathbf{A}}$ is correct:

$$\hat{\mathbf{x}} \in \underset{\mathbf{x}\in\{0,1\}^n}{\arg\max}\ \mathbf{p}^T(\mathbf{q}^b \circ \hat{\mathbf{A}}\mathbf{x}).$$

This approximate $\hat{\mathbf{x}}$ can be found in a similar manner as $\mathbf{x}^*$: simply substitute the estimated columns $\hat{\mathbf{A}}_i$ for the true columns $\mathbf{A}_i$ in Step 2 of the algorithm described in Section 8.3.1.

8.3.3 Performance of profit maximization

As with the case of optimizing pricing decisions, we focus on the fraction of the maximum profit gain that we can capture. Using a similar procedure as the one described in Section 7.2.3, we perform simulations to determine the number of experiments needed to achieve 50% of the maximum possible profit gain. Because we are now considering the promotional setting, we draw the elements of the $\mathbf{A}$ matrix from different seed distributions. The study that was used to obtain seed distributions under the pricing



(a) This plot illustrates the ideal thresholding scenario, with $k = \kappa = 10$. We see clear sublinear growth, so that only around 160 experiments are needed even with 10,000 products.

(b) In this scenario, we choose $\kappa = 2$ when the true k is 10. Despite this incorrect choice, we still see a clear sublinear growth relationship.

(c) Here, we choose $\kappa = 50$ when the true k is 10. Despite this incorrect choice, we again see a clear sublinear growth relationship.

Figure 8-1: In these plots, we illustrate the number of experiments needed to achieve 50% of the optimal profit gain under the promotional setting for matrices of increasing size and under various choices of κ.

setting also experimentally manipulated promotional decisions, and hence we can also obtain seed distributions for the promotional setting from that dataset.

As Figure 8-1 illustrates, the number of experiments needed is relatively small and grows sublinearly with the number of products. Moreover, the performance is similar for different choices of κ, demonstrating that our method is robust.

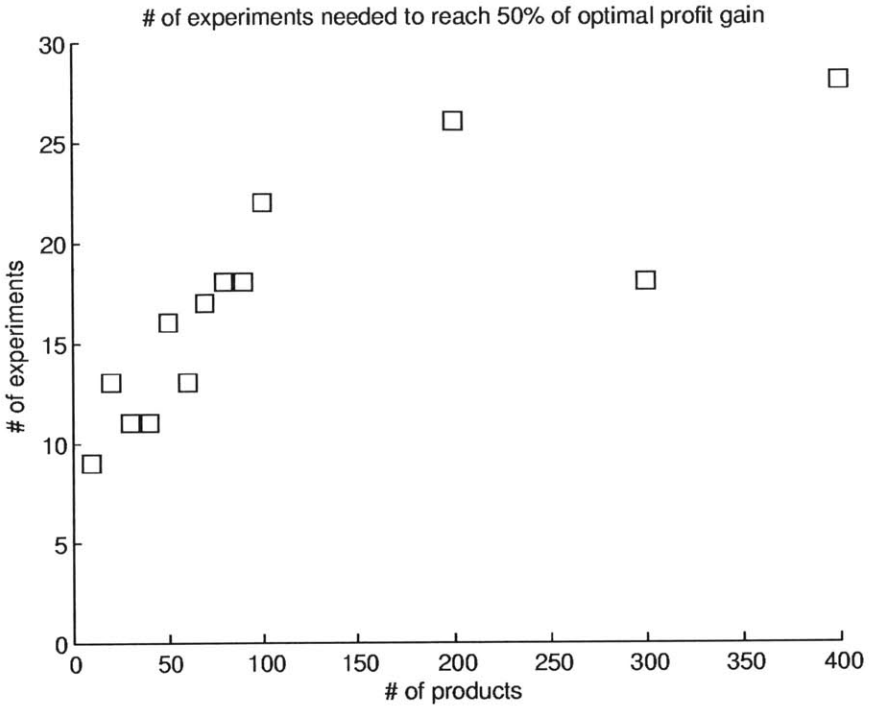


Figure 8-2: Under the promotional setting, using Lasso estimates of the **A** matrix also achieves 50% of the optimal profit gain with relatively few experiments. The number of experiments needed also grows sublinearly with the number of products.

8.3.4 Using Lasso estimates

As in Section 7.2.4, we can also use Lasso to estimate **A** and use the resulting $\hat{\mathbf{A}}$ in the profit maximization algorithm. As Figure 8-2 illustrates, by using Lasso estimates, we again require few experiments in order to achieve 50% of the optimal profit gain. Moreover, the number of experiments needed also grows sublinearly with the number of products. Therefore, for the problem of choosing which products to advertise or promote, managers can also achieve significant profit gains using a practically feasible number of experiments.

Chapter 9

Conclusions

While many firms lack the capabilities to estimate sophisticated econometric models, almost any firm can compare the results between experimental treatment and control groups. We have investigated whether conducting these simple comparisons can help firms improve their managerial decisions even as the complexity of the problem grows. In particular, we consider settings where actions taken to impact the sales of one product tend to spill over and also affect sales of other products. As the number of products, n, grows, the number of parameters to estimate grows as $O(n^2)$. This suggests that the number of experiments required to estimate these parameters will quickly grow beyond what is feasible.

However, we show that if the category exhibits a favorable structure, then firms can learn these parameters accurately using a relatively small number of experiments. We investigate two such structures. The first is sparsity, in which any one product can be affected by at most k products. An important point is that we do not need to know which specific products affect that one product's demand, only that there is a limit to how many such products there are. Given this restriction, the number of experiments required to estimate the matrix of elasticities drops from $O(n \log n)$ to $O(k \log n)$.

We also describe a second restriction that yields similar results. Rather than limiting the number of products that can affect any one product, it may be more appropriate to restrict how much the total percentage change in sales of one product can be affected by actions on all of the products. As long as there is a limit to the aggregate magnitude of these interactions, then we again have a favorable scaling of the number of experiments with the number of products.

To investigate whether these favorable structures exist, we propose a method for estimating the level of sparsity in a given category. We use this method to analyze actual historical sales data and estimate the sparsity parameters for the "Cold remedies" category. The empirical results show that sparse structures do appear to exist. In estimating the sparsity parameters, we also obtain estimates of elasticities. Using these preliminary elasticity estimates to help design subsequent experiments is an interesting opportunity for future research.

Given actual estimates of the sparsity parameters, we then test our main method of estimating the matrix of elasticities using a simulation seeded from real experimental data. The results verify that the number of experiments needed to obtain accurate estimates does indeed grow logarithmically with the number of products, as our theory predicts.

The use of historical data is convenient because it is often readily available. However, care must be taken because historical prices are often endogenous, which may lead to biased parameter estimates. To account for potential endogeneity, we modify our method of estimating the sparsity parameters to use an instrumental variable approach. The results show that our data does exhibit endogeneity but that our estimation methodology is robust to this source of bias.

In addition to estimating elasticity parameters, we have also explored the problem of choosing prices in order to maximize profit. We propose an algorithm for making this decision and use simulations to show that it achieves significant profit gains using relatively few experiments. Hence, experiments are valuable not only for learning elasticities but also for making decisions and realizing tangible benefits.

Our findings provide guarantees about the rate of learning from experiments. These guarantees are obtained using randomized experiments and simple comparisons of outcomes between treatment and control conditions. Firms may increase the rate of learning by optimizing the experimental designs and/or using more sophisticated analyses to estimate the parameters. While our guarantees will continue to hold under these alternative approaches, future research may investigate the extent to which the bounds can be improved in these circumstances.

We have framed our findings by focusing on pricing decisions. However, the results can be extended to other marketing decisions in which actions targeted at an individual product spill over to affect other products as well. In the context of learning demand elasticities, we have extended our findings to choosing which products to promote and demonstrated that similar results are obtained as in the case of choosing prices. Other applications to which our model could apply include the allocation of sales force resources across products or the focus of future investments in product development. It may also be possible to extend the results to settings in which marketing actions targeted at one customer (or group of customers) also impact the decisions of other customers. Spillovers between customers may arise when customers can observe the decisions of other customers, or when their decisions depend on the recommendations of other customers. Extending our results to these forms of externalities may present fertile opportunities for future research.

Bibliography

Anderson, E. T., Cho, E., Harlam, B. A., and Simester, D. I. (2010). What affects price and price cue elasticities? Evidence from a field experiment. Working paper.

Anderson, E. T. and Simester, D. I. (2001). Are sale signs less effective when more products have them? *Marketing Science*, 20(2):121–142.

Anderson, E. T. and Simester, D. I. (2013). Advertising in a competitive market: The role of product standards, customer learning, and switching costs. *Journal of Marketing Research*, 50(4):489–504.

Candès, E. J. (2006). Compressive sampling. *Proceedings of the International Congress of Mathematicians.*

Candès, E. J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509.

Candès, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215.

Chintagunta, P. K. (2002). Investigating category pricing behavior at a retail chain. *Journal of Marketing Research*, 39(2):141–154.

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

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* John Wiley and Sons, New York.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. *Proc. 21st Annual Conf. Learn. Theory (COLT 2008).*

Dubé, J.-P., Hitch, G. J., Rossi, P. E., and Vitorino, M. A. (2008). Category pricing with state-dependent utility. *Marketing Science*, 27(3):417–429.

Erdem, T. (1996). A dynamic analysis of market structure based on panel data. *Marketing Science*, 15(4):359–378.

Farias, V., Jagabathula, S., and Shah, D. (2013). A nonparametric approach to modeling choice with limited data. *Management Science*, 59(2):305–322.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Hauser, J. R., Toubia, O., Evgeniou, T., Befurt, R., and Dzyabura, D. (2010). Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *Journal of Marketing Research*, 47(3):485–496.

Huber, J. and Zwerina, K. (1996). The importance of utility balance in efficient choice designs. *Journal of Marketing Research*, 33(3):307–317.

Keane, M. P. (1997). Modeling heterogeneity and state dependence in consumer choice behavior. *Journal of Business & Economic Statistics*, 15(3):310–327.

Lewis, R. A. and Rao, J. M. (2012). On the near impossibility of measuring advertising effectiveness. Working paper.

Liu, Q. and Arora, N. (2011). Efficient choice designs for a consider-then-choose model. *Marketing Science*, 30(2):321–338.

Louviere, J. J., Hensher, D. A., and Swait, J. D. (2000). *Stated Choice Methods: Analysis and Application.* Cambridge University Press, Cambridge, UK.

Louviere, J. J., Street, D., and Burgess, L. (2004). A 20+ years' retrospective on choice experiments. In Wind, J., editor, *Tribute to Paul Green*, International Series in Quantitative Marketing, chapter 8. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Manchanda, P., Ansari, A., and Gupta, S. (1999). The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing Science*, 18(2):95–114.

Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995). *Microeconomic Theory.* Oxford University Press, New York.

Mersereau, A. J., Rusmevichientong, P., and Tsitsiklis, J. N. (2009). A structured multi-armed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12):2787–2802.

Sandor, Z. and Wedel, M. (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research*, 38(4):430–444.

Seetharaman, P. B., Ainslie, A., and Chintagunta, P. K. (1999). Investigating household state dependence effects across categories. *Journal of Marketing Research*, 36(4):488–500.

Sriram, S., Balachander, S., and Kalwani, M. U. (2007). Monitoring the dynamics of brand equity using store-level data. *Journal of Marketing*, 71(2):61–78.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288.

Toubia, O., Simester, D. I., Hauser, J. R., and Dahan, E. (2003). Fast polyhedral adaptive conjoint estimation. *Marketing Science*, 22(3):273–303.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, 79(4):281–299.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

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

Wainwright, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741.

Appendix A

Multiplicative demand model

Consider an alternative multiplicate demand model of the following form:

$$\Delta q_i = x_1^{a_{i1}} x_2^{a_{i2}} \cdots x_n^{a_{in}} w_i.$$

Taking the logarithm of both sides, we obtain

$$\begin{aligned}\log(\Delta q_i) &= a_{i1}\log(x_1) + a_{i2}\log(x_2) + \cdots + a_{in}\log(x_n) + \log(w_i) \\ &= \sum_{\ell=1}^{n} a_{i\ell}\log(x_\ell) + \log(w_i).\end{aligned}$$

By defining $\Delta\tilde{q}_i \triangleq \log(\Delta q_i), \tilde{x}_\ell \triangleq \log(x_\ell)$, and $\tilde{w}_i \triangleq \log(w_i)$, we can rewrite the above as

$$\Delta\tilde{q}_i = \sum_{\ell=1}^{n} a_{i\ell}\tilde{x}_\ell + \tilde{w}_i,$$

which is of the same form as our standard linear additive model.

Suppose that the noise term w_i is log-normally distributed and hence $\tilde{w}_i \sim N(0, c^2)$.[1] We are free to choose the decisions x_ℓ, and so let us choose each one randomly by first choosing u_ℓ uniformly from the interval $[-\rho, \rho]$ and then assigning $x_\ell = e^{u_\ell}$. Thus, $\tilde{x}_\ell \sim$

[1] More generally, we can relax this assumption – we require only that $\tilde{w}_i$ is zero-mean and sub-Gaussian with parameter c.

$U[-\rho, \rho]$. We continue to assume independence among the x's and w's, which translates into independence among the $\tilde{x}$'s and $\tilde{w}$'s. Therefore, we can apply the same estimation method as described in Chapter 3 to learn the $\mathbf{A}$ matrix under this multiplicative model. In particular, the statistic defined in Section 3.5 becomes $\tilde{y}_{ij} \triangleq \beta \cdot \Delta\tilde{q}_i \cdot \tilde{x}_j$, which would again be an unbiased estimator of a_{ij}. In addition, our methodology for estimating k and d and our profit maximization algorithm, presented in Chapters 4 and 7, can be similarly adapted to fit the multiplicative model.

Appendix B

Asymptotic notation

Let $\mathbf{n}$ be a vector of variables; then we say:

(i) $f(\mathbf{n}) \in O(g(\mathbf{n}))$ if there exist constants N and $C > 0$ such that $|f(\mathbf{n})| \leq C|g(\mathbf{n})|$ for all $\mathbf{n}$ such that $n_i > N$, $\forall i$;

(ii) $f(\mathbf{n}) \in \Omega(g(\mathbf{n}))$ if there exist constants N and $C > 0$ such that $|f(\mathbf{n})| \geq C|g(\mathbf{n})|$ for all $\mathbf{n}$ such that $n_i > N$, $\forall i$;

(iii) $f(\mathbf{n}) \in \Theta(g(\mathbf{n}))$ if $f(\mathbf{n}) \in O(g(\mathbf{n}))$ and $f(\mathbf{n}) \in \Omega(g(\mathbf{n}))$.

In the first case, $f(n) \in O(g(n))$ essentially means that $f(n)$ grows *no faster* than $g(n)$ as n becomes large. In this sense, $g(n)$ can be thought of as an "upper bound" on the rate of growth of $f(n)$. An example is $f(n) = 100n$ and $g(n) = n^2$.

In the second case, $f(n) \in \Omega(g(n))$ essentially means that $f(n)$ grows *at least as fast* as $g(n)$ as n becomes large. And so in this case, $g(n)$ can be thought of as a "lower bound" on the rate of growth of $f(n)$. An example is $f(n) = n$ and $g(n) = \log n + 100\sqrt{n}$.

In the last case, $f(n) \in \Theta(g(n))$ means that $f(n)$ and $g(n)$ grow at essentially the same rate as n becomes large. An example is $f(n) = n + \sqrt{n}$ and $g(n) = 2n - 1$, as both grow linearly with n. We say that $f(n) \in \Theta(n)$ and $g(n) \in \Theta(n)$.

As illustrated above, asymptotic notation focuses on the order of growth and ignores constants. To justify the importance of focusing on the order of growth in the regime of large numbers of products, let us consider the following example.

Example 1 (Impact of linear vs. logarithmic growth). Suppose that there are two estimation methods, requiring $s_1(n) = n$ and $s_2(n) = 10 \log n$ experiments, respectively, in order to estimate an $\mathbf{A}$ matrix for n products. For a small number of products, such as $n = 10$, the first method requires just 10 experiments, whereas the second method requires $10 \log(10) \approx 23$ experiments. However, with a large number of products, such as $n = 100$, the first method now requires 100 experiments, whereas the second method requires $10 \log(100) \approx 46$ experiments, a much smaller number. As the number of products increases further, the difference between the two methods becomes more and more pronounced.

The purpose of asymptotic notation is to focus on the dominant scaling factor and ignore constants, such as 10 in method 2 of the example above. Although these constants have a relatively larger impact when n is small, they become insignificant as n becomes large. Specifically, we say that for method 1, $s_1(n) \in \Theta(n)$, and for method 2, $s_2(n) \in \Theta(\log n)$.

Appendix C

Proof of Theorem 1

Proof. Let our decisions be i.i.d. continuous random variables x distributed uniformly on $[-\rho, \rho]$, so that $\mathbb{E}[x] = 0$ and $\text{var}(x) = \mathbb{E}[x^2] = \rho^2/3$. We perform an experiment using a vector of decisions $\mathbf{x}$. Let Δq_i be the observed percentage change in demand for product i, and let x_j be the pricing decision for product j.

Having defined $\beta \triangleq 3/\rho^2$, we consider the statistic

$$y_{ij} = \beta(\Delta q_i x_j) = \beta \left(\sum_{\ell=1}^{n} a_{i\ell} x_\ell + w_i \right) x_j.$$

A simple calculation shows that it satisfies $\mathbb{E}[y_{ij}] = a_{ij}$. Therefore, y_{ij} is an unbiased estimator of a_{ij}. Let $y_{ij}(t)$ be the statistic calculated from the t^{th} experiment. By Assumption 1, for each (i, j), the statistics $y_{ij}(t)$ are independent and identically distributed across different experiments t. By the laws of large numbers, the sample mean $\hat{a}_{ij} \triangleq \frac{1}{s} \sum_{t=1}^{s} y_{ij}(t)$ converges to a_{ij} as we take many samples from many experiments. We wish to bound the concentration of $\hat{a}_{ij}$ around its mean, a_{ij}.

To do so, we show that $\hat{a}_{ij}$ is sub-Gaussian. A random variable X is sub-Gaussian with parameter $\sigma > 0$ if

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(\sigma^2 \lambda^2 / 2), \quad \forall \lambda \in \mathbb{R}. \tag{C.1}$$

We make use of the following well-known properties:

1. If X is sub-Gaussian with parameter σ, then $aX+b$ is sub-Gaussian with parameter $|a|\sigma$.

2. If X is bounded a.s. in an interval $[a,b]$, then X is sub-Gaussian with parameter at most $(b-a)/2$.

3. If X_1 and X_2 are sub-Gaussian with parameters σ_1 and σ_2, respectively,

   (a) and if X_1 and X_2 are independent, then $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2+\sigma_2^2}$.

   (b) and if X_1 and X_2 are not independent, then $X_1 + X_2$ is sub-Gaussian with parameter at most $\sqrt{2(\sigma_1^2+\sigma_2^2)}$.

4. If X is sub-Gaussian with parameter σ, then it satisfies the following concentration bound:

$$\mathbf{P}(|X-\mathbb{E}[X]| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right), \quad \forall\epsilon \geq 0. \tag{C.2}$$

We first consider the random variable y_{ij}:

$$\begin{aligned} y_{ij} &= \beta\left(\sum_{\ell=1}^{n} a_{i\ell}x_\ell + w_i\right)x_j \\ &= \beta\left\{\left(\sum_{\ell\neq j} a_{i\ell}x_\ell + w_i\right)x_j + a_{ij}x_j^2\right\} \\ &= \beta\left\{Vx_j + a_{ij}x_j^2\right\}, \end{aligned}$$

where we have defined

$$V \triangleq \sum_{\ell\neq j} a_{i\ell}x_\ell + w_i.$$

We now show that V is sub-Gaussian. For each ℓ, x_ℓ is bounded on $[-\rho,\rho]$ and therefore sub-Gaussian with parameter ρ. Hence, $a_{i\ell}x_\ell$ is sub-Gaussian with parameter $|a_{i\ell}|\rho$. Also, under Assumption 1, w_i is sub-Gaussian with parameter c. The random variables $a_{i\ell}x_\ell$ and w_i are all independent. Therefore, their sum, V, is also sub-Gaussian with parameter $\sigma_V \triangleq \sqrt{\sum_{\ell \neq j} a_{i\ell}^2 \rho^2 + c^2}$.

Next, we show that Vx_j is sub-Gaussian using the definition. For any $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\lambda(Vx_j - \mathbb{E}[Vx_j])\right\}\right] &= \mathbb{E}\left[\exp\left\{\lambda(Vx_j)\right\}\right] && \text{(C.3)}\\
&= \int_{-\rho}^{\rho} \mathbb{E}[\exp\{\lambda(Vx)\}] \frac{1}{2\rho}\, dx && \text{(C.4)}\\
&\leq \int_{-\rho}^{\rho} \exp\left\{(|x|\sigma_V)^2\lambda^2/2\right\} \frac{1}{2\rho}\, dx && \text{(C.5)}\\
&\leq \int_{-\rho}^{\rho} \exp\left\{(\rho\sigma_V)^2\lambda^2/2\right\} \frac{1}{2\rho}\, dx \\
&= \exp\left\{(\rho\sigma_V)^2\lambda^2/2\right\},
\end{aligned}$$

where (C.3) holds because Vx_j has zero mean; (C.4) is obtained by conditioning on the values of x_j; and (C.5) follows from (C.1) and the fact that for any $x \in [-\rho, \rho]$, Vx is zero-mean and sub-Gaussian with parameter $|x|\sigma_V$. Therefore, Vx_j is also sub-Gaussian with parameter $\rho\sigma_V$.

Next, we show that $a_{ij}x_j^2$ is sub-Gaussian. Since x_j^2 is bounded a.s. in $[0, \rho^2]$, it is sub-Gaussian with parameter $\rho^2/2$. Therefore, $a_{ij}x_j^2$ is sub-Gaussian with parameter $\rho^2|a_{ij}|/2$.

Finally, y_{ij} is a sum of two (dependent) sub-Gaussian random variables: βVx_j with parameter $\beta\rho\sigma_V$, and $\beta a_{ij}x_j^2$ with parameter $\beta\rho^2|a_{ij}|/2$. Therefore, y_{ij} is also sub-Gaussian with parameter

$$\begin{aligned}
\sigma_Y \triangleq \sqrt{2(\beta^2\rho^2\sigma_V^2 + \beta^2\rho^4 a_{ij}^2/4)} &= \sqrt{2\left\{\beta^2\rho^2\left(\sum_{\ell \neq j} a_{i\ell}^2\rho^2 + c^2\right) + \beta^2\rho^4 a_{ij}^2/4\right\}}\\
&\leq \sqrt{2\beta^2\rho^4\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}\\
&= \sqrt{18\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}.
\end{aligned}$$

Since $\hat{a}_{ij} = \frac{1}{s}\sum_{t=1}^{s} y_{ij}(t)$ is a sample mean of s independent y_{ij}'s, $\hat{a}_{ij}$ is sub-Gaussian with parameter

$$\sigma_{ij} \triangleq \frac{1}{s}\sqrt{s\sigma_Y^2} \leq \sqrt{\frac{18}{s} \cdot \left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}.$$

We can then bound the concentration of our estimator $\hat{a}_{ij}$ around the true parameter a_{ij} using (C.2):

$$\begin{aligned} \mathbf{P}(|\hat{a}_{ij} - a_{ij}| \geq \epsilon) &\leq 2\exp\left\{-\frac{\epsilon^2}{2\sigma_{ij}^2}\right\} \\ &\leq 2\exp\left\{-\frac{s\epsilon^2}{36\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}\right\}. \end{aligned}$$

This gives a concentration bound for the error of a particular a_{ij}. To arrive at the final result, which bounds the maximum error over all a_{ij}'s, we apply the union bound and conclude that

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon\right) \leq 2n^2 \exp\left\{-\frac{s\epsilon^2}{\max_i 36\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}\right\}.$$

□

Appendix D

Proof of Theorem 2

Proof. For any $\lambda > 0$, consider the class $\mathcal{A}_{n,k}(\lambda)$. Fix some $\epsilon \in (0, \lambda/2)$ and $\delta \in (0, 1/2)$. In what follows, all estimators use the results of s experiments, for some arbitrary s.

Define the sub-class $\mathcal{A}_{n,k}^{\text{const}}(\lambda) \triangleq \{\mathbf{A} \in \mathbb{R}^{n\times n} : |\{j : a_{ij} \neq 0\}| = k, \forall i = 1, \ldots, n; a_{ij} = \lambda, \forall i, j \text{ s.t. } a_{ij} \neq 0\} \subset \mathcal{A}_{n,k}(\lambda)$, which is the class of all $n \times n$ $\mathbf{A}$ matrices whose rows are k-sparse and whose nonzero entries are all exactly equal to λ.

The desired specification is an estimator that for any $\mathbf{A}$ matrix in $\mathcal{A}_{n,k}(\lambda)$ achieves uniformly ϵ-accurate estimates with probability $1-\delta$. In order to obtain a lower bound on the number of experiments needed to meet this specification, it suffices to obtain a lower bound on the number of experiments needed to meet the following looser specification: we let the $\mathbf{A}$ matrix be generated uniformly at random from the sub-class $\mathcal{A}_{n,k}^{\text{const}}(\lambda)$ and require that with probability at least $1-\delta$ the first row of $\mathbf{A}$ is correctly estimated to uniform ϵ-accuracy. Because $\mathbf{A} \in \mathcal{A}_{n,k}^{\text{const}}(\lambda)$, all elements of $\mathbf{A}$ are either exactly 0 or λ, and since $\epsilon \in (0, \lambda/2)$, achieving uniform ϵ-accuracy is equivalent to perfectly recovering $\mathbf{A}$, which is also equivalent to perfectly recovering the sparsity pattern of $\mathbf{A}$ (i.e., identifying the locations of all nonzero entries). Let R_1^{const} denote the event of exactly recovering the sparsity pattern of the first row of an $\mathbf{A}$ matrix chosen uniformly at random from $\mathcal{A}_{n,k}^{\text{const}}(\lambda)$.

We now focus on the event R_1^{const} and find an upper bound on its probability.

Within the sub-class $\mathcal{A}_{n,k}^{\text{const}}(\lambda)$, there are exactly $N \triangleq \binom{n}{k}$ possible sparsity patterns for the first row of any $\mathbf{A}$ matrix. Moreover, because all nonzero entries are equal to the same value λ, each unique sparsity pattern corresponds to a unique row vector, and vice versa. Suppose that we randomly choose the first row $\mathbf{a}_1^T$ by choosing one of the N possible sparsity patterns uniformly at random. We can then view the sparsity pattern recovery problem as a channel coding problem. The randomly selected sparsity pattern $\theta \in \{1, \dots, N\}$ is encoded, using a sequence of s experimental decisions $\mathbf{X} \in \mathbb{R}^{n \times s}$, into codewords $\mathbf{r} = \mathbf{a}_1^T\mathbf{X} = (r_1, r_2, \dots, r_s) \in \mathbb{R}^s$. These codewords represent the uncorrupted percentage changes in demand for product 1 in each of the s experiments. The codewords are sent over a Gaussian channel subject to noise $\mathbf{w} = (w_1, w_2, \dots, w_s) \sim \mathcal{N}(0, c^2 I)$ and received as noisy measurements $\mathbf{y} = \mathbf{r} + \mathbf{w} = (y_1, y_2, \dots, y_s) \in \mathbb{R}^s$, which are equal to the observed noisy percentage changes in demand, $\Delta\mathbf{q}_1$. The goal is to recover the pattern θ from the measurements $\mathbf{y}$.

The power of a Gaussian channel is given by $P = \frac{1}{s}\sum_{t=1}^{s} r_t^2$. Since $\mathbf{a}_1^T$ is k-sparse and any decision x is bounded in $[-1, 1]$, we have that $|r_t| \leq k\lambda$ for all t, and hence $P \leq k^2\lambda^2$. From standard results (Cover and Thomas 1991), the capacity of a Gaussian channel with power P and noise variance c^2 is $\frac{1}{2}\log\left(1 + \frac{P}{c^2}\right)$. Therefore, the capacity of our particular channel is

$$C \leq \frac{1}{2}\log\left(1 + \frac{k^2\lambda^2}{c^2}\right).$$

From Fano's inequality (Cover and Thomas 1991), we know that the probability of error, P_e, of a decoder that decodes the sparsity pattern θ from noisy measurements $\mathbf{y}$

is lower bounded as

$$\begin{aligned} P_e &\geq \frac{H(\theta \mid \mathbf{y}) - 1}{\log N} \\ &= \frac{H(\theta) - I(\theta; \mathbf{y}) - 1}{\log N} \\ &= \frac{\log N - I(\theta; \mathbf{y}) - 1}{\log N} \\ &= 1 - \frac{I(\theta; \mathbf{y}) + 1}{\log N}, \end{aligned}$$

where H denotes entropy and I denotes mutual information. The first equality is by the definition of mutual information, and the second equality follows from the fact that θ is chosen uniformly over a set of cardinality N. We can upper bound the mutual information between θ and $\mathbf{y}$ as

$$\begin{aligned} I(\theta; \mathbf{y}) &\leq I(\mathbf{r}; \mathbf{y}) && \text{(D.1)} \\ &= h(\mathbf{y}) - h(\mathbf{y} \mid \mathbf{r}) \\ &= h(\mathbf{y}) - h(\mathbf{w}) \\ &\leq \sum_{t=1}^{s} h(y_t) - \sum_{t=1}^{s} h(w_t) && \text{(D.2)} \\ &= \sum_{t=1}^{s} [h(y_t) - h(y_t \mid r_t)] \\ &= \sum_{t=1}^{s} I(r_t; y_t) \\ &\leq sC, && \text{(D.3)} \end{aligned}$$

where h denotes differential entropy, (D.1) follows from the data processing inequality, (D.2) follows from the independence of the w_t's and the fact that the entropy of a collection of random variables $\{y_t\}$ is no more than the sum of their individual entropies, and (D.3) follows from the definition of channel capacity as the maximal mutual information. And so by Fano's inequality, the probability of error is lower bounded by

$$P_e \geq 1 - \frac{sC + 1}{\log N},$$

which immediately gives the following upper bound on the probability of R_1^{const}:

$$\mathbf{P}(R_1^{\text{const}}) = 1 - P_e \leq \frac{sC + 1}{\log N}.$$

Therefore, achieving the looser specification of uniform ϵ-accurate estimates of the first row of a random $\mathbf{A} \in \mathcal{A}_{n,k}^{\text{const}}(\lambda)$ with probability $1-\delta$ implies the following condition on the number of experiments, s:

$$1 - \delta \leq \frac{sC + 1}{\log N} \Longrightarrow s \geq \frac{(1-\delta)\log N - 1}{C}.$$

Consequently, achieving the stricter original specification of an estimator that for all $\mathbf{A}$ matrices in $\mathcal{A}_{n,k}(\lambda)$ achieves uniformly ϵ-accurate estimates with probability $1 - \delta$ also requires the number of experiments to satisfy the above condition.

With some simple rearrangement, and noting that $\log N = \log \binom{n}{k} \geq k \log(n/k)$ and $\delta \in (0, 1/2)$, we obtain the desired lower bound:

$$s \geq \frac{(1-\delta)\log N - 1}{C} \geq \frac{2(1-\delta)k\log(n/k) - 2}{\log(1 + k^2\lambda^2/c^2)} \geq \frac{k\log(n/k) - 2}{\log(1 + k^2\lambda^2/c^2)}.$$

□