Learning the Structure of Price Stickiness in Scanner Data

Amy H. Hua

Advised by Professor Christopher A. Sims

Department of Economics

Princeton University

April 12, 2017

A senior thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Arts in Economics at Princeton University

This thesis represents my own work in accordance with University regulations.

Amy H. Hua

∼ Amy H. Hua

Abstract

A substantial portion of the macroeconomics literature suggests that demand shocks have noteworthy effects on real output. One prevailing theory for rationalizing the large effects of demand shocks on output is that nominal prices are not perfectly flexible. Our work aims to extend the empirical literature studying price stickiness by using models and techniques from Bayesian statistics and machine learning. We analyze a high-frequency scanner dataset and focus on algorithmically identifying price setting periods for any given product and characterizing a set of regular prices within price setting periods, given only the values of the price time series. Our work departs from the existing literature in three important ways. For each product, we flexibly identify price setting periods without making any assumptions on the length of the periods or the number of periods. Additionally, we do not simply study one regular, or reference, price. Instead, we identify several "modal" prices from the data that correspond to the peaks of the price distributions for the identified price setting periods. Finally, we develop highly clusterable and interpretable metrics for price stickiness, which we demonstrate with an easily separable clustering of the products into two clean groups of highly sticky and non-sticky products.

Acknowledgements

First and foremost, I would like to thank my advisor Professor Christopher Sims for being the most thoughtful and dedicated advisor that I could have asked for. Thank you for generously lending your time and insights throughout the year and for guiding me throughout the entire research process. It is an honor to have had the opportunity to learn from you.

I am also grateful to the professors of whom I have been privileged to be a student, including Professors Bo Honoré and Han Liu, who introduced me to the world of econometrics and machine learning; Professor Ilyana Kuziemko for providing crucial advice throughout my first ventures into research; Professors Benjamin Moll, Peter Ramadge, and Emmanuel Abbe, from whom I learned the tools with which I tackled this thesis; and Professor Harvey Rosen for inspiring me to become an economics major in the first place.

Thank you to my dearest friends who have stood by my side all of these years. You each mean so much to me, and I am excited to continue growing alongside you. From my incredible roommates and pset buddies to the amazing women surrounding and inspiring me: I am beyond proud to graduate with you.

Above all, I am fortunate to have the most supportive family in my mom, dad, and brother, behind me every step of the way. I would not be here without your confidence in me and your unwavering dedication.

Contents

List of Figures

Chapter 1

Introduction

One of the major questions that macroeconomists consider is understanding the effects of demand shocks on the economy. Under the assumptions made in a large number of macroeconomic models, if the economy is able to swiftly reach an equilibrium in response to shocks, thus responding efficiently, then demand shocks should produce minimal effects on output. However, a large portion of the empirical literature generally finds evidence that significant monetary non-neutrality ensues in response to monetary shocks (Christiano, Eichenbaum, and Evans 1999 and Romer and Romer 2004) and that shocks to government spending increase aggregate output significantly (Ramey 2011 and Nakamura and Steinsson 2014).

A prevailing theory for rationalizing the large effects of demand shocks on output is that nominal prices are not perfectly flexible; rather, they adjust slowly to variations in money supply, which effectually allows aggregate output to deviate significantly from efficient levels. Assessing the adjustment speed of the aggregate price level to exogenous shocks is thus critical to understanding the effects of demand shocks on output. While one major problem faced by researchers is how to identify exogenous demand shocks in data, what economists can do is collect evidence on price stickiness at the micro price level. The empirical objective, then, is twofold: research in this area works on identifying and characterizing rigid price adjustment in the economy and, more fundamentally, understanding the important dynamics that drive the existence of sticky prices in the first place (Nakamura and Steinsson 2013).

In the past, empirical research on price dynamics was limited in scope due to the lack of micro data on prices; however, in recent decades, data with individual prices have become more readily available, allowing a more wide-ranging examination of individual price settings. The disaggregated nature of these data poses new issues, including the treatment of temporary price changes, seasonality trends, cross-sectional heterogeneity, and interactions between retailers and manufacturers, that, if not accounted for, skew analysis of the aggregate price level. Thus, much of the empirical research on price stickiness has focused on the treatment of these issues and on determining which dynamics governing price trends are the most significant in measuring the rigidity of aggregate prices. Stickiness measures originally focused on the extent to which prices varied, namely in terms of the frequency of price changes, but a key modern idea is the notion of "regular prices," which generally refers to the standard price of a good during a given time interval. For specific intervals and corresponding regular prices, several measures of stickiness can be constructed and used in the calibration of macroeconomic pricing models such as multisector models (Bils and Klenow 2004 and Carvalho 2006), the Calvo model (Kehoe and Midrigan 2015), and the menu-cost model (Eichenbaum et al. 2011 and Kehoe and Midrigan 2015). Then, the validity of the stickiness measures is analyzed by comparing model outputs to the real economy.

Analyzing a scanner dataset, Eichenbaum et al. (2011) offer a new approach to identifying regular prices by defining "reference prices" to be the most common price within a fixed quarter of time. Eichenbaum et al. (2011) demonstrate that reference prices are a good measure of price stickiness across quarters and, in fact, reveal more about price rigidities than standard frequency measures of price change, since reference prices are quite inertial, with an average duration of roughly one year, despite the fact that weekly prices shift much more frequently. In order to also understand price dynamics within quarters, Eichenbaum et al. (2011) estimate a two-state Markov model for the weekly prices within a given quarter, where the two states are given by a price being equal to the reference price and the price not being equal to the reference price. The resulting transition matrices for all products are estimated from data averaged across categories of goods.

We take a closer examination of four components governing the structure of Eichenbaum et al. (2011)'s contributions. First, they examine price time series broken into fixed time intervals (quarters), making a defining assumption that firms make decisions on a quarterly basis. Next, they associate with each of these time intervals a measure of the regular price throughout the interval (the reference price). Thirdly, they calculate a price stickiness measure for product categories across the intervals (frequency of reference price changes). Finally, they analyze within-interval price stickiness by calculating a stickiness measure for each quarter (namely, the transition matrix of the Markov model and the frequency and duration of weekly price changes).

Using a similar dataset to the one studied by Eichenbaum et al. (2011), we extend the study of price stickiness further by building upon these existing methods, with the objective of understanding the fundamental trends that occur in these micro price data. In particular, several modern techniques from statistics and machine learning are useful for analyzing this dataset and providing insights. Importantly, we are consequently able to separate ourselves from presumptions regarding price dynamics to focus on extracting insights from the data. Our ultimate goal is to use these findings to update the existing approach toward empirical pricing analysis in economic literature.

Identifying Price Setting Periods

We extend the work of Eichenbaum et al. (2011) by developing new methods to algorithmically detect the periods of time over which prices are set for each individual product, which frees us from making any assumption of fixed quarters. The fact that we can learn the periods of time over which firms make price decisions also allows us to bypass having to make assumptions about the costs of changing prices (as in the menu-cost model) or any notions of randomized ability to change prices (the Calvo model). Furthermore, learning price setting periods allows for the possibility of further investigating price variation within price setting periods. For instance, we could ask whether different classes of products go on sale at different times. By looking at the broad trends outlined by the standard prices for different price setting periods, we can also investigate product properties like seasonality.

Finding a Set of Regular Prices

In the previous section, we outlined our goal of automatically identifying price setting periods. Our objective now is to detect a set of interesting prices *within* the identified price setting periods. We justify our separation of these two goals since regular price decisions and sale price decisions may be made orthogonally to each other (Anderson et al. 2016). Eichenbaum et al. (2011) suggest that some non-reference prices might not correspond to sale prices, but their model does not explain what these non-regular prices might be nor further try to distinguish between them; thus, our work extends the depth of their analysis.

Our approach is to, given a price setting period, identify price clusters with significant prices to capture the most frequently visited, and thus likely to be the most important, prices. We achieve this goal by applying a Gaussian mixture model within a Bayesian framework to identify the interesting prices within a price setting period.

Metrics of Price Stickiness

For any given product, we want to be able to capture two sets of measures of how sticky its price trend is: (1) how frequently and how much the regular prices fluctuate from one price setting period to another and (2) how much prices within price setting periods fluctuate. The first set of measures are comparable with the existing measures of frequencies in price changes and durations of rigidity used commonly throughout the price stickiness literature (Nakamura and Steinsson 2013), though we make some adjustments. The second set of price stickiness measures extends the simple two-state Markov model that Eichenbaum et al. (2011) use to characterize the price stickiness of all products. We use a larger transition matrix, taking into account the entire set of interesting regular prices, and we also use the entropy rate of the transition matrix to characterize overall stickiness with a single number. Our measures of stickiness are also able to decompose the price stickiness of product price trends into a set of measures corresponding to the kinds of interesting regular prices. Consequently, we provide a more thorough characterization of what it means for a product to be sticky.

Price-Cost Relations

We fit a ridge regression model to estimate a linear relationship between cost data and price data. The point of this analysis is to identify the links between costs and prices, as previously examined by Eichenbaum et al. (2011).

Clustering

Eichenbaum et al. (2011)'s analysis focuses on the aggregate nature of the product categories found in their dataset. We would like to consider more fine-grained patterns at a product level. Given the price setting periods, the set of significant regular prices for each price setting period, and the stickiness measures for each product (both across and within price setting periods), how can we identify patterns across the products? Our hypothesis is that there is a set of common types of trends that can be found in our data, and the goal is to identify these shared dynamics. It turns out that our models are able to identify fine-grained clusters of the dynamics of price setting periods. Furthermore, our measures of price stickiness are able to stratify products into categories corresponding to various degrees of stickiness.

Organization

The remainder of this paper proceeds as follows. Chapter 2 reviews the empirical literature on price stickiness and, in particular, research analyzing scanner data. Chapter 3 summarizes the dataset we use, preprocessing steps we take, and results from exploratory data analysis. Next, Chapter 4 provides an overview of the methods that we use to explore price dynamics. Chapter 5 provides an overview and interpretation of the results from our analysis, and finally, Chapter 6 concludes with some remarks on possible future directions for research on this scanner dataset and price stickiness in general. Included at the end of the paper is an appendix that goes into detail about some of the theory behind the methodology. The code for this paper can be found at `https://github.com/amyhhua/princeton-thesis`.

Chapter 2

Background

Until recently, empirical research on price rigidity was limited in quantitative scope as large micro price datasets were difficult to secure. However, new sources of price data containing large amounts of information on individual products are becoming more readily available to researchers. In particular, scanner data have provided new opportunities to extract insights from high-frequency price dynamics for a wide-ranging array of consumer goods. These datasets bring several new issues into consideration that were not as relevant in the previous era of empirical research on parsimonious macroeconomic pricing models. We can roughly categorize the empirical literature into studying price stickiness and the treatment of the following key issues: temporary price changes such as sales; regular price changes, including reference prices; small price changes in scanner data that might not represent price changes due to discounts; seasonality trends; cross-sectional heterogeneity across products; and regression analysis of costs and prices. Nakamura and Steinsson (2013) provide a comprehensive overview of the literature on price stickiness that we draw upon throughout this section.

2.1 Temporary Price Changes

One research area of significance concerns the treatment of temporary sales and other short-term price fluctuations. The increased accessibility to scanner data over the past few decades has introduced new opportunities to study demand and competitive strategy models. As retailers have begun to analyze scanner data more, some pricing power potentially has shifted from manufacturers to retailers. Dutta et al. (2002), for instance, analyze price flexibility using a dataset with actual retail transaction prices, wholesale transaction prices, and a measure of manufacturers' costs. Their results indicate that retail prices respond not only to their direct costs, but also to the upstream manufacturer costs. In particular, temporary retail store promotions have been shown to greatly impact brand and store substitution (Kumar and Leone 1988 and Walters 1991).

Measuring price stickiness can be highly sensitive to how these temporary sales and other short-term price fluctuations are treated. The median implied duration for all posted prices is roughly half of that for regular prices in isolation (Nakamura and Steinsson 2013), so the empirical literature considers the decision regarding whether sales should be excluded when measuring price stickiness to be critical.

Before we can address how to treat sales in constructing a measure of price stickiness, we must first be able to accurately identify sales, even if price data are not labeled with sale markers. If no such indicators to identify sales are included in the data, researchers often turn to other methods of detecting temporary prices, such as asserting that sale prices have a particular shape. A common identifier of sales is V-shaped temporary discounts, or large temporary drops in prices followed by prices returning exactly to their former levels, with only few other types of price changes (Campbell and Eden 2014 and Guimaraes and Sheedy 2011). One major implication of this trend is that though a product might experience high-frequency price flexibility, it can simultaneously exhibit a low-frequency price stickiness in the regular price. Kehoe and Midrigan (2015) develop a model that attempts to extend the Calvo model and the standard menu cost model by allowing firms to temporarily deviate from a sticky preexisting price.

However, not all datasets contain such well-defined sale price behavior, and constructing a dependable algorithm for identifying sales has proven challenging for researchers. Other authors (Nakamura and Steinsson 2008 and Kehoe and Midrigan 2015, for instance) consider more complex sale filter algorithms that encode modifications to the basic notion of a sale filter. Kehoe and Midrigan (2015)'s algorithm categorizes each price shift as either temporary or regular, based on each price's relative position to the mode price over a given window of time within a particular time series; notably, the algorithm differentiates between price increases and price decreases. The filter used by Eichenbaum et al. (2011) sets the non-sale price for a given product to be the most commonly observed price in a particular quarter.

Chahrour (2011) compares many of these different filters applied to the same dataset—the Dominick's Finer Food dataset—and demonstrates that though several pricing facts are robust to the type of sale filter or identification algorithm used, implications for price stickiness depend on filter specification. He proposes an alternative sale filter that determines non-sale prices based on a 13-week (plus or minus 6 weeks) window, instead of fixed, non-overlapping periods. Chahrou finds two major trends in temporary price changes: he captures large decreases in prices, as the other papers also accomplished, but additionally, he finds small increases in price that are far less common and significantly smaller than sales. We return to such small fluctuations in prices later in this chapter. In general, however, we recognize the necessity of developing a sound algorithm that identifies price fluctuations with high accuracy, to which we return in our methodology chapter.

2.2 Regular Price Changes

In the price stickiness literature, there is an ongoing debate on whether temporary prices and regular prices should be treated symmetrically. One advantage of working with regular prices, for the purpose of estimating stickiness across time, is that we do not need to form prior assumptions on whether sales are special events that should be treated differently.

Another rationalization of using regular prices is that sets of regular prices make sense in the context of a menu cost model, where firms can choose a pricing setting from a small, fixed menu of prices between which they are free to move. Furthermore, Kehoe and Midrigan (2015) find that regular price changes contribute to aggregate price level considerably more than temporary prices. Finally, menu-cost models may be a decent way to think about regular price changes (though not temporary price changes). In fact, the processes for deciding regular and temporary prices may be orthogonal (Eichenbaum et al. 2011 and Anderson et al. 2016).

2.2.1 Reference Prices

Eichenbaum et al. (2011) define a "reference" price to be the most common regular price found in a fixed window of time (a quarter). Thus, they group all prices into two clusters: (1) reference prices (defined for a certain quarter) and (2) non-reference prices. They find evidence that while actual prices are quite flexible, experiencing a median duration of three weeks, nominal rigidities take the form of reference prices that experience significantly more inertia with a duration of nearly two quarters.

One important observation that the researchers make concerns their assumption of fixed intervals of quarters, over which they define each quarter's reference price. They find that non-reference prices do not necessarily correspond to sale prices, and in fact, 21 percent of non-reference prices are higher than the corresponding reference price.

2.2.2 Markov Model

In order to understand price dynamics across weeks within quarters, Eichenbaum et al. (2011) use a single two-state Markov process to estimate the likelihoods of products' prices returning to and deviating from their reference prices, a modal measurement of price adjustments. A Markov model is a linear, memoryless process that describes how a particular quantity transitions between various states. At each point in time, a price can either be the reference price or a "non-reference" price, and the model estimates the probability distribution over these two states. By examining the resulting transition matrix (describing probabilistic transitions between states), Eichenbaum et al. (2011) conclude that prices are sticky. Notably, they estimate the transition matrix for each item in every quarter and then average these transition matrices over all quarters. With this average in hand, they then compute the average transition matrix for items within categories and, lastly, compute the average transition matrix across categories. Specifically, by examining the first diagonal entry of the average transition matrix (the probability that a reference price state will return to itself), the authors of Eichenbaum et al. (2011) claim price stickiness.

2.3 Small Price Changes and Scanner Data

An important finding in the empirical literature over the past decade is that firms often appear to make small price changes (Klenow and Kryvtsov 2008, Midrigan 2011, and Bhattarai and Schoenle 2014). Eichenbaum et al. (2014) use a scanner data similar to our dataset and to the dataset that they used in Eichenbaum et al. (2011)—though their data span only one year and a few states in the U.S.—and micro data from the Consumer Price Index to measure the prevalence of small price changes. They propose that the majority of small price changes are due to a type of measurement error that arises from the lack of explicitly recorded prices in many scanner data and the subsequent use of price measures constructed as the ratio of sales revenue from a product to the quantity sold of that product. Indeed, suppose that a range of consumers buy the same good at different prices; a small change in the breakdown of consumers could lead to a misleading small price change, though the true price might not have changed. This measurement error is particularly pronounced with supermarket transactions since some products are sold at a discount to customers with loyalty cards, and some products are discounted with coupons or qualify for specific promotions.

Other scanner datasets come with labeled prices, in the sense that each price is marked as a regular price or as a sale price (Anderson et al. 2016). The nice property that these datasets have is that they have a ground truth that is discernable independent of noise.

Our scanner dataset does not contain exact pricing information and also does not contain price labels, which makes our task more difficult: we must sift through the noise and find signal. Eichenbaum et al. (2014) study price changes that are smaller, in absolute terms, than 1, 2.5, and 5 percent and find that their conclusions hold irrespective of which of these values is used to define a small price change; thus, we assume that there is not too much noise in the scanner data.

2.4 Seasonality Trends

While there is considerable evidence of some price changes—such as sales for particular products—following a regular pricing schedule, micro data often exhibit a significant amount of seasonality. Chevalier et al. (2003) examine a dataset with retail and wholesale prices covering over seven years, finding that, on average, prices fall during peaks in seasonal demand for a given product, due in large part to decreases in retail margins, which is consistent with "loss-leader" pricing models of retailer competition. Furthermore, trends of seasonality motivated by factors including weather conditions, production cycles, and holidays can affect price shift distributions.

Nakamura and Steinsson (2008) also find notable seasonality trends in the United States: for consumer prices, they observe a median frequency of regular price change of 11.1% in the first quarter, which drops monotonically to 8.4% in the fourth quarter, and for producer prices, the median frequency of price change in the first quarter is 15.9%, which falls by an even greater magnitude to 8.2% by the fourth quarter. This demonstration of seasonality in price variations indicates that monetary non-neutrality might be more significant for shocks that occur earlier in the year rather than later.

2.5 Cross-Sectional Heterogeneity

There is a significant amount of heterogeneity across sectors of consumer goods. Nakamura and Steinsson (2013) generate a histogram of the frequencies of non-sale price changes across different product categories in the Consumer Price Index (CPI). They find that the mean frequency of non-sale price changes is roughly double the weighted median frequency of non-sale price changes across sectors.

Bils and Klenow (2004), among the earlier work analyzing the newly-available micro data that catalyzed a shift in the conventional belief of the twentieth century that prices changed roughly once a year (Blinder 1998), study a component of the U.S. Consumer Price Index (CPI) with 350 categories of goods and services covering roughly 70 percent of consumer spending. Bils and Klenow (2004) find a median duration of price stickiness of merely 4.3 months and, importantly, find evidence that the frequency of price variations differs wildly across different goods. Carvalho (2006) develops a model that introduces sectoral heterogeneity in terms of nominal rigidities into an otherwise standard sticky price model. He finds that heterogeneity in price stickiness amplifies both the magnitude and the persistence of the real effects resulting from monetary shocks, as compared with parallel effects in identical-firm economies with more homogeneous sectors in terms of price rigidities. Ceteris paribus, sectors in which prices change frequently should experience a swifter response of inflation, as compared to sectors with stickier prices, in response to an expansionary demand shock. Since cross-sectional heterogeneity in the frequency of price variations across sectors has testable implications regarding price changes and relative inflation rates across sectors, the cross-sectional distribution of sectoral relative adjustment prices merits further empirical analysis.

2.6 Cost-Price Regression Analysis

Eichenbaum et al. (2011) study the relations between reference prices and reference costs, finding a distinctive form of state dependence in reference prices—namely, that the duration of reference prices is chosen by the retailer to limit markup variation. Moreover, they find that prices do not change unless there is a motivating change in costs, and they do not discover a significant lag or lead in the relation between cost changes and price changes. Other papers have also explored the relations between producer prices and consumer prices (Nakamura and Zerom 2009 and Anderson et al. 2016) and generally find swift pass-through of changes in retail costs to the prices of the goods.

Chapter 3

Data

3.1 Scanner Dataset

We use a similar dataset to the scanner dataset used by Eichenbaum et al. (2011). The scanner data that we analyze is well-suited for our research due to its high frequency and information it includes about quantities sold and wholesale prices that cannot be found in all scanner datasets. Specifically, we analyze scanner data provided by a large grocery store retailer operating in nearly 2,200 stores, mostly across the United States but also in Canada. We focus on the U.S. dataset, which contains about 2.5 million transactions for each week from the beginning of 2004 to mid-2007, with over a million unique products covered and about 50,000 unique products sold weekly.

3.2 Transactions in the Dataset

We define a valid transaction to be a transaction involving both a positive amount of gross sales and a positive quantity sold of the product. In particular, transactions that are returns or sales in non-positive quantities, which constitute less than 0.5 percent of the total set of transactions, are excluded from our analysis in this paper.

Figure 3.1 illustrates the distribution of the number of weeks that products are sold. The distribution implies two significant clusters of products: products that are consumed at least every week throughout the duration of the dataset and products that are purchased very sparsely over this timeline. While we will focus our work on the first cluster, there may be a different set of price dynamics governing each of these two groups, which may be an interesting area to explore for future work.

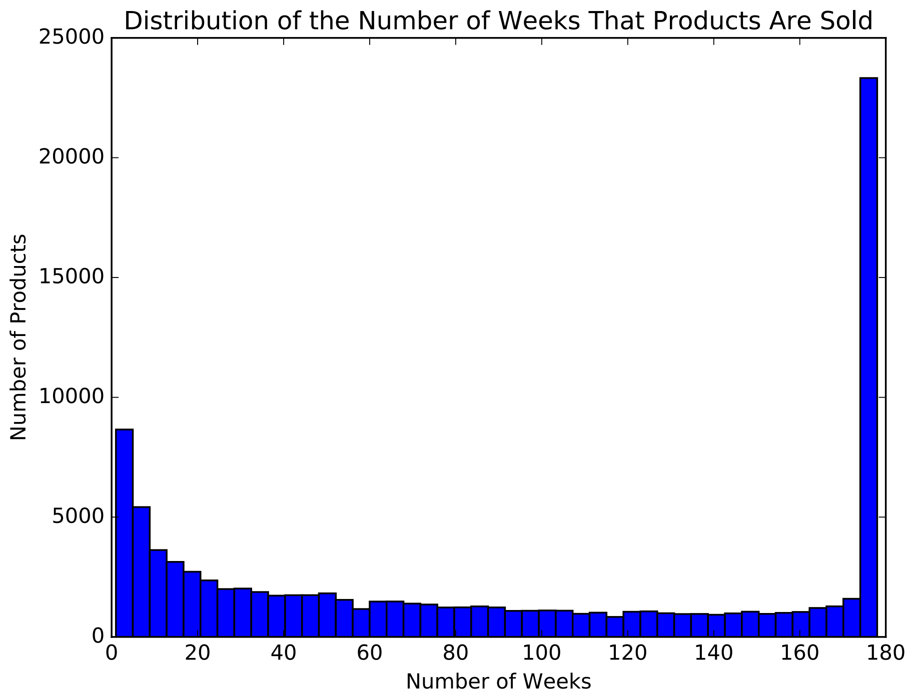


Figure 3.1: Distribution of the number of weeks that products are sold.

For the remainder of this paper, we apply the following filter for transactions that we analyze: we include only products that are sold in at least one valid transaction for every week throughout the entire dataset's timeline. This selection of products still consists of over 18,500 unique products, and restricted to this subset of products, the dataset contains approximately 1.6 million valid weekly transactions for each week. We also select this group of products because we are interested in duration measures, in addition to frequency measures, characterizing price stickiness. Since our products show up every week, we do not need to account for issues stemming from product dropout, as suggested by Nakamura and Steinsson (2014).

Figure 3.2 illustrates the volume of transactions in which each product is involved.

A significant majority of products partake in high numbers of transactions throughout the dataset.

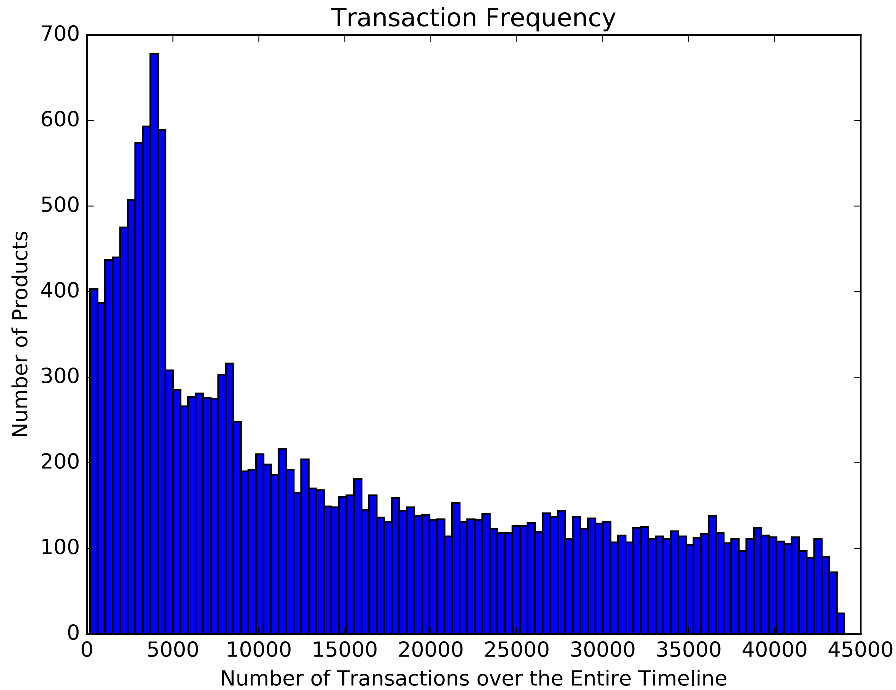


Figure 3.2: Distribution of the volume of transactions in which products are involved, taken across the entire timeline of the dataset.

3.3 Price Indices

To explore how an aggregate price measure of the goods in our dataset changes over time, we examine several constructions of a mean price over all transactions or goods. We first track the Laspeyres and Paasche price indices, defined, respectively, as

$$\begin{aligned} P_{\text{Laspeyres}}(t) &= \frac{\sum_i p_{i,t} \cdot q_{i,0}}{\sum_i p_{i,0} \cdot q_{i,0}}, \\ P_{\text{Paasche}}(t) &= \frac{\sum_i p_{i,t} \cdot q_{i,t}}{\sum_i p_{i,0} \cdot q_{i,t}}, \end{aligned} \tag{3.1}$$

where i indexes products and t indexes weeks. As a result of potential item substitution bias, the Laspeyres price index tends to overstate the price level, while the Paasche price index tends to understate the price level. Thus, we also analyze the Fisher price index, defined as the geometric mean of the Laspeyres and Paasche indices:

$$P_{\text{Fisher}}(t) = \sqrt{P_{\text{Laspeyres}}(t) \cdot P_{\text{Paasche}}(t)}. \tag{3.2}$$

Figures 3.3, 3.4, and 3.5 illustrate how the Laspeyres price index, the Paasche price index, and the Fisher price index, respectively, change across the timeline of the dataset. Not surprisingly, each of these price indices displays a gradual upward trend across time, with the exception of a few spikes in the price indices, particularly the Paasche index, which generally occur right after the start of a new year in the timeline. We can also interpret these figures in terms of seasonality. We note that there are drops in January of each year, which could be explained by the possibility that wages are more likely to change in January rather than in other months of the year—especially since we see these drops in aggregate price levels and not for any specific kind of product (Nakamura and Steinsson 2014).

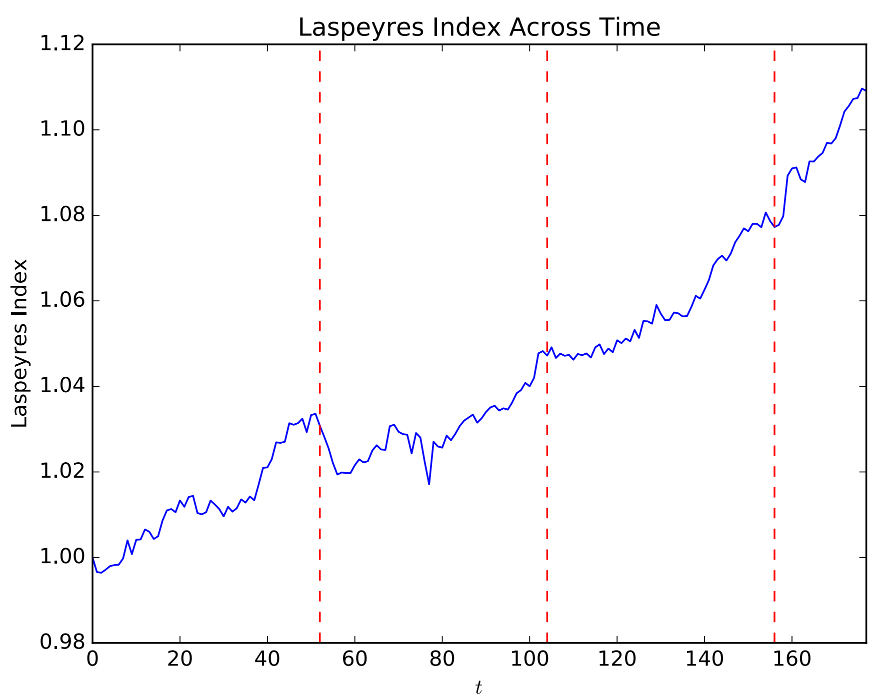


Figure 3.3: Plot of the Laspeyres price index over time. The dashed red lines mark the first week of a new year in the timeline.

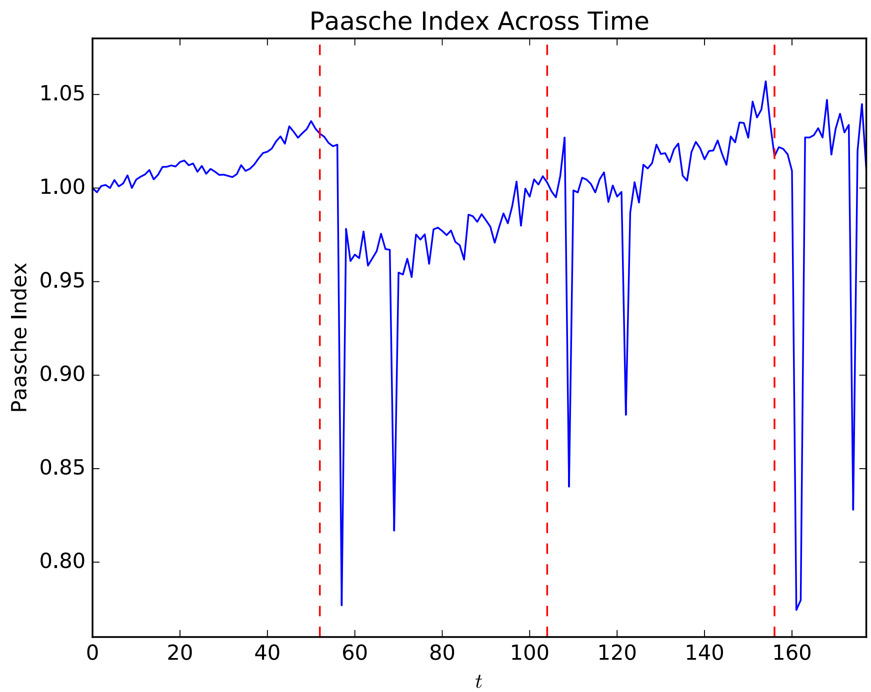


Figure 3.4: Plot of the Paasche price index over time. The dashed red lines mark the first week of a new year in the timeline.

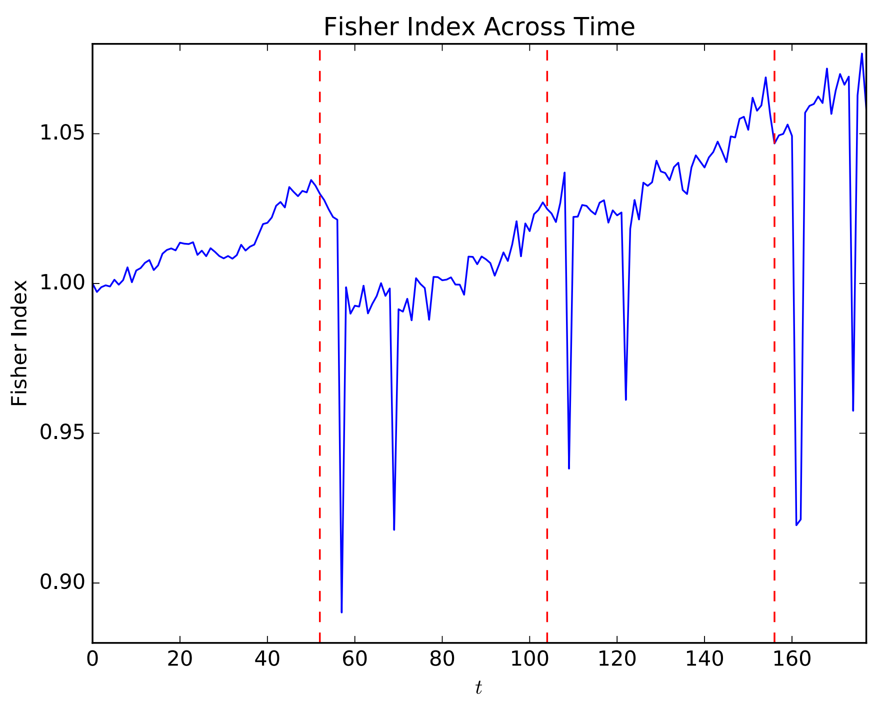


Figure 3.5: Plot of the Fisher price index over time. The dashed red lines mark the first week of a new year in the timeline.

3.4 Preprocessing

3.4.1 Standardization of Data

For each product's trend of prices and trend of costs, we calculate the mean and the standard deviation of each trend. We then "z-score" the data by subtracting the mean from each point and dividing each point by the standard deviation. This preprocessing step has the effect of setting the mean of the empirical distribution to be 0 and the standard deviation to be 1. Thus, we are now able to compare products' trends on the same scale. As shown in Figure 3.6, most of the standard deviations of both raw price trends and raw cost trends are less than 1. Thus, a potential downside of this preprocessing is that small deviations are amplified if the standard deviation of the raw trend is very small. For instance, consider a trend where all values are the same except one point, which is only slightly different from the rest of the points. Our standardization will exaggerate this difference. This property may not be desirable if the deviation was due to noise. However, if we assume that small deviations are not significantly due to noise, then standardization avoids the problem of mistaking true deviations for noise solely due to scale issues.

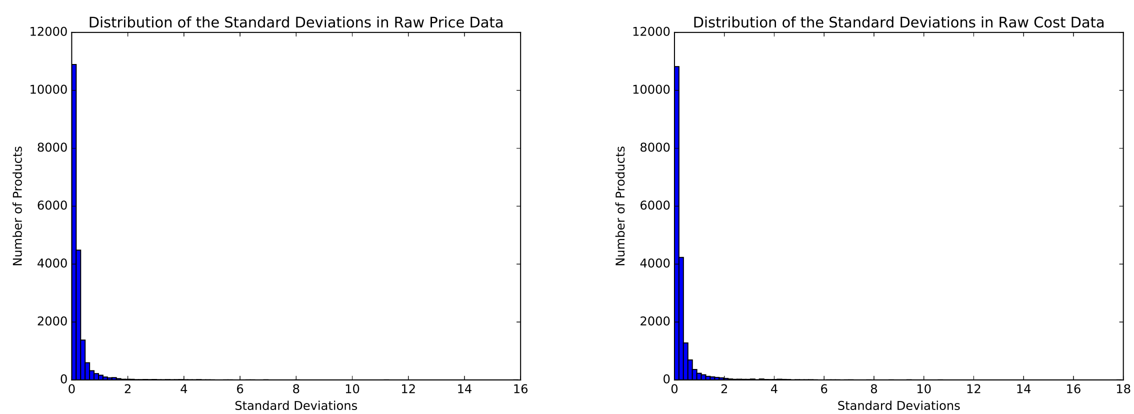


Figure 3.6: Distributions of the standard deviations in raw price and cost trends across all products. Notably, the standard deviations are generally less than 1.

3.4.2 Calculating Prices and Costs

Since our dataset does not contain direct information on prices or costs, we construct a measure of prices and costs based on the following formulas, for product i and week t.

$$\begin{aligned}\text{Price}_{i,t} &= \frac{\text{Total revenue involving product } i \text{ in week } t}{\text{Total quantities sold of product } i \text{ in week } t}. \\ \text{Cost}_{i,t} &= \frac{\text{Total wholesale costs for product } i \text{ in week } t}{\text{Total quantities sold of product } i \text{ in week } t}.\end{aligned} \tag{3.3}$$

3.5 Summary Statistics

3.5.1 Price Trends

Unprocessed Modes

Figure 3.7 displays the distribution of the number of weeks that the mode of price trends shows up. We observe that the actual mode does not appear to capture the notion of repeated regular prices, as outlined in the literature described in Chapter 2.

Mean Prices

Figure 3.8 illustrates the distribution of mean prices of products, where the mean price of each product is calculated by aggregating the gross sales and quantities sold of all transactions involving the product over all weeks.

Correlations between Products using Price Vectors

Given two products' price time series, where each point in a product's time series is the average price of the product across all transactions involving the product during a particular week, how can we measure the similarity between these two price data vectors? In the following analysis, we use the pairwise Pearson correlation metric to measure the correlation between two particular products' price time series, where

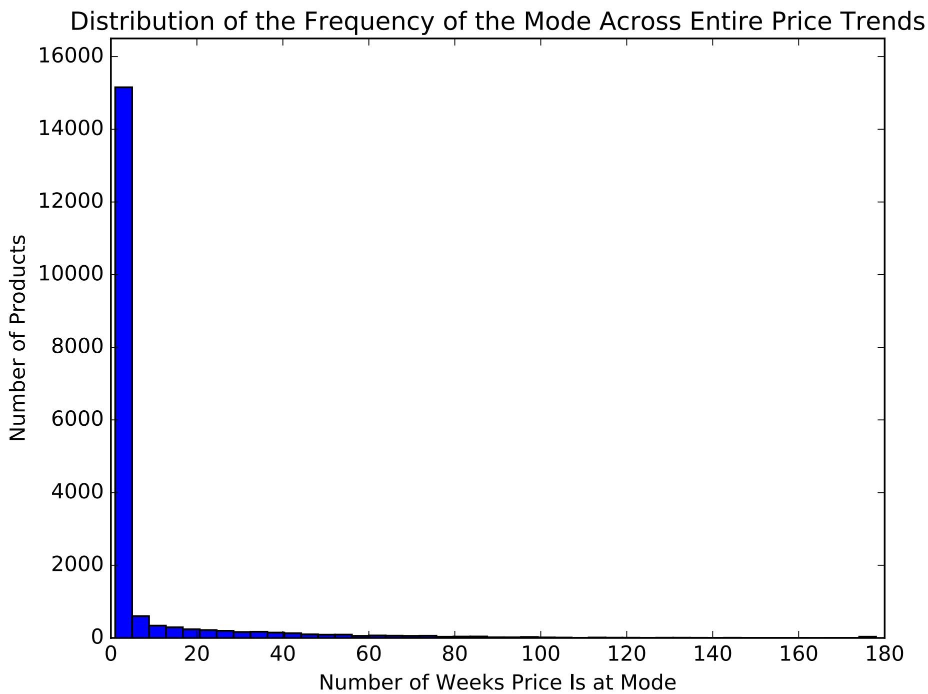


Figure 3.7: Distribution of the frequency of the mode across price trends for all products, where the frequency of the mode is defined by the number of weeks that the price equals the mode.

$\rho_{X,Y}$, the Pearson correlation coefficient between two vectors X and Y, is defined as

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{3.4}$$

We compute the pairwise Pearson correlation coefficient for pairs of products drawn from a set of 1,000 random products. As shown in Figure 3.9, a fair number of pairs of products are positively correlated, while fewer pairs of products are negative correlated. Of particular interest are the pairs of products with correlations beyond some threshold γ (for instance, product pairs (i, j) such that $\left|\rho_{\{t_i\},\{t_j\}}\right| \geq \gamma$), which might aid in clustering products that are not obviously related but share highly similar or dissimilar price trends.

Figure 3.9 illustrates the correlation values $c(i, j)$ between pairs (i, j) of 1,000 random products. Before calculating the correlation values, we first sort the products by the category to which they belong, with the hope that products belonging to the same category will have similar correlation metrics. We note an interesting block-like

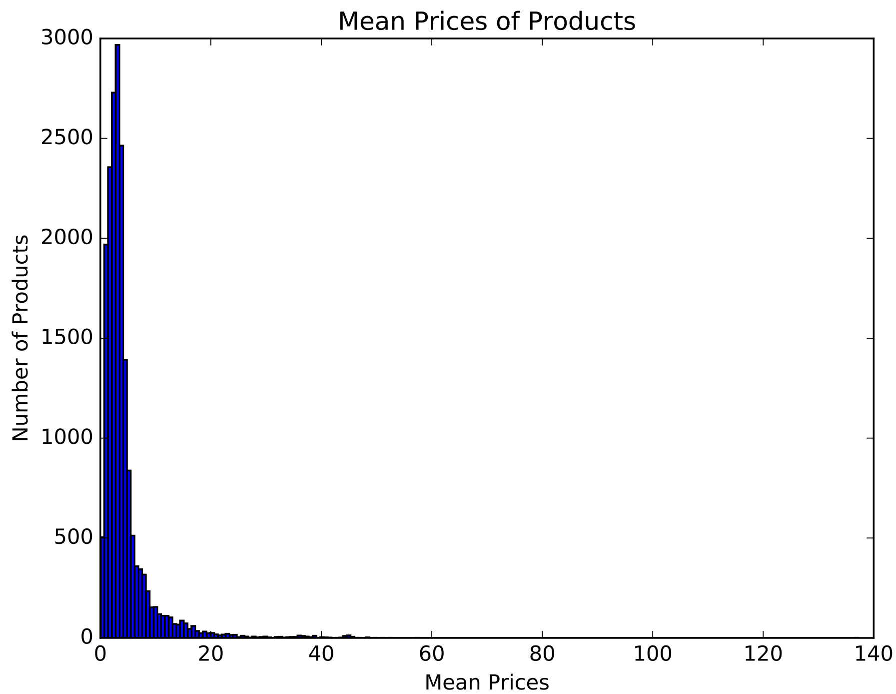
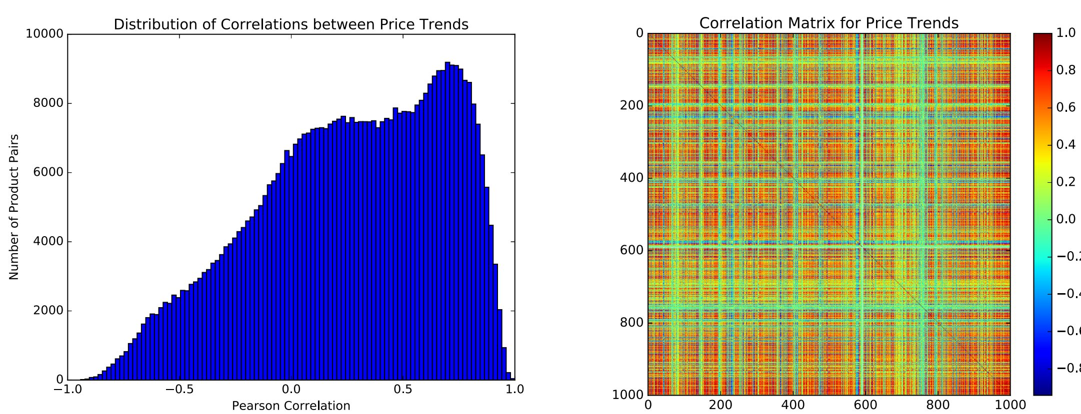


Figure 3.8: Distributions of the mean prices of products over all weeks, where the mean price of each product is calculated by aggregating the gross sales and quantities sold of all transactions involving the product over all weeks.

structure in the correlation visualization, which indicates that significant correlation exists across different categories of products.



(a) Distribution of correlations

(b) Correlation matrix

Figure 3.9: Distribution of Pearson correlation coefficients for pairs of price trends drawn from a set of 1,000 random products. Correlation matrix of 1,000 random products, organized by category.

3.5.2 Cost Trends

In this section, we apply the previous analysis on prices to costs.

Mean Costs

Figure 3.10 illustrates the distribution of mean costs of products. We see a similar shape for the distribution of mean costs as we did for mean prices and small magnitudes of costs than we saw with prices.

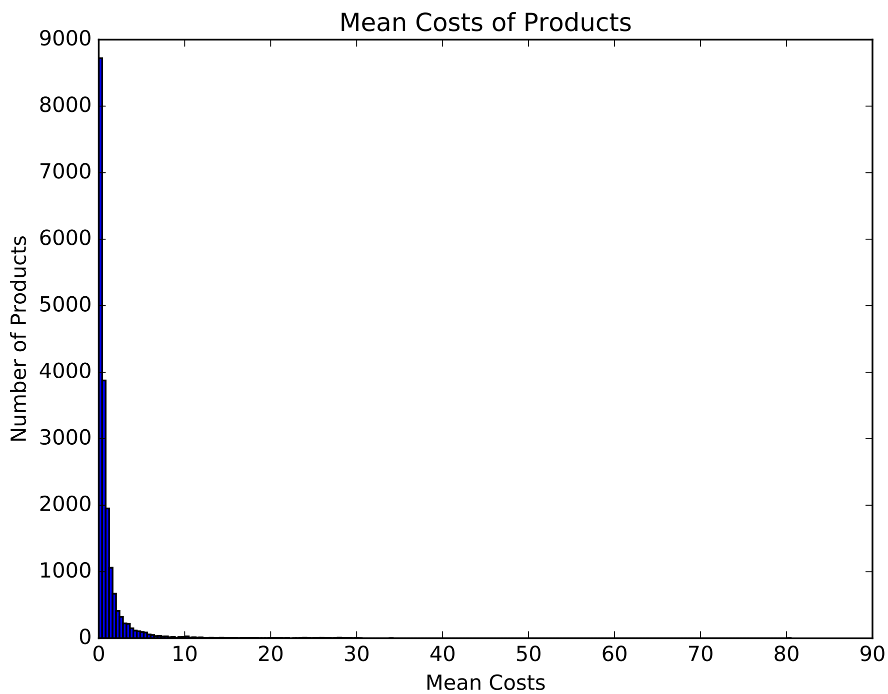
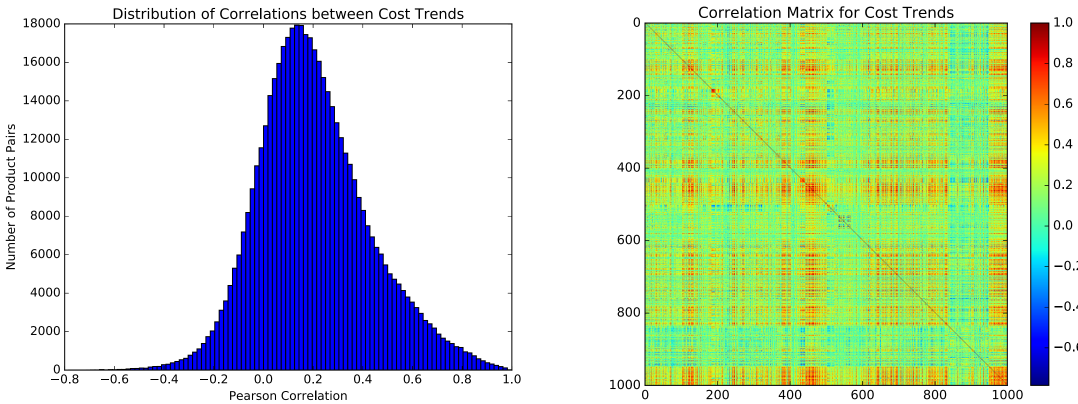


Figure 3.10: Distributions of the mean costs of products over all weeks, where the mean cost of each product is calculated by aggregating the wholesale costs and quantities sold of all transactions involving the product over all weeks.

Correlations between Products using Cost Vectors

Figure 3.11 plots the distribution of Pearson correlation coefficients for pairs of cost trends drawn from a set of 1,000 random products, alongside a correlation matrix for the same set of pairs, where the products are organized by category. While fewer correlations in Figure 3.11(a) are highly positive, the distribution still has most of its mass above a Pearson correlation of 0.



(a) Distribution of correlations

(b) Correlation matrix

Figure 3.11: Distribution of Pearson correlation coefficients for pairs of cost trends drawn from a set of 1,000 random products. Correlation matrix of 1,000 random products, organized by category.

We repeat the exploratory analysis for markups and do not observe products that are significantly correlated based on their markup vectors.

Chapter 4

Methodology

4.1 Outline of Goals and Experiments

Our goal in this section is to describe our methodology for (1) learning periods of time over which price setting decisions are made and (2) learning interesting regular prices for each price setting period (PSP). The first task is a generalization of Eichenbaum et al. (2011)'s ad-hoc choice of fixed-size quarters for price setting periods. The second task explores a generalization of Eichenbaum et al. (2011)'s reference price measure, which refers specifically to the mode of the prices in a given quarter. As demonstrated in Chapter 3, modes by themselves are not a good measure of reference price due to noisy measurements. We replace the exact definition of reference price by a similar notion: instead, we learn a Gaussian mixture model over the trend of each price setting period in a manner that tends to align cluster centers with modes of different orders. That is, we learn an approximation of the mode of order 1 (the most common price), the mode of order 2 (the second most common price), and so on up to the mode of order k (where k is the number of clusters learned by the mixture model). It is important to note that we are truly using a mixture of the notion of mode and mean when we learn our "modal prices." Then, we are able to define measures of cross-PSP and within-PSP stickiness using the models we have previously defined.

We define notions similar to price duration, examine the total absolute deviation between regular prices, and fit Markov models to each price setting period. We also use the entropy rate of the Markov transition matrix as a single-dimensional value for characterizing the internal stickiness of price setting periods. Finally, we are able to use these stickiness metrics as well as trend data to cluster the products according to price stickiness. Along the way, we also check the linear relationships between costs and prices by fitting a linear regression and examining the effect of using previous time steps to learn the model, allowing us to test if there is a significant delay in response between changes in costs and changes in prices.

We thus address the following questions:

1. *Can we automatically identify price setting periods for any product?*
2. *Do "reference prices" exist? Can we characterize and identify reference prices?*
3. *How do we measure price stickiness?*
4. *Can we capture the effect of costs on prices?*
5. *Which products are similar in general, and which products are similar with respect to price stickiness?*

4.2 Learning Price Setting Periods

Eichenbaum et al. (2011) define the reference price of a product to be the most commonly observed price within a business quarter. One weakness of this methodology is that this approach assumes that these segments of time are fixed in length across products and are aligned with quarters. This assumption may not hold since some products experience more frequent price changes than others. Furthermore, we generally observe in the data that prices do not change according to fixed quarters of time.

Instead of assuming fixed quarters, we aim to infer a set of time periods for a given product's price trend in which the prices found in any particular time period remain relatively fixed within that period. Thus, our first objective is to systematically identify potentially uneven time periods over which prices remain within a relatively fixed range. In the remainder of the paper, we refer to these variable-length time periods that we identify for each product as price setting periods (PSPs).

4.2.1 Event-Based Hidden Markov Model

We apply the model developed by Baldassano et al. (2016) to our scanner dataset. Originally, this model was created by neuroscientists studying how the human brain changes state over time in response to stimulus in the form of a movie, which is easy to segment into different scenes or "events." Baldassano et al. (2016) investigate a particular question: when an input stimulus has shifts in context, is it possible to recover the boundaries between the scenes of the movie from the brain data? Their methods are able to successfully infer the appropriate scene boundaries, given time series data collected from participants in their experiments.

From Neuroscience to Economics

Given the time series of prices for a product, our goal is to infer the various price setting periods such that the prices in any particular PSP remain relatively fixed within that PSP. Our objective applied to scanner data is thus quite similar to the goal of identifying movie events from brain data. However, there is a key difference between the neuroscience and economic settings: brain data are high-dimensional in nature, typically on the order of thousands of dimensions, while our price data are one-dimensional. Due to the brain's division into separate structures for different tasks, the high-dimensional time series carry multi-timescale information (in other words, the data are a superposition of many frequencies). While it is possible that economic price data can be decomposed into multiple frequencies as well, we hypothesize that the trends are much simpler. Another difference between the neuroscience and economic contexts lies in the number of time segments that we expect the brain data and the economic data to have. Throughout a movie, there are a relatively large number of scenes (approximately 50). Over the course of nearly three and a half years, the timeline spanned by our scanner data, we expect there to be far fewer changes in price, based on the literature (Eichenbaum et al. 2011). As we describe in the next chapter, it turns out that these differences are sufficiently insignificant in terms of performance of the model.

Model

The model is very similar to a hidden Markov model (summarized in the Appendix), with a key additional assumption. Here, the hidden states correspond to the price setting periods. Instead of allowing positive transition probabilities between any pair of hidden states, this model allows only the transition between state i and itself and the transition between state i and state $i+1$ to have nonzero probability. This fact can also be expressed by a restriction of the transition matrix to be upper bidiagonal. We are required to provide as input a parameter N_{states} indicating the number of events, or in our context, price setting periods.

Baldassano et al. (2016) model the HMM as having isotropic Gaussian outputs and restrict the transmission matrix to be upper bidiagonal to enforce the constraint that, given a current state, the process can only stay in the same state or transition to the next state. The actual model is as follows: given that the process is current in state k at time t,

$$\mathbb{P}(\text{output}_t \mid \text{mean}_k) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left\|z(\text{output}_t) - z(\text{mean}_k)\right\|_2^2\right) \tag{4.1}$$

where $z(\cdot)$ is a function denoting the z-score (mean of 0, standard deviation of 1) representation over the data, where output_t is the output price at time t and mean_k is the mean price of cluster k. Note that in the model by Baldassano et al. (2016), they make the assumption that brain outputs are high-dimensional. In our model, the $\|\cdot\|_2^2$ term is replaced by a simple one-dimensional squared term. They optimize the log likelihood as is standard.

The transition matrix and output distribution are learned using an annealed version of the Baum-Welch algorithm, where the variance is scheduled to decay at the following rate, where i indexes the iteration: $\sigma^2(i) = 4 * 0.98^i$. This technique is standard in optimization, the intuition being that for high values of the variance, we want to explore the parameter space, and for low values of the variance, we want to focus on optimizing a certain part of parameter space. In practice, this procedure is helpful for obtaining solutions of higher likelihood.

To fit the model, we use code provided by the authors of Baldassano et al. (2016), which can be found at `https://github.com/IntelPNI/brainiak`. Their approach is essentially a modified version of the Baum-Welch algorithm, which is a standard technique used in fitting hidden Markov models.

4.2.2 Model Selection Procedure

Our model is unsupervised in the sense that we do not see hidden state labels (the PSPs) for the prices. Since the number of total hidden states j is a fixed input to the model, we need to tune this value for each product trend. Thus, the necessity of a procedure for automatically determining j arises.

We solve the following optimization problem to determine j:

$$\begin{aligned}\hat{j} &= \underset{j\in\mathbb{Z}_+,\ 1\leq j\leq J_{\max}}{\operatorname{argmin}} \left\{ \frac{1}{A_{\max}} \cdot \frac{1}{j} \sum_{\text{PSP } i} \sum_{t\in\text{PSP } i} (p(t)-\mu_i)^2 + \frac{1}{J_{\max}} \cdot \gamma j \right\}, \\ A_{\max} &= \max_{j\in\mathbb{Z}_+,\ 1\leq j\leq J_{\max}} \frac{1}{j} \sum_{\text{PSP } i} \sum_{t\in\text{PSP } i} (p(t)-\mu_i)^2.\end{aligned} \tag{4.2}$$

$J_{\max}$ is a fixed input parameter to limit the number of hidden states that our model can possibly identify. μ_i is the mean of the prices contained in a given PSP i, and γ is a parameter for penalizing the identification of an additional price setting period.

The term $\frac{1}{j}\sum_{\text{PSP } i}\ \sum_{t\in\text{PSP } i}\,(p(t)-\mu_i)^2$ measures the sum of the squared residuals of price points within a given PSP i from μ_i, the mean price in PSP i, averaged across all j PSPs. Finally, $A_{\max}$ and $J_{\max}$ are included in the objective function to normalize the first and second terms so that they can be directly comparable with each other.

This model selection procedure captures a couple of goals that we have. First, for each identified PSP, we want to minimize the aggregate amount of deviation of prices in the PSP from the average price of the PSP, so we use the mean squared error. On the other hand, we do not wish to identify too many PSPs: we want to avoid potential overfitting, and we do not expect there to be a significant number of PSPs in a period of less than four years. Finally, we want to balance these goals against each other equally, subject only to paramter γ. For that reason, we normalize each term to take value in $[0, 1]$. Thus, the two terms in the objective, together, find an optimal $\hat{j}$ for our purposes.

4.3 Identifying Reference and Other Modal Prices

Our goal in this section is to determine the "modal prices" of a product. We define a modal price of order i to be the price that appears the i^{th} most frequently over some period of time. The reference price is defined as the modal price of order 1.

There are now several ways in which we can proceed to identify these modal prices. First, we can completely ignore the price setting periods that we previously learned for each product and attempt to evaluate the modal prices over the whole timecourse for each product. Second, we can use the previously identified price setting periods as the periods over which we identify the modal prices. We now discuss the methodology for each procedure.

4.3.1 Unsupervised Hidden Markov Model

We use the Baum-Welch algorithm to find the unknown parameters of a hidden Markov model (HMM), which we describe in more detail in the Appendix. Notably, we fit the HMM both across the entire time trends and within price setting periods. We use the unsupervised HMM implementation by Stratos, Collins, and Hsu (2016), with the code available at `https://github.com/karlstratos/anchor`.

4.3.2 Kullback-Leibler Divergence

For a given PSP, we want to be able to characterize how "close" its distribution of prices is to a uniform distribution over the same set of prices—that is, we would like to estimate how much randomness that the data trend seems to exhibit. We can calculate the Kullback-Leibler (KL) divergence between the price distribution P of a given PSP and a uniform distribution U over the set of prices in the PSP, where the KL divergence is defined as the following:

$$D_{\mathrm{KL}}(P\|U) = \sum_i P(i) \log \frac{P(i)}{U(i)}. \tag{4.3}$$

The KL divergence is nonnegative, attaining the lower bound of 0 if and only if $P = U$. Additionally, we note that the KL divergence is not symmetric; in other words, generally, $D_{\mathrm{KL}}(P\|U) \neq D_{\mathrm{KL}}(U\|P)$. We choose to use $D_{\mathrm{KL}}(P\|U)$ since we want to take the expectation over the distribution of P and not over the uniform distribution: it is more informative to weight by more probable prices rather than treat all prices equally.

We can use the KL divergence to identify PSP trends that are less random and therefore more likely to have interesting, multi-modal distributions over prices. As we will discuss in the next chapter, we can find an appropriate cutoff such that PSPs higher than this cutoff have modal prices that merit further analysis.

4.3.3 Dirichlet Process Gaussian Mixture Model

After applying the methods from the previous section, each product's price trend is now split into distinct price setting periods. Here, our goal is to determine the number of modal prices in each price setting period that passes the KL divergence threshold that we previously discussed.

The purpose of using these quantities is to identify the reference price (the modal price of order 1) and further extend the work of Eichenbaum et al. (2011), who stop after identifying reference prices and group all other prices as "non-reference prices," by analyzing the set of non-reference prices for economic significance. For instance, a low modal price with a high order might indicate a temporary sale. Furthermore, Kehoe and Midrigan (2015) suggest that other temporary prices that do not correspond to sales might exist and would be interesting to examine. Identifying the number of useful modes also provides an additional lens into understanding price stickiness: Eichenbaum et al. (2011) analyze price stickiness by only examining the reference price, whereas with our model, we can now investigate the persistence of prices at multiple modal levels. We set out to identify these modal prices by using a clustering technique: Gaussian mixture models.

Model Specification

A one-dimensional K-Gaussian mixture model approximates the true distribution of a dataset by assuming that it can be represented as a weighted combination of Gaussian densities. Thus, the parameters of this model that we seek to learn are a set of nonnegative weights, means, and variances (w_k, μ_k, σ_k^2), for each of the K Gaussians, such that the weights are on the probability simplex. Our estimation of the distribution $D(K, w, \mu, \sigma^2)$ of a price setting period using a mixture of K Gaussians can thus take the following form:

$$D(K, w, \mu, \sigma^2) \approx \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\mu_k, \sigma_k^2). \tag{4.4}$$

This formulation can be fit using the Expectation-Maximization (EM) algorithm, which optimizes the log-likelihood of the parameters. We observe that in this formulation, we are required to decide on K, the number of Gaussian clusters we want to use, prior to fitting the model.

However, we would like to avoid choosing K. Thus, we replace the finite K-mixture model with an infinite mixture model of Gaussians and incorporate a distribution over weight distributions into the model. Based on our observations of the data, we expect that there are only a few modal prices; thus, we want to impose a constraint on the values of the weights. Namely, only a small fraction of the weights should be significantly large. We can encode this belief by assuming a Dirichlet process prior distribution on the weight distribution. Thus, our formulation of the model can be summarized as an infinite mixture model with a Dirichlet process prior on the weights.

In practice, we still have to approximate the Dirichlet process prior, and so we need to set a parameter governing the maximum number of components that we are willing to use. However, we find empirically that we can set this maximum value low enough and still ensure that the Dirichlet process never uses the maximum number of components when drawing simplex distributions.

We now outline the updated model and refer the reader to Frigyik et al. (2010) and Görür and Rasmussen (2010) for a more detailed presentation. Let p_n denote the n^{th} data point, an observed price in a PSP, and let μ refer to the mean of a Gaussian and σ^2 refer to the variance of a Gaussian. We define our cluster parameter distribution prior (the base distribution) $H((\mu, \sigma^2))$ jointly over parameters μ and σ^2 by

$$H((\mu, \sigma^2)) \sim \mathcal{N}(\bar{\mu}, \bar{\sigma}^2) \times \chi^2(1), \tag{4.5}$$

The part of the joint distribution focusing on the mean is initialized as a normal distribution with mean $\bar{\mu}$, the empirical average of the price data, and with variance $\bar{\sigma}^2$, the empirical variance of the price data. The distribution over variances is a χ^2 distribution initialized with the degrees of freedom of the data, which is taken to be

1, the dimension of the price data. Many other choices of priors are possible here; however, our particular choices are simple and make computation easier.

Now, we define the Dirichlet process $\mathrm{DP}(H((\mu, \sigma^2)), \alpha)$, a distribution over distributions, with α denoting the "concentration parameter." Using a larger α means that we tend to draw distributions that are closer to uniform, and using a smaller α means that we draw distributions that put most of their mass on a few points. Consider a draw $P \sim \mathrm{DP}(H((\mu, \sigma^2)), \alpha)$. Let θ denote the paired mean and variance parameters (μ, σ^2). We have that for any finite partition of the parameter space $(\Theta_1, \cdots, \Theta_m)$ of $\Theta = \{\theta\}$, the distribution P has the property that

$$(P(\Theta_1), \cdots, P(\Theta_m)) \sim \text{Dirichlet}\left(\alpha H(\Theta_1), \cdots, \alpha H(\Theta_m)\right). \tag{4.6}$$

It turns out that we can write P as the distribution

$$P(\cdot) = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}(\cdot), \tag{4.7}$$

where the positions of the mean-variance pairs are θ_k and are drawn from H randomly, and where the w_k are probability weights associated with each Gaussian (μ_k, σ_k^2). It remains to define a sampling procedure for the w_k such that we satisfy the condition outlined above (namely, that any realization of P is a Dirichlet distribution). Then, we can sample the θ_k from H and the w_k from our procedure to generate a distribution P, from which we can finally draw $(\mu_{i(n)}, \sigma_{i(n)}^2)$ for each data point n, where $i(n)$ denotes the identity of the cluster for data point n.

So far, the way that we have described the process assumes draws of an infinite-length vector w from the infinite-dimensional probability simplex. However, the resolution is simple: it turns out we can draw w_k one at a time until we reach a point when we decide to stop. We draw parameters $\beta_1, \beta_2, \ldots$ from a "stick-breaking" distribution, where each $\beta_j \sim \mathrm{Beta}(1, \alpha)$. It turns out that if, after sampling $\beta_1, \beta_2, \ldots$

and $\theta_1, \theta_2, \ldots$, we set the weights to be

$$\begin{aligned}
w_1 &= \beta_1, \\
w_2 &= \beta_2 * (1 - \beta_1), \\
w_3 &= \beta_3 * (1 - \beta_1) * (1 - \beta_2), \\
&\vdots
\end{aligned} \tag{4.8}$$

then each of these weights is part of a true draw w from the infinite-dimensional probability simplex.

Now, we can define our model in its entirety, describing precisely the process we assume that the data is drawn from. We assume that prices are normally distributed and formulate the following:

$$\begin{aligned}
p_{i(n)} \mid \mu_{i(n)}, \sigma^2_{i(n)} &\sim \mathcal{N}(\mu_{i(n)}, \sigma^2_{i(n)}) \\
(\mu_{i(n)}, \sigma^2_{i(n)}) &\sim P = \sum_{k=1}^{\infty} w_k \delta_{\theta_k} \\
\theta_k &\sim H(\mu, \sigma^2) \text{ i.i.d.} \\
w_k &\sim \left(\beta_k \cdot \prod_{j=1}^{k-1} (1 - \beta_j) \right) \mid \{\beta_j\}_{j \in [k]} \\
\beta_j &\sim \text{Beta}(1, \alpha).
\end{aligned} \tag{4.9}$$

Note that P is in fact a discrete distribution over the list of cluster parameters $\theta_1, \theta_2, \ldots$. We use $i(n)$ to denote the cluster label.

We now summarize by clarifying the data generation process:

1. Draw $\beta_k \mid \alpha \sim \text{Beta}(1, \alpha)$ for $k = 1, 2, \ldots$. Use these to form the w_k.

2. Draw $\theta_k \mid H \sim H$ for $k = 1, 2, \ldots$. These are the Gaussian mixture parameters $\theta_k = (\mu_k, \sigma^2_k)$.

3. Now, P, the distribution over clusters, can be formed.

4. For the n^{th} data point,

(a) Draw the cluster label $Z_n \mid w_1, w_2, \ldots \sim \text{Multinomial}(w_1, w_2, \ldots)$. All that we are doing here is drawing the class to which the current data point belongs. In the specification above, $i(n) \sim Z_n$.

(b) Draw the price realization for the class from the (Gaussian) price distribution for the class: $p_n \mid i(n) \sim \mathcal{N}(\mu_{i(n)}, \sigma^2_{i(n)})$.

Variational Inference for Fitting the DPGMM

Given the data generation process outlined above, we seek to compute the posterior distribution $\mathbb{P}(V|p,\theta) = \exp\left(\log \mathbb{P}(p, V|\theta) - \log \mathbb{P}(p|\theta)\right)$ for our Dirichlet process Gaussian mixture model (DPGMM) with latent variables V, price outputs p, and Gaussian cluster parameters θ, for which no direct computation exists due to the intractability of evaluating the log marginal probability $\log \int \mathbb{P}(p, V|\theta)dV$. However, we can approximate likelihoods and the posterior distribution using the procedure of variational inference. We summarize the key concepts behind variational inference and refer the reader to Blei and Jordan (2004) for a detailed treatment.

Essentially, the idea is that we parameterize a tractable class of "variational distributions" $q_\nu(\tilde{V})$ as an approximation to the true posterior, with variational parameters ν and variational latent variables $\tilde{V}$. Then, we minimize the KL divergence over ν

$$D\left(q_\nu(\tilde{V}) \parallel \mathbb{P}(V|p,\theta)\right) = \mathbb{E}_q\left[\log q_\nu(\tilde{V})\right] - \mathbb{E}_q\left[\log \mathbb{P}(V|p,\theta)\right] + \log \mathbb{P}(p|\theta), \tag{4.10}$$

trying to bring the true posterior and our approximation $q(\tilde{V})$ as close together as possible.

In our model, we are trying to learn the β-parameters, the μ-parameters, and the σ^2-parameters, as well as the class labels. The concentration parameter α and the χ^2 parameter η are parameters that must be decided upon by the user. We must now find a family of variational distributions $\{q_\nu\}_\nu$ that approximate the infinite-dimensional distribution P. We achieve this goal by truncating the number of β_k, θ_k, and σ^2_k: we cap the value of k at $K_{\max}$. We also must set $q(\beta_{K_{\max}} = 1) = 1$, ensuring that mixture proportions w_k for $k > K_{\max}$ will be set to zero. Notably, we do not truncate the true model—the model is still an untruncated Dirichlet process Gaussian Mixture Model—as we only truncate the approximation to the true model that we are trying to learn. In our experiments, we fix $K_{\max} = 6$ and find that we never actually find that 6 clusters are necessary.

We use the following tractable class for our variational distribution family for q, supposing that there are N data points in our dataset:

$$q_\nu(\beta_1, \ldots, \beta_{K_{\max}}, \theta_1, \ldots, \theta_{K_{\max}}, i_1, \ldots, i_N) = \prod_{k=1}^{K_{\max}-1} q_{\gamma_k}(\beta_k) \prod_{k=1}^{K_{\max}} q_{\tau_k}(\theta_k) \prod_{n=1}^{N} q_{\phi_n}(i_n), \tag{4.11}$$

where q_{γ_k} are beta distributions with parameters γ_k, q_{τ_k} are normal distributions with parameters τ_k, and q_{ϕ_n} are multinomial distributions with parameters ϕ_n.

This class (known as a "mean-field approximation") is tractable since it assumes independence between the parameters. Here, we have a different variational parameter in $\nu = (\gamma_1, \ldots, \gamma_{K_{\max}-1}, \tau_1, \ldots, \tau_{K_{\max}}, \phi_1, \ldots, \phi_N)$ for every single variational latent variable $\tilde{V} = (\beta_1, \ldots, \beta_{K_{\max}}, \theta_1, \ldots, \theta_{K_{\max}}, i_1, \ldots, i_N)$ from the variational distribution q.

To optimize the KL divergence loss function, a standard coordinate descent algorithm is used. In practice, we use a pre-existing implementation of the Dirichlet process Gaussian Mixture Model from the `scikit-learn` Python module.

4.3.4 Label Identification

After fitting the Dirichlet process Gaussian mixture model for each PSP with modal prices, we need to identify which Gaussians are significant enough for their means to be marked as modal prices within the PSP. Moreover, we aim to formalize the notion of "modal prices of order i" in the context of the DPGMM.

The algorithm that we use to identify labels for each price in a given PSP follows; notably, we require the specification of λ, a threshold factor to test the significance of an identified Gaussian. For each PSP with modal prices, we run the following algorithm.

1. Fit the Dirichlet process Gaussian mixture model to the PSP. Identify the set of weights, means, and variances, $\{(w_k, \mu_k, \sigma_k^2)\}_{k=1}^{K_{\max}}$.

2. Sort $\{(w_k, \mu_k, \sigma_k^2)\}_{k=1}^{K_{\max}}$ in decreasing magnitude of w_k. Let $w_{\max}$ be the maximum value over all weights.

3. Calculate $\lambda w_{\max}$.

4. For each (w_k, μ_k, σ_k^2), $1 \leq k \leq K_{\max}$, we have two cases:

   (a) If $w_k \geq \lambda w_{\max}$, then we label μ_k as a modal price.

   (b) If $w_k < \lambda w_{\max}$, then we label μ_k as a non-modal price.

   We note that the modal prices have a natural ordering: the modal price of order 1 is the first modal price in the sorted list of triples (in other words, the mean price of the cluster with highest significance), the modal price of order 2 is the second modal price in the sorted list of triples, and so on.

Now that we have identified a set of significant Gaussians, we need to assign labels to each price in the PSP. Specifically, since we have now identified a set of modal prices, each of which has its own associated label, we can now identify a price as a modal price with a certain order or a non-modal price. We outline the algorithm that we use to identify the label for each price in a specific PSP. In this procedure, we require another thresholding parameter η for determining if a price is close enough to the mean of a significant Gaussian. For each price p_t in a given PSP with modal prices, we run the following algorithm.

1. Let $\{w_s, \mu_s, \sigma_s^2\}_{s=1}^{S_{\max}}$ denote the parameter set for the Gaussians that we identify to be significant in the last algorithm, where the Gaussians are sorted in decreasing order of w_s. We now iterate through $\{w_s, \mu_s, \sigma_s^2\}_{s=1}^{S_{\max}}$.

2. For the current Gaussian s, calculate $d = \|p_t - \mu_s\|_2^2$, the distance between the price and μ_s.

3. We have two cases:

   (a) If $d \leq \eta\sigma_s$, then we check if the current Gaussian is the first Gaussian in the ordered list that satisfies this threshold. If so, we label p_t with the label of $\{w_s, \mu_s, \sigma_s^2\}$. If we have previously found a Gaussian—in other words, if we have found a Gaussian with higher weighting that is close enough—then we keep the previous Gaussian's label. This check ensures that if a price is close enough to more than one significant Gaussian, we choose to assign the price to the most significant Gaussian.

   (b) If $d > \eta\sigma_s$, then p_t is not close enough to the current Gaussian, and we move to the next Gaussian in the ordered list.

4. After iterating through the entire list of significant Gaussians, if we do not find that p_t is close enough to any significant Gaussian, then we assign p_t a non-modal price label.

4.3.5 Summary

To recap our models, we first identify price setting periods with the event-based hidden Markov model. We then calculate the KL divergence of each PSP, and if the PSP has a KL divergence over a significance threshold, we fit its prices to a Dirichlet process Gaussian mixture model, from which we learn the significant Gaussians. We then assign prices to corresponding modal prices of certain orders, if the prices are close enough to a mean of a significant Gaussian, and if not, label them as non-modal. If the PSP does not satisfy the KL divergence cutoff, then we do not need to fit the DPGMM and simply estimate the reference price with the mean price of the PSP. We provide an example product's price trend and how we tackle its analysis in Figure 4.1.

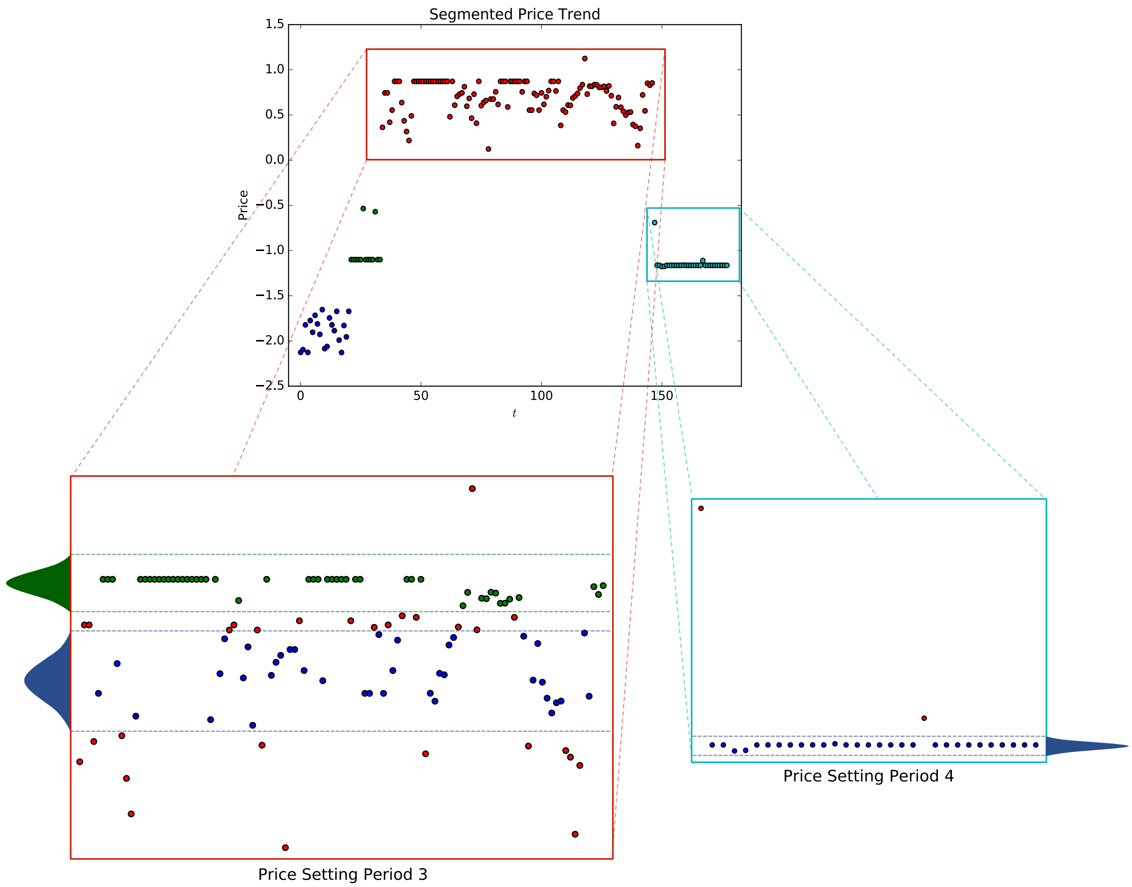


Figure 4.1: The results of our methodology applied to an example price trend. We first identify four price setting periods (PSPs), shown in the top visualization, and then calculate the KL divergence of each PSP. The third and fourth PSPs have a KL divergence over a significance threshold, so we fit PSP 3 and PSP 4 each to a Dirichlet process Gaussian mixture model. In PSP 3, we identify two significant Gaussians and, thus, two modal prices. Prices close enough to the green Gaussian, which has the highest significance, are labeled as modal prices of order 1 (reference prices), prices close enough to the blue Gaussian are labeled as modal prices of order 2, and the rest of the prices are non-modal. In PSP 4, we identify one significant Gaussian, to which most prices are assigned.

4.4 Measuring Price Stickiness

After finding price setting periods and identifying a set of modal prices, we can explore several measures of the rigidity of the prices for a given product. We first note that the upper bidiagonal matrix found during the identification of PSPs is not particularly interesting in terms of the entropy rate (to be defined in this section) due to the nature of the matrix. Instead, we discuss a set of measures that we define for calculating price stickiness within PSPs and across entire trends for products.

4.4.1 Measures of Stickiness Within Price Setting Periods

Duration of PSP

We define the duration of a PSP to be the number of weeks included in the PSP. If the prices in a given PSP are sticky, we expect the duration of that PSP to be long, since the PSP is optimized to have a set of prices that do not vary significantly. Thus, duration is a simple measure of the price stickiness for the prices in a PSP, which is slightly different from the typical measure that uses the inverse of frequency of price changes across time points.

Entropy Rate

For a given PSP, we can use the labels learned from the Gaussian mixture model to fit a Markov model M to the PSP. We construct a transition matrix P of the Markov model by estimating each entry $P_{i,j}$ with the following formula:

$$P_{i,j} = \frac{\text{Number of occurrences of state } i \text{ after state } j}{\text{Number of occurrences of state } j}. \tag{4.12}$$

Note that this construction enforces the columns to be probability distributions.

We can then calculate the entropy rate $H(M)$, defined as

$$H(M) = \sum_j \mu_j H(P_j) = -\sum_{i,j} \mu_j P_{i,j} \log P_{i,j}, \tag{4.13}$$

where $H(P_j)$ is the entropy of the distribution over transition states for state j, and μ_j is the j^{th} entry of μ, the stationary distribution of M.

The key intuition behind using the entropy rate as a measure of stickiness stems from the definition of entropy. Distributions that have a significant amount of randomness have higher entropy than nearly deterministic distributions; thus, we expect highly sticky trends to be nearly deterministic and thereby have a lower entropy.

The second aspect to the formulation of entropy rate as a price stickiness measure comes from the weights μ_j. A transition matrix has a probability density P_j over output states for each state j, and we can calculate the entropy for each density. We combine these entropies by weighting each $H(P_j)$ according to μ_j, the limit of the proportion of time that the Markov chain spends in state j.

As a result, we can view the entropy rate as the expected entropy of the output probability distribution over possible states. The intuition behind why lower entropy implies stickier prices thus transfers to the entropy rate.

4.4.2 Measures of Stickiness Across Products' Trends

Translating Per-PSP to Per-Product Stickiness Measures

For each PSP in a given product's price trend, we can calculate the duration. We can also calculate the entropy rate for PSPs that have a KL divergence that is high enough (modal PSPs), as we previously discussed. For the PSPs that do not have a large enough KL divergence (non-modal PSPs), we can approximate the stickiness of the prices in the PSP with the following method. For a given product, we can do one of the following calculations:

1. If there are other PSPs that are modal, we define N to be the maximum number of states across the modal PSPs and construct a Markov transition matrix, $M \in \mathbb{R}^{N\times N}$, such that each output distribution is uniform over all states. Consequently, since $M\mu = \mu$, the stationary probability distribution is uniform, so the entropy rate is $\log N$, the entropy of the uniform distribution over N states.

2. If there are no other PSPs that are modal, we let the entropy rate be $\log \tilde{N}$, where $\tilde{N}$ is the number of time points in the PSP.

If the product does have modal PSPs, we expect to be able to extract some information from these PSPs about typical fluctuation trends. Therefore, we can approximate the entropy rate of the non-modal PSPs by assuming that the distribution over states is completely random and, moreover, that there are no more states found in the non-modal PSPs than the ones found in any of the modal PSPs. Thus, our measure in case 1 allows us to fill in the missing entropy rates with a "maximal" entropy rate. In case 2, we do not have any information on the number of states, so we simply estimate the entropy rate using the number of time points, the maximal number of possible states.

ℓ_2-Norm of Per-PSP Duration Vector and of Per-PSP Entropy Rate Vector

From here, we can form $v_{\text{durations}}$, a vector of durations over a product's PSPs, and $v_{\text{entropy rates}}$, a vector of entropy rates over the PSPs. We can then calculate the ℓ_2-norms of $v_{\text{durations}}$ and $v_{\text{entropy rates}}$:

$$\begin{aligned} \|v_{\text{durations}}\|_2 &= \left(\sum_{\text{PSP } i} \left(v_{\text{durations}}(i)\right)^2 \right)^{1/2} \\ \|v_{\text{entropy rates}}\|_2 &= \left(\sum_{\text{PSP } i} \left(v_{\text{entropy rates}}(i)\right)^2 \right)^{1/2} . \end{aligned} \tag{4.14}$$

Note that our formulation incorporates both the dimension of the vectors, or the number of PSPs, and the magnitudes of the durations and entropy rates.

$\|\Delta\|_1$-Measures of Mean and Modal Prices

We define the $\|\Delta\|_1$-measure of a vector of prices to be the sum of the absolute values of the differences between the sequential pairs of prices. (Δ in the name refers to the vector of differences between prices.) For a given product, we have several relevant price vectors: we can construct a vector of the mean prices for each PSP, a vector containing the modal price of order 1 (the reference price) for each PSP, a vector containing the modal price of order 2 for each PSP, a vector containing the modal price of order 3 for each PSP, and so on. Note that if we do not have the relevant modal price—in other words, if the Gaussian mixture model finds only k modal prices for the PSP, and we are trying to calculate the $\|\Delta\|_1$-measure for the modal price of order greater than k—then we use the mean of the PSP as a substitute measure.

The point of this construction is to test if we can identify different levels of stickiness by looking at mean prices and at modal prices of different order. If we notice that modal prices of a certain order have a lower $\|\Delta\|_1$-measure, we can conclude that the modal prices of that particular order are stickier than the other sets of prices with which we are comparing the $\|\Delta\|_1$-measure.

4.5 Predicting Price Trends with Cost Trends

Eichenbaum et al. (2011) suggest that prices shift as a result of changes in costs. We set out with a more complicated regression technique to verify that costs indeed do have predictive power. We also attempt to learn a time lag by utilizing the k previous values of costs to predict prices. To make this procedure tractable, we first apply dimensionality reduction on both the price data and the cost data and repeat the analysis for a range of values for k. We also attempt this analysis on the full, high-dimensional price and cost data for a fewer number of values for k. Finally, we repeat our procedure using k previous values of prices to predict costs to test for potential reverse correlations.

4.5.1 Ridge Regression Optimization

Our procedure is explained visually in Figure 4.2. We define the number of products to be N. We then define a cost data matrix $X \in \mathbb{R}^{N \times T}$ and a price data matrix $Y \in \mathbb{R}^{N \times T}$, where T is the number of time points. We now apply dimensionality reduction with principal component analysis (summarized in the Appendix) to X and Y, choosing the minimum number of principal components that are need to capture 95 percent of the variances for each of X and Y. Let n_{cost} and n_{price} denote the new dimensions of the cost matrix and the price matrix, respectively. We denote

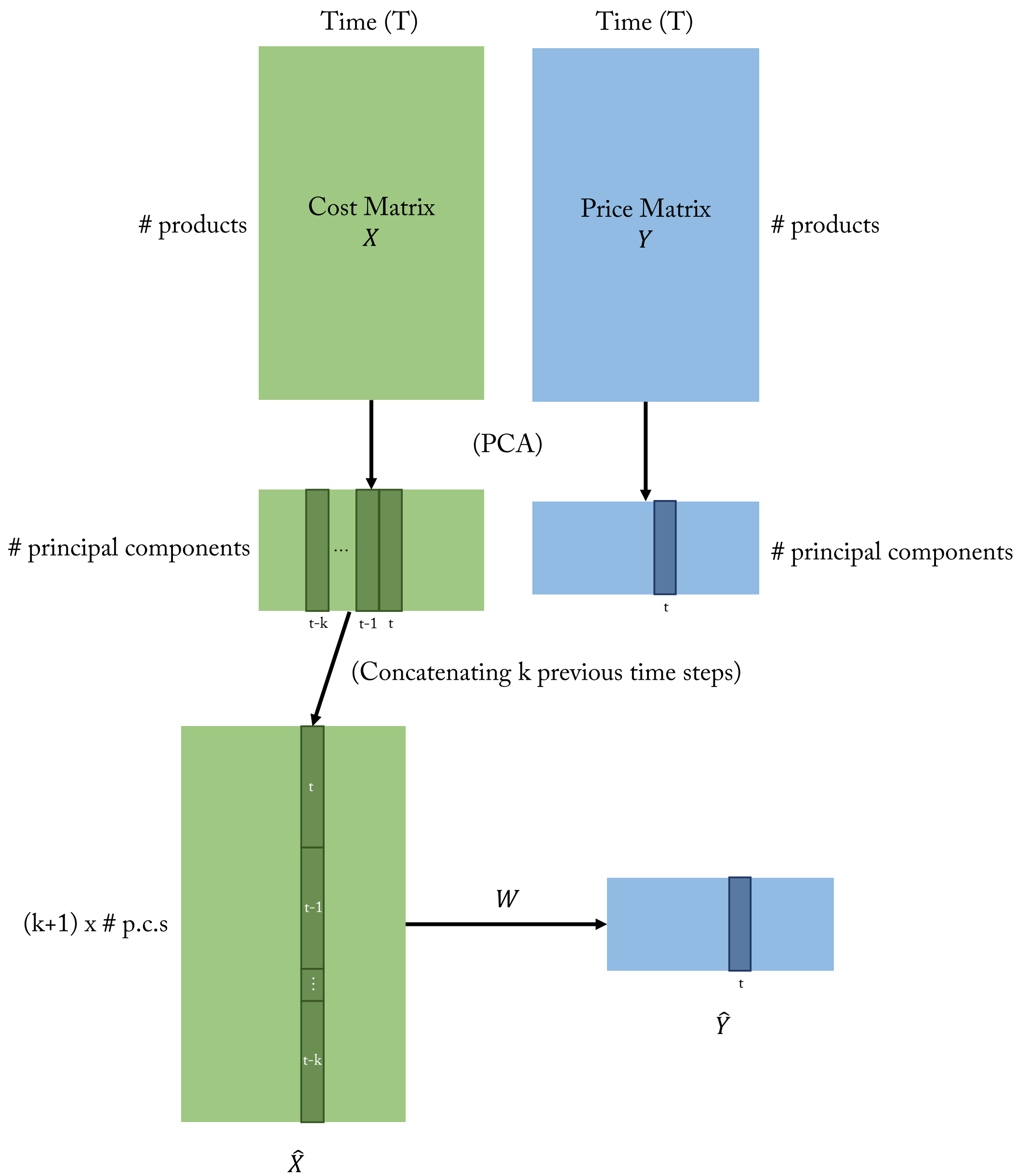


Figure 4.2: Methodology for ridge regression. We first apply dimensionality reduction to the cost data X and to the price data Y and then form $\hat{X}$ and $\hat{Y}$. We aim to learn a map W such that $W\hat{X} \approx \hat{Y}$.

by $\hat{Y} \in \mathbb{R}^{n_{\text{price}} \times T}$ the dimension-reduced version of Y.

Now, we define $\hat{X} \in \mathbb{R}^{(k+1)n_{\text{cost}} \times T}$ as follows, recalling that k is the number of previous time points we use in our prediction.

1. First, for each $n_{\text{cost}} \times 1$ time point t in X, consider the previous k time points at

locations $t-1, \ldots, t-k$. If $t-k<0$ (zero-indexing), then set that time point to be the all-zeros vector of dimension n_{cost}. Now, we have a $n_{\text{cost}} \times (k+1)$ matrix subset of the dimension-reduced cost matrix.

2. Concatenate the columns of the subset defined in the previous step so that they form a $(k+1)n_{\text{cost}}$ dimensional vector.

3. The resulting vectors form a matrix $\hat{X} \in \mathbb{R}^{(k+1)n_{\text{cost}} \times T}$, which represents the temporal dynamics of cost over the past $k+1$ time points (including the current point).

We want to learn $W \in \mathbb{R}^{(k+1)n_{\text{price}} \times n_{\text{cost}}}$ such that $W\hat{X} \approx \hat{Y}$. We seek to find the optimal $\hat{w_i}$, for each dimension of data in $\hat{Y}$. To solve this problem, we formulate the ridge regression objective in terms of each row w_i of W:

$$\hat{w_i} = \underset{w_i}{\operatorname{argmin}} \, \|w_i^T X - Y_i\|_2^2 + \|w_i\|_2^2. \tag{4.15}$$

Here, we penalize the reconstruction loss $\|w_i^T X - Y_i\|_2^2$ with an ℓ_2-penalty term, which encourages solutions w_i with smaller norm. This penalty is for the purposes of regularization* as well as to make the problem identifiable. Standard least squares regression does not necessarily have a unique solution due to a mismatch in rank between $\hat{X}$ and $\hat{Y}$; adding the ℓ_2 ridge penalty ensures that a unique solution exists.

Now, we can build the full matrix of weights W by stacking the w_i vectors:

$$W = \begin{pmatrix} | & | & & | \\ w_1 & w_2 & \ldots & w_{n_{\text{cost}}} \\ | & | & & | \end{pmatrix}^T. \tag{4.16}$$

Note that we train on the first 80% of the data and test upon the last 20% of the data. We measure reconstruction error, namely,

$$\text{Reconstruction Error} = \frac{1}{n_{\text{price}}} \left\| W\hat{X} - \hat{Y} \right\|_2^2, \tag{4.17}$$

*Regularization is a term we use when we deliberately introduce model mismatches in order to avoid overfitting—that is, cases when the model fits nearly exactly to the training data but performs poorly on testing data.

where we normalize by the number of entries in the matrix. This metric captures the goodness of W's fit and measures the average mean squared error over the elements of the reconstructed data matrix.

4.6 Clustering Products

We summarize two types of analyses for clustering products that we use on our dataset. First, we cluster products based on their time series of prices, using a dimensionality reduction method. We then cluster products based on the measures of price stickiness that we have defined.

4.6.1 Clustering with Dimensionality Reduction

Clustering with Price Trends

One method of clustering products lies in identifying similarity between products based completely on their time series of prices. We use principal component analysis (PCA), a method for dimensionality reduction, to identify a small subset of "principal components" that capture a significant amount of the total variance across the price data for all products. More details about the theory behind PCA can be found in the Appendix.

4.6.2 Clustering with Price Stickiness Features

Our second objective in clustering products is to be able to identify clusters of products with highly sticky price trends and clusters of products with less persistent prices found throughout the timeline of our dataset. We define each product in terms of ($\|\Delta\|_1$-measure for reference prices, ℓ_2-norm of entropy rates for PSPs), a vector of two of the price stickiness metrics, noting that the two-dimensional nature of this vector makes it highly interpretable and easy to visualize.

Chapter 5

Results and Discussion

We summarize the key insights gained from applying our methodology to our scanner dataset.

5.1 Discovered Price Setting Periods

We evaluate the performance of the key models to which we fit our price trends. Notably, the event-based hidden Markov model is able to confirm that significant price setting periods exist and, moreover, is able to identify PSPs with reasonable success. The event-based HMM generally performs well across different kinds of trends, finding varying numbers and shapes of PSPs across products. Furthermore, the identified PSPs capture interesting internal dynamics, which we analyze to present additional results discussed in the subsequent sections of this chapter.

Figure 5.1 illustrates the distribution over all products of the number of PSPs identified by fitting each product's price trend to the event-based HMM. The distribution captures a range of small numbers of identified PSPs, so the model selection procedure that we covered in the previous chapter is successful in helping us avoid overfitting from the identification of too many PSPs.

Figure 5.2 displays the PSPs identified using the event-based HMM for several

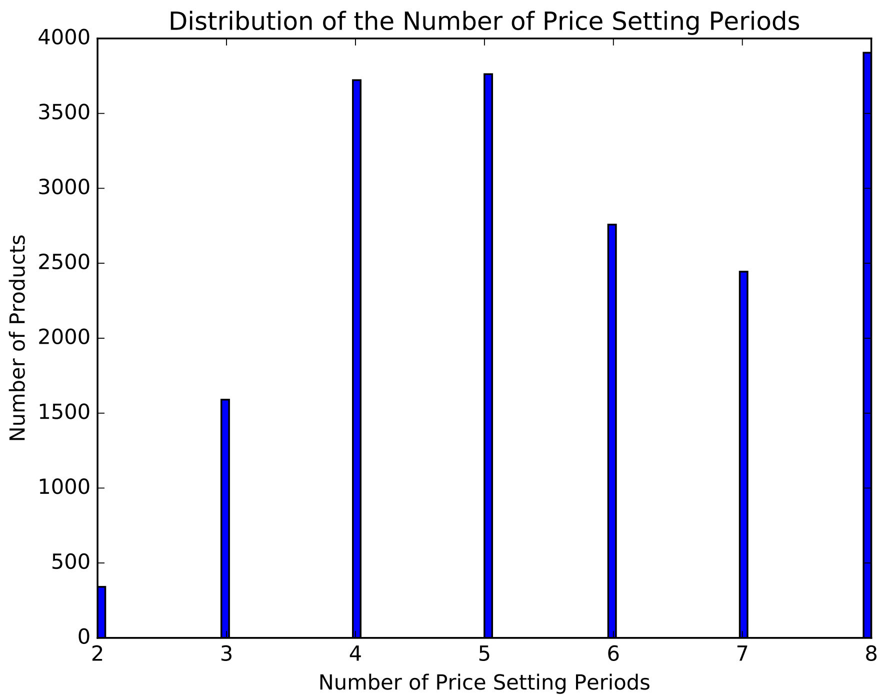


Figure 5.1: Distribution over all products of the number of price setting periods identified by fitting each product's price trend to the event-based hidden Markov model.

different products, with each color denoting a different PSP in the product's trend. We generally see that each PSP includes a set of prices that have reasonably low variance and that merging a PSP with the preceding PSP would, for the most part, increase the variance of prices included. In other words, for each example in Figure 5.2, if more PSPs were identified, then we likely would be overestimating the number of periods.

We also observe from Figure 5.2 that the shapes of many PSPs can be described as "clouds" of points and "lines" of points, where clouds of points have higher variances but generally trend in a stabilized manner around a fixed price level. Lines are the low-variance version of clouds: we often notice that the prices over PSPs with this shape do not change at all, or if they do, only vary in small amounts. One surprising observation is that three broadly-defined types of products exist: (1) products with PSPs that are all shaped as lines, (2) products with PSPs that are all shaped as clouds, and (3) products with a mix of these two PSP shapes.

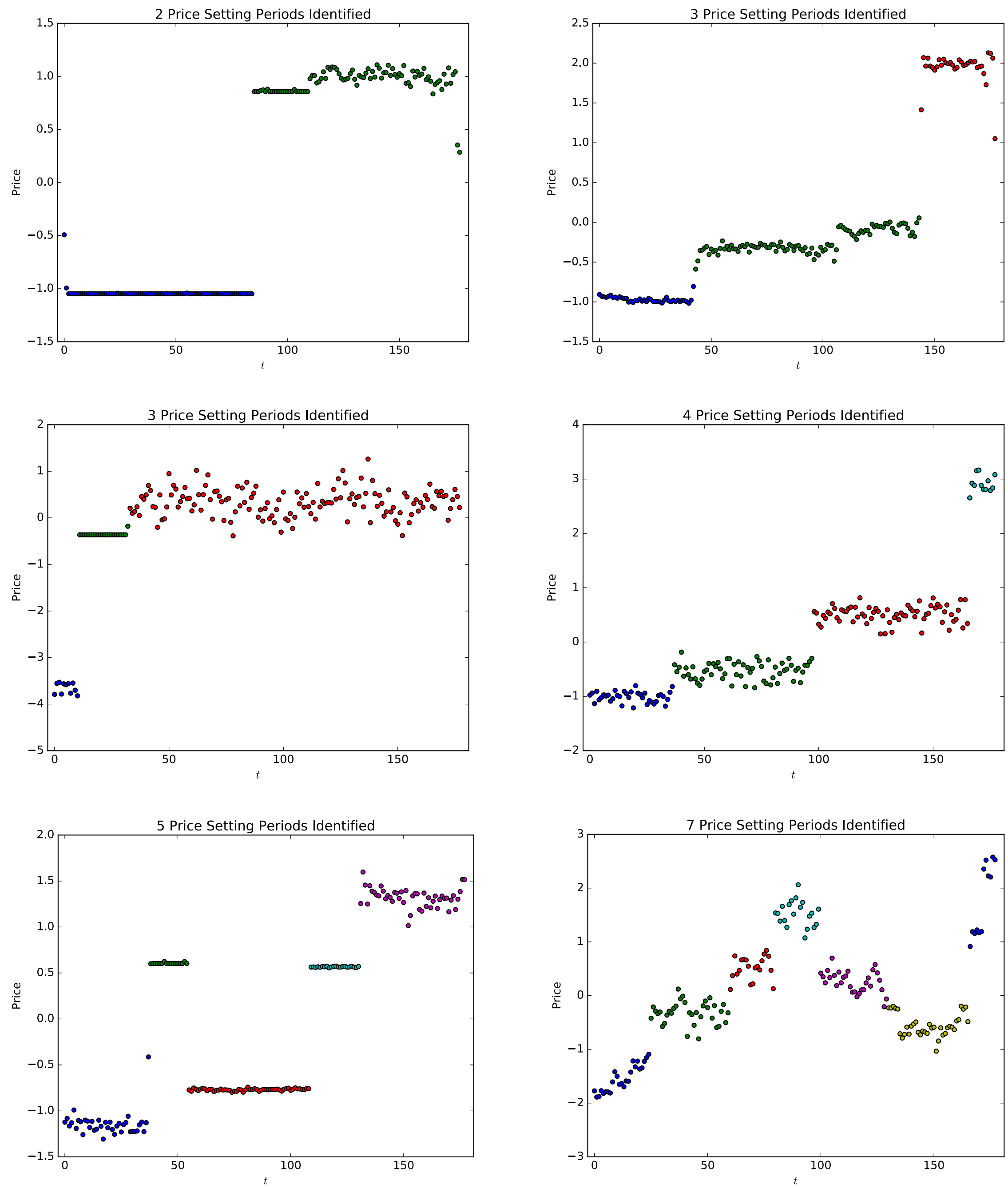
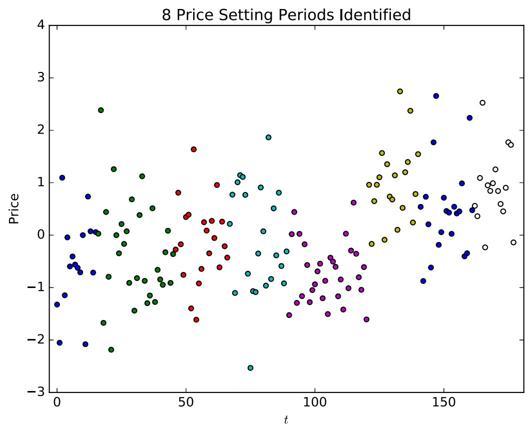
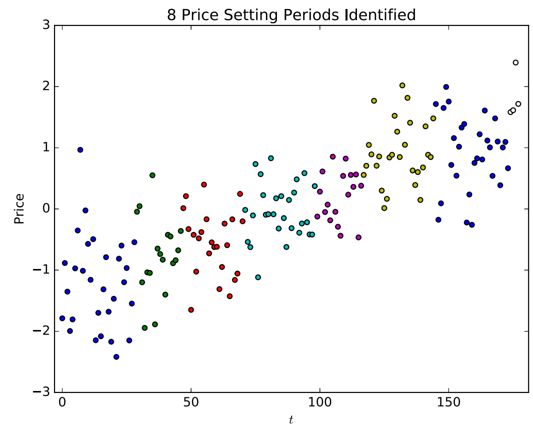


Figure 5.2: Examples of the price setting periods identified from fitting price trends to the event-based hidden Markov model, with each color denoting a different PSP.

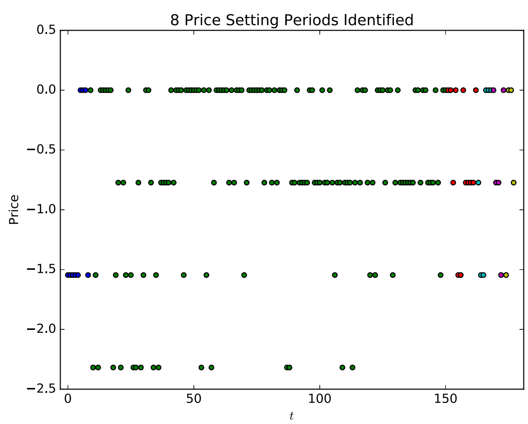
Figure 5.3 illustrates the PSPs identified for examples corresponding to four key types of trends that we encounter frequently throughout the dataset. Each example presents an issue with our PSP identification model that we address in subsequent steps of our methodology—namely, we describe how we address noisy data, steadily increasing or decreasing trends, and trends with prices that rarely repeat from one week to the next week, and how we can consolidate PSPs when we identify too many PSPs.



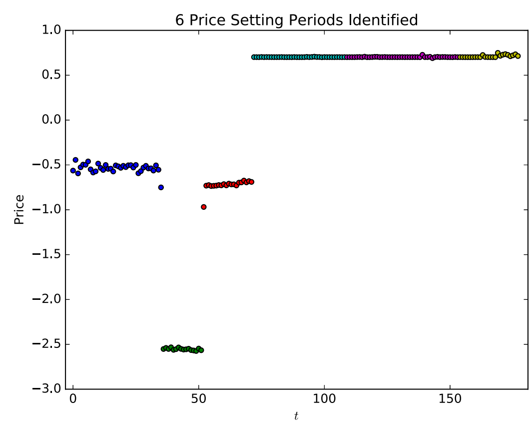
(a) Noisy data trend



(b) Steadily increasing trend



(c) Violently fluctuating data



(d) Too many price setting periods identified

Figure 5.3: Examples of the price setting periods identified from fitting price trends to the event-based hidden Markov model, with each color denoting a different PSP. Each example presents an issue with our PSP identification model that we address with subsequent steps of our methodology—namely, we are able to address noisy data, steadily increasing or decreasing trends, and trends with prices that rarely repeat from week to week, and we can consolidate PSPs when we identify too many PSPs.

Noisy Data

In Figure 5.3(a), we plot an example of noisy data, which we encounter frequently in our scanner dataset. Our event-based HMM identifies eight PSPs, which is high compared against other products. The main issue, however, is that within each identified PSP, the variance of the prices is quite high. We will later account for this aspect of the data by essentially determining whether each PSP has prices too close to uniform noise. In the case where the prices are too uniformly distributed, the best option for getting an idea of the typical price of the product is to simply take an average (since the mode is not a useful measurement in this kind of data).

Steadily Increasing or Decreasing Trends

In Figure 5.3(b), we plot an example of a steadily increasing trend. For these trends, either the notion of price setting periods does not apply (in other words, price is, generally, steadily increasing) or the price setting periods are very short in length, in which case, a large number of PSPs is more appropriate. In either case, analyzing a set of uniformly increasing time segments separately is equivalent to analyzing the entire trend.

Violently Fluctuating Data

In Figure 5.3(c), we plot an example of a trend with prices that alternate through a set of prices with a high frequency. We find that our model is unable to realize that the pattern of price activity remains relatively constant across all time points, likely due to the fact that the pattern in examples like this trend is highly oscillatory across a small set of fixed prices. The weakness of our model is that it also assumes that prices within the same price setting period will be close together, an assumption that examples like this trend violates. However, the same pattern of price behavior is present in all splits of the timeline, so breaking up the timeline into different PSPs will not affect the analysis much.

Too Many Price Setting Periods Identified

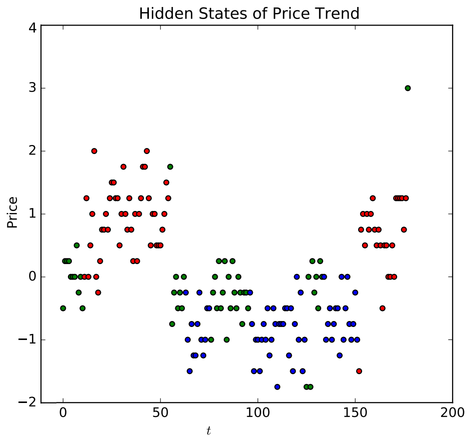
In Figure 5.3(d), we plot an example of a trend for which we identify too many PSPs. Specifically, the last three PSPs that the model identifies should be consolidated, which is an error in the identification of the number of price setting periods. However, splitting up the timeline into multiple PSPs does not affect our examination of the internal dynamics of PSPs at all.

5.2 Structure of Reference Prices and Other Modal Prices

Now that we have identified price setting periods, we present the results of our models that uncover modal prices, including reference prices (modal prices of order 1). We are able to identify PSPs that have relevant modal prices, and within those PSPs, we can identify specific reference prices and other modal prices. We find that only 11.2 percent of PSPs have modal prices and that among these PSPs containing modal prices, most PSPs have only one or two orders of modal prices.

5.2.1 Performance of Hidden Markov Model

Figure 5.4 plots example results from fitting a hidden Markov model on the entire price trend of a product and on a price setting period from the price trend of a different product. Each price is labeled with a color corresponding to a certain state learned from the HMM. Debatably, the HMM is able to detect broad bands of prices, apparent in the figure on the left: roughly, red corresponds to prices between 0 and 2, green corresponds to prices between -0.5 and 0.5, and blue corresponds to prices between -1 and 0, though there are errors. The figure on the right, however, does not seem to convey any useful information on modal prices.



(a) Entire trend

(b) Price setting period

Figure 5.4: Example results of fitting a hidden Markov model on the entire price trend of a product and on a price setting period from the price trend of a different product. Each price is labeled with a color corresponding to a certain state learned from the HMM.

5.2.2 Finding Price Setting Periods with Multi-Modal Price Distributions

Figure 5.5 plots the distribution over price setting periods of KL divergences, where each measures how close the price distribution in each PSP is to the uniform distribution over the set of prices found in that price distribution. The closer a KL divergence of a price distribution is to 0, the closer the distribution is to uniform. We can interpret this feature of the metric in terms of the number of modes of the price distribution (in other words, modal prices): A uniform distribution has the property that every value drawn from the distribution is a mode. The other extreme is a deterministic distribution, where there is only one value (in other words, the mode). Relaxing these conditions, we get that increasing values of KL divergence between the price distribution and the uniform distribution correspond to fewer and fewer modes of the price distribution and thus fewer and fewer modal prices. We want to choose a cutoff such that we capture the interesting situation where there are a few (for instance, fewer than six) modal prices.

Empirically, we examine random PSPs with KL divergences from every band capturing 10 percent of the distribution and find that a KL divergence threshold of 1.2 is appropriate to ensure that we only examine PSPs with potentially multi-modal price distributions. We find that 11.2 percent of price setting periods meet the KL divergence threshold criterion. PSPs below this threshold tend to be completely random, in which case, an average value summarizes the "modal" price behavior relatively well. Thus, after setting this threshold, we treat every PSP less than this threshold as though it has zero modal prices and summarize the price behavior in the PSP with the average price value. In practice, this procedure causes us to miss on on a few PSPs which may have a mode—however, in these cases, there tends to be only one mode that we miss out on due to the noise of the price distribution, which makes the distribution appear more random than it truly is. In these cases, summarizing the price distribution in the PSP by its average also gives us essentially the information we want from the PSP.

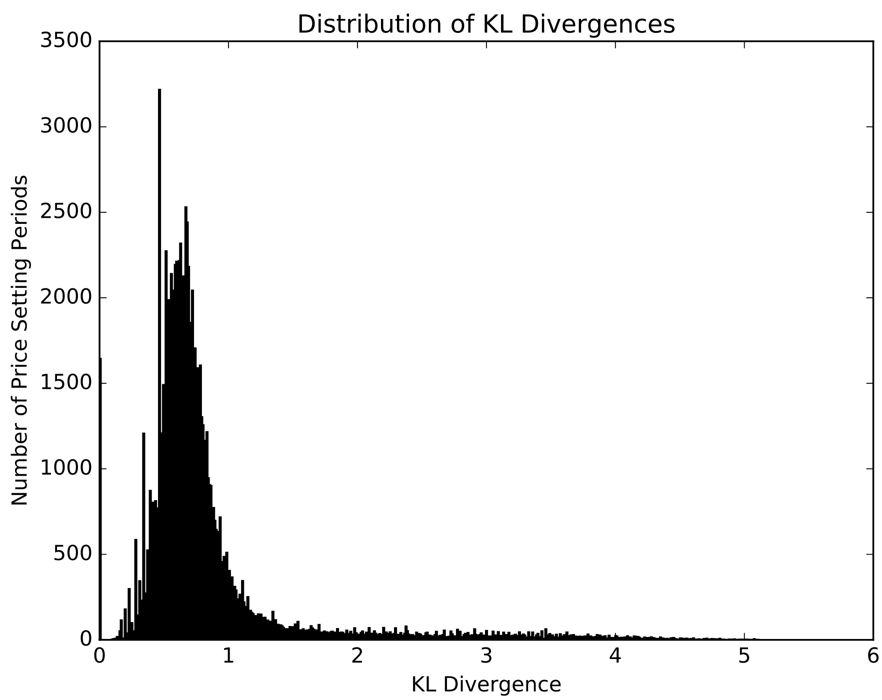


Figure 5.5: Distribution of KL divergences, each measuring how close the price distribution in each price setting period is to the uniform distribution over the set of prices found in that price distribution.

5.2.3 Modal Prices Within Price Setting Periods

Figure 5.6 plots the distribution of the numbers of modal prices found in "modal PSPs," or price setting periods with modal prices. We observe that most of the PSPs with modal prices have exactly one or two modal prices, with a significant dropoff in the number of modal PSPs having any more modal prices, which is consistent with the assumptions we previously made and trends we see from this dataset. We also see that no more than six modal prices within a single PSP are identified.

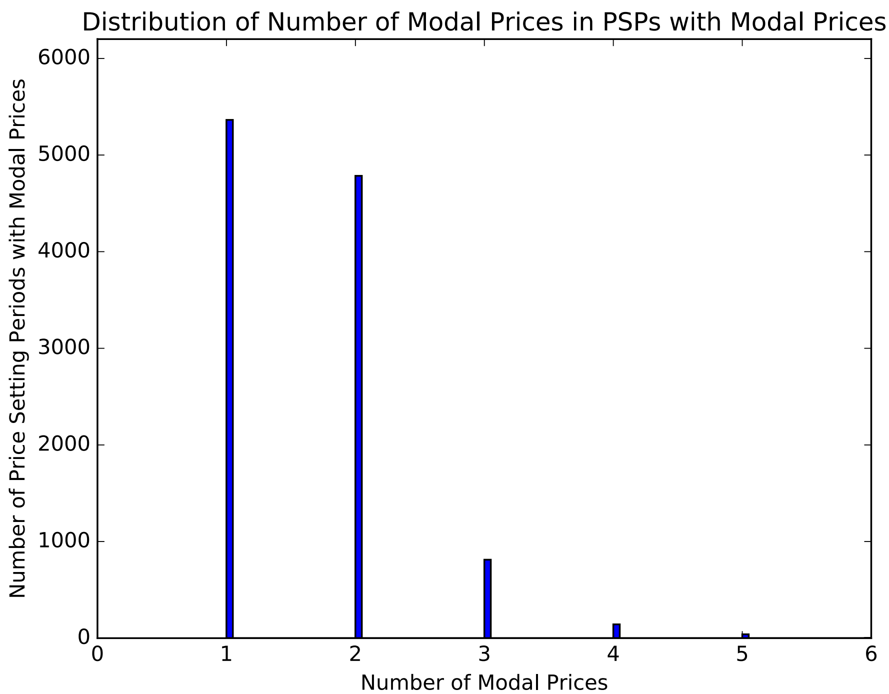


Figure 5.6: We find that 11.2 percent of price setting periods have at least one modal price. For these PSPs with modal prices, we plot the distribution of the number of modal prices found in each PSP.

We now fit the Dirichlet process Gaussian mixture model to all PSPs with modal prices, as described in Chapter 4. Figure 5.7 depicts an example price trend fit to the Dirichlet process Gausian mixture model. Prices in the trend that the model identifies to correspond to the modal price of order 1 are plotted in blue, prices corresponding to the modal price of order 2 are plotted in green, and non-modal prices are shown in red.

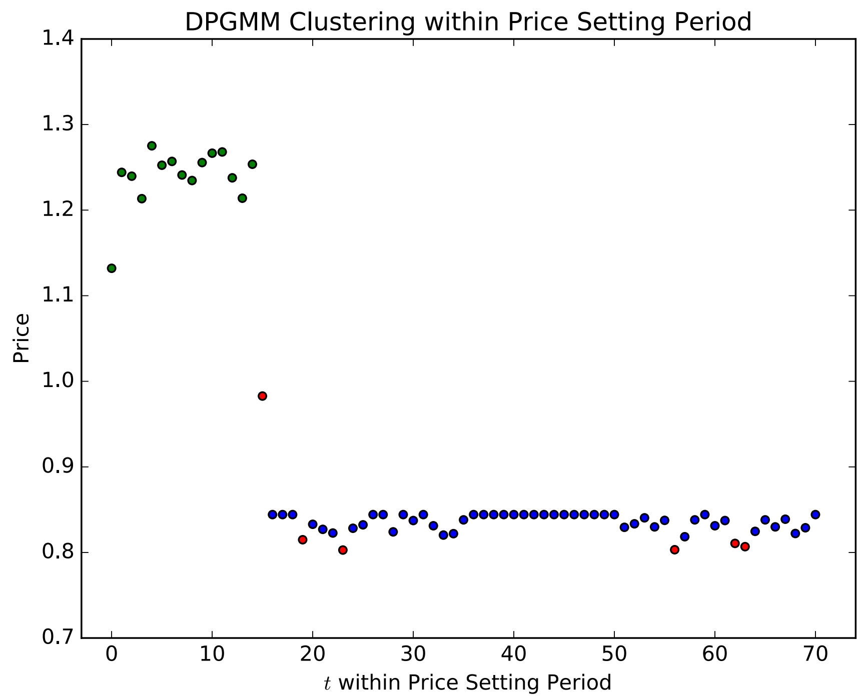


Figure 5.7: Example price trend fit to the Dirichlet process Gausian mixture model. Prices in the trend identified to correspond to the modal price of order 1 are plotted in blue, prices corresponding to the modal price of order 2 are in green, and non-modal prices are in red.

Figure 5.8 tracks the distributions of the fractions of modal PSPs that have modal prices of order i for $i = 1, 2, 3, 4$. For each product, the modal fraction of order i is calculated by dividing the number of PSPs that have a modal price of order i by the total number of PSPs for that product. As expected, a significant portion of modal PSPs have modal fractions of order 1 equal to exactly 1. In other words, the PSPs for a large number of these products that have modal PSPs all have a modal price of order 1. This trend also holds for modal prices of order 2, though to a considerably lesser extent. Furthermore, the fractions of modal PSPs with modal prices of order i decreases with increasing i.

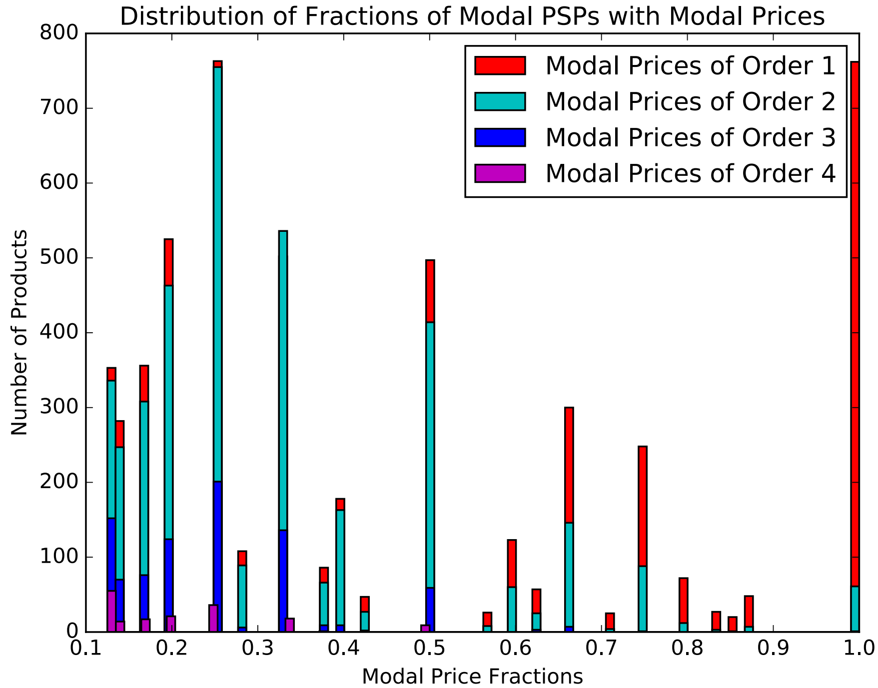


Figure 5.8: Distributions over products of the fractions of price setting periods with modal prices of order i for $i = 1, 2, 3, 4$. For each product, the modal fraction of order i is calculated by dividing the number of PSPs that have a modal price of order i by the total number of PSPs for that product.

5.3 Analysis of Price Stickiness

Our main insights from this section stem directly from analysis of our three price stickiness measures: (1) durations, (2) entropy rates, and (3) $\|\Delta\|_1$-measures for modal prices of different orders. We summarize our key findings below.

1. We find that the durations of most price setting periods are longer than a quarter. In fact, half of the price setting periods have a duration of at least 27 weeks. Nearly 15 percent of PSPs have a duration of at least one year.

2. We find that the entropy rate works well as a measure of price stickiness. The entropy rate has several appealing properties: (a) it successfully identifies clusters of price distributions for price setting periods of different shapes, (b) it is a one-dimensional measure and is thus easily visualizable, and (c) it is easily interpretable as a measure of the amount of randomness in the price distribution.

3. Modal prices are stickier than mean prices.

5.3.1 Durations of Price Setting Periods

Figure 5.9 plots the distribution of durations of price setting periods, where the duration of a PSP is the number of weeks that the PSP spans. We find that half of PSPs are at least half a year in duration, which is significantly longer than the use of quarters by Eichenbaum et al. (2011) and others in which to analyze the internal dynamics of segmented time periods. Eichenbaum et al. (2011) find that the duration of reference prices is nearly a year, while we find that nearly 15 percent of price setting periods are a year or longer in duration. We note that our measure of the duration of a PSP is in fact a lower bound estimate of the true duration of modal prices. As discussed in the first section of this chapter, our model sometimes identifies too many PSPs, and so the true duration would be the sum of the durations for multiple PSPs. Thus, we find evidence supporting a high level of rigidity at the micro price level.

5.3.2 Entropy Rates of Price Setting Periods

Figure 5.10 illustrates the distribution of the entropy rates over all PSPs along with a few representative examples with entropy rates from different parts of the distribution. As expected, the PSP with prices that do not change at all has an entropy rate of 0, while the PSP with prices that oscillate between a small set of prices has one of the highest entropy rates found across all PSPs. Furthermore, the number of modes is increasing in the entropy rate: the example with an entropy rate of 0 naturally has one modal price. Notably, over $3,500$ price setting periods have an entropy rate that is close to 0, suggesting that these PSPs have very well-defined reference prices. The example trends with entropy rates of 0.36, 0.78, and 1.34 have two modal prices. We observe a bump in the distribution of entropy rates across PSPs around 0.8 to 1.0, which suggests that a high number of PSPs are multi-modal in nature. Finally, the PSP with an entropy rate of 1.91 has four modal prices. In these examples, the

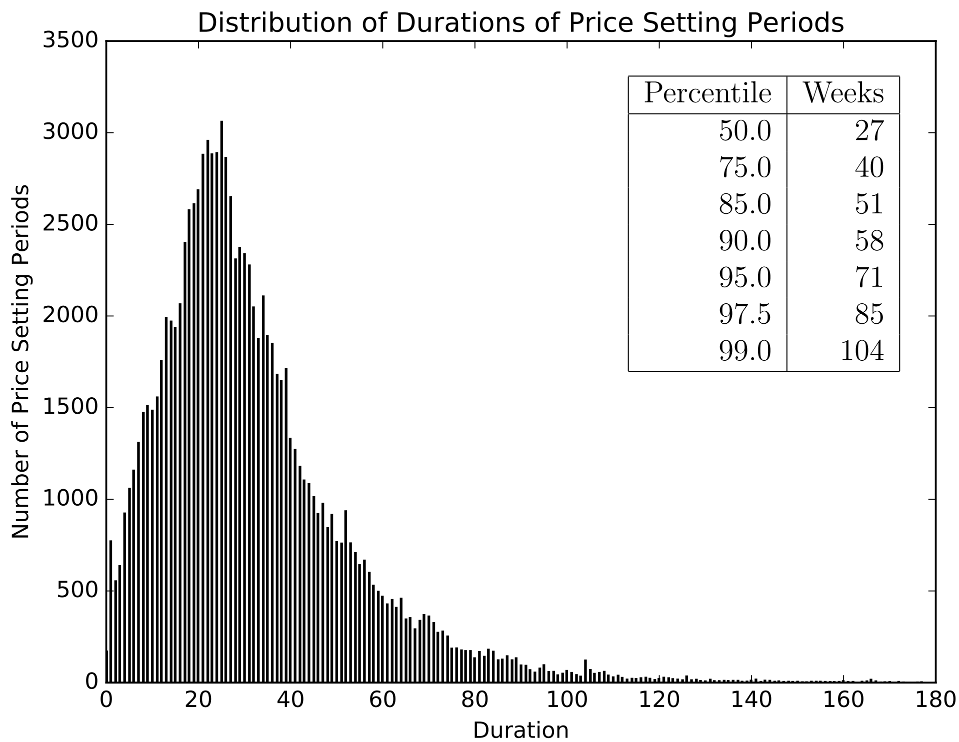


Figure 5.9: Distribution of the durations of price setting periods in weeks.

reference price, or modal price with order 1, generally shows up throughout the length of the PSP.

Clustering Using Entropy Rate

Figures 5.11 and 5.12 show several example results from fitting price setting periods to a Dirichlet process Gaussian mixture model, where the first cluster of results all have low entropy rates between 0.31 and 0.34 and the second cluster of results all have high entropy rates between 1.90 and 1.99. We observe that the first cluster is entirely characterized by a small set of modal prices, with only a small number of points that are not found elsewhere in the trend. Furthermore, we note that each example from the second set is a violently fluctuating trend that we discussed in the first section of this chapter—in other words, each example in this cluster oscillates

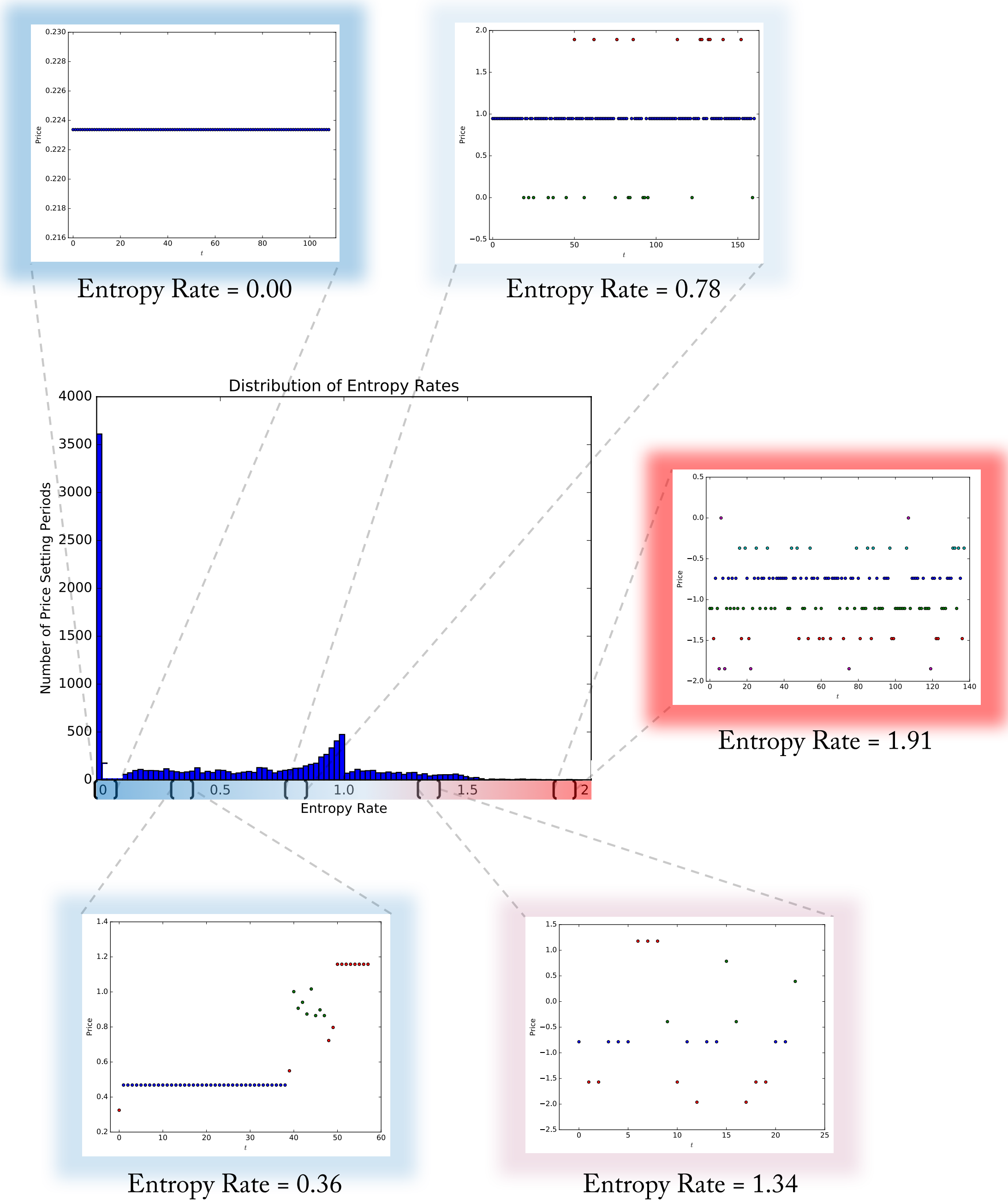


Figure 5.10: Distribution of the entropy rates over all price setting periods. We illustrate a few examples of price setting periods, each with the entropy rate of its price distribution.

from week to week between a set of four or five modal prices.

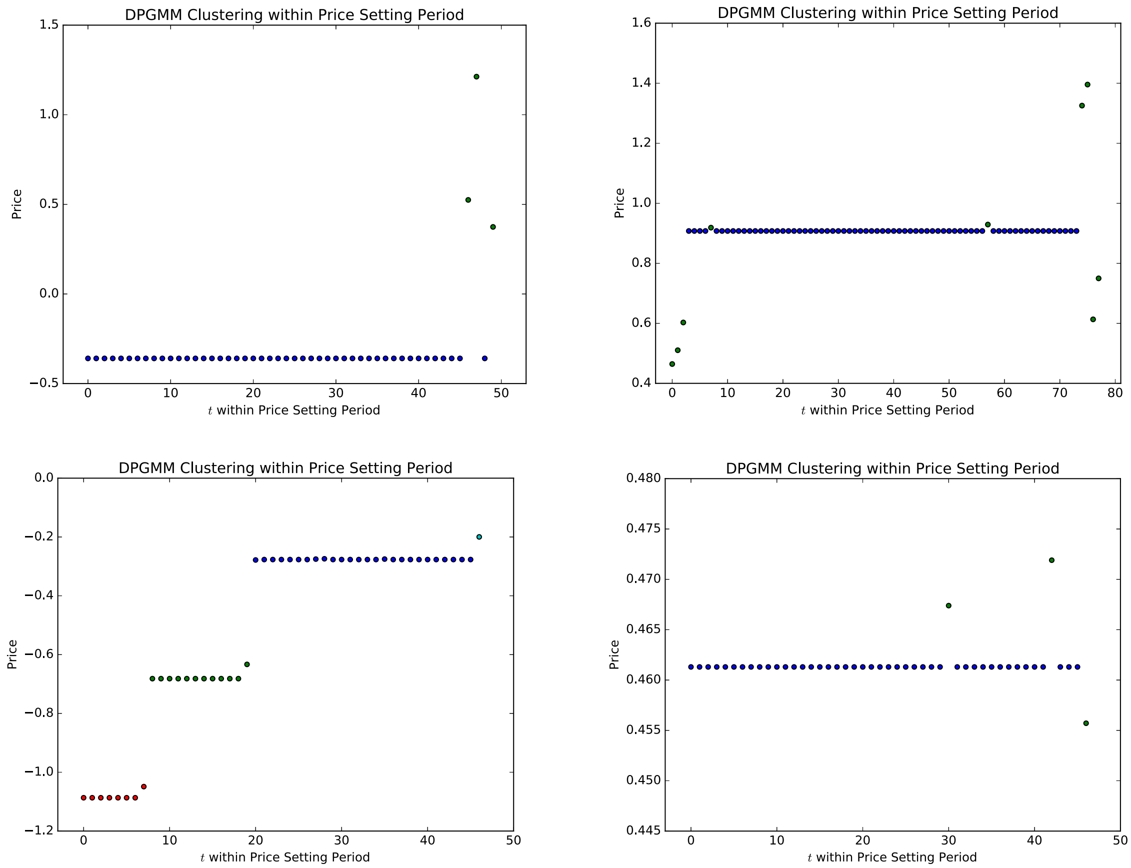


Figure 5.11: The results of fitting various price setting periods to a Dirichlet process Gaussian mixture model. These examples each have a low entropy rate between 0.31 and 0.34.

5.3.3 Modal Prices Are Persistent

Figure 5.13 plots the distribution of the $\|\Delta\|_1$-measures for mean prices along with the distribution of the $\|\Delta\|_1$-measures for modal prices of order 1, both across all PSPs. We observe that the distribution of $\|\Delta\|_1$-measures for modal prices of order 1 has more mass in the range of low values than the distribution of $\|\Delta\|_1$-measures for mean prices. Since $\|\Delta\|_1$ sums the magnitudes of changes between the modal or mean prices for each PSP from PSP to PSP, this result suggests that modal prices are stickier than mean prices.

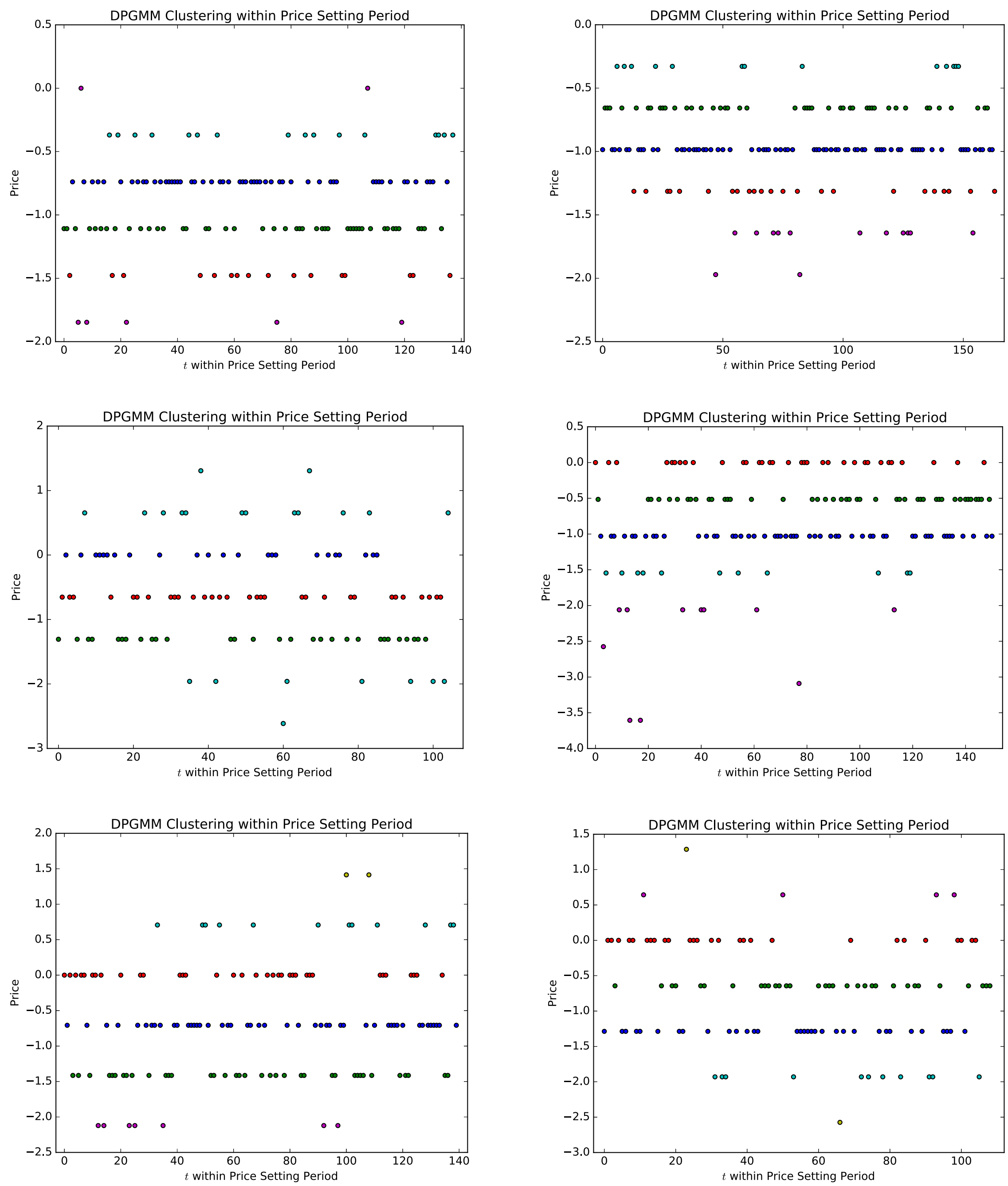


Figure 5.12: The results of fitting various price setting periods to a Dirichlet process Gaussian mixture model. These examples each have a high entropy rate between 1.90 and 1.99.

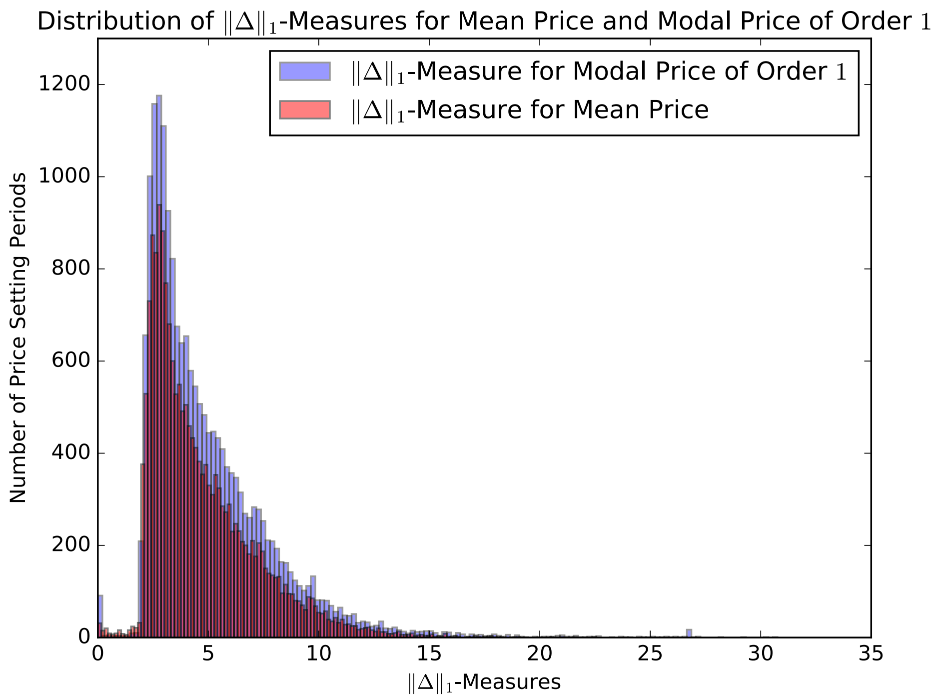


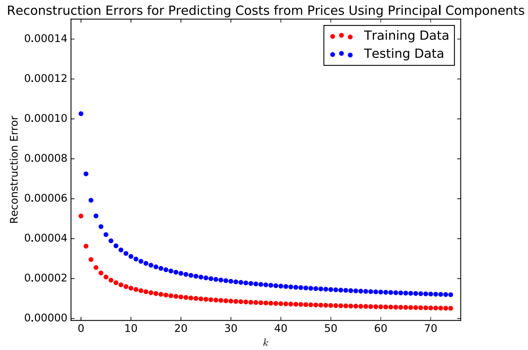
Figure 5.13: Distributions of $\|\Delta\|_1$-measures for mean prices and for modal prices of order 1 across price setting periods. Modal prices appear to be stickier than mean prices.

5.4 How Well Do Costs Predict Prices?

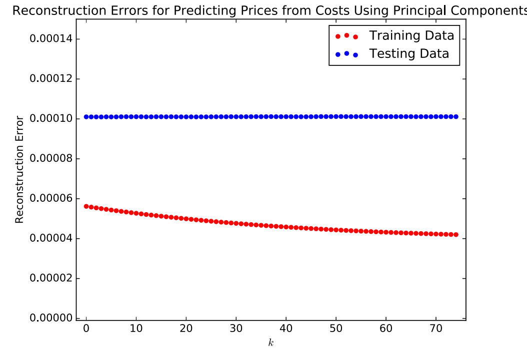
Figure 5.14 plots the reconstructions errors for predicting costs from prices and for predicting prices from costs, where the reconstruction error is a measure of the average mean squared error over the elements of the reconstructed data matrix. Plots (a) and (b) illustrate the results from when we fit the model on only the principal components identified for the price data and for cost data, while plots (c) and (d) are results from fitting the model on all trends. We observe that the reconstruction errors generally decrease in k, the number of previous time steps that we use in the model. We also note that k is a measure of model complexity: the larger k is, the more parameters we need to fit the model, since the matrix W of map weights increases linearly in k. Our result implies that using more time steps tends to improve the model, suggesting that a delay exists between costs and prices. Furthermore, the reconstruction errors for using our model to predict training data are generally lower than the errors for predicting testing data. Surprisingly, the models for predicting prices from costs seem to perform slightly better than the models for predicting costs from prices.

A possible reason for these results may be due to the relative variance of the price and cost values. We refer back to Figures 3.8 and 3.10, the plots of the distributions of the mean prices and mean costs, respectively, over all weeks. We can see that the distribution of costs over all products is much more peaked and is thus lower variance than the price distribution; thus, price may better predict cost simply because the cost values have less variance than the prices. Empirically, this statement seems to be true.

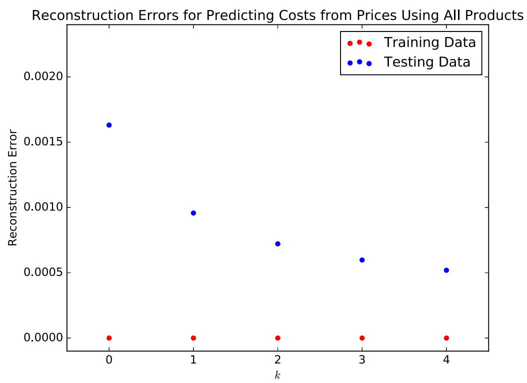
Overall, we can see that the learned maps generalize (the testing errors decrease alongside the decreasing training errors). The exception is with plot (b), where the reconstruction errors for testing data do not improve with increasing k, while reconstruction errors for training data decrease slightly. This result could be a sign of overfitting in this case: model complexity increases, but only training errors improve.



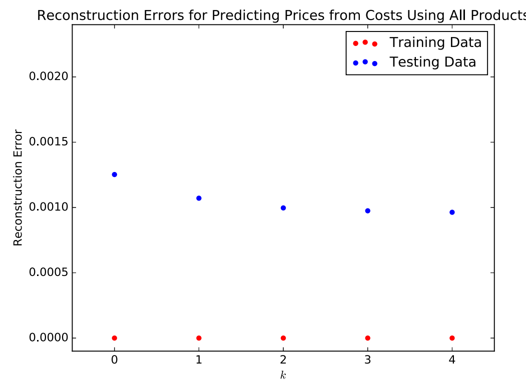
(a) Predicting costs from prices using principal components



(b) Predicting prices from costs using principal components



(c) Predicting costs from prices using all data



(d) Predicting prices from costs using all data

Figure 5.14: Reconstruction errors of the models predicting costs from prices and prices from costs, plotted against k, the number of previous time steps used in prediction. Plots (a) and (b) depict the results of fitting the model to only the principal components for costs and for prices. Plots (c) and (d) depict the results of fitting the model to all data.

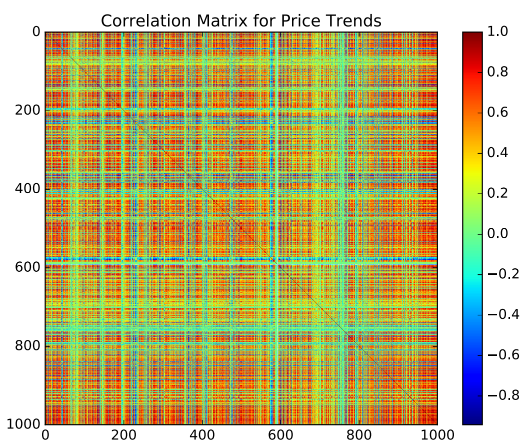
5.5 Product Clusters

We discuss the results that we obtain by clustering products by two different measures: according to time trends and according to price stickiness. In both cases, we find that products naturally cluster.

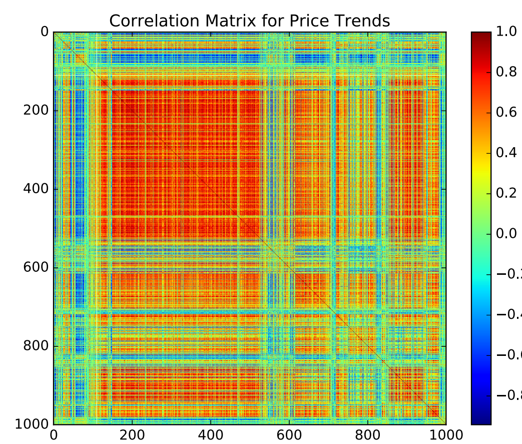
5.5.1 Clustered by Trends

One method of clustering the products is to perform principal component analysis on all of the price trends. We find that $d = 102$ principal components are sufficient for capturing 95 percent of the variance in the price data across all products (nearly $20,000$ products in total).

We then perform d-nearest neighbor clustering, where the d clusters are determined based on Euclidean distance to the d principal components. In other words, each product is assigned to the closest principal component. Figure 5.15 displays correlation matrices after sorting by the assigned category and sorting by the closest principal component, respectively—notably, we label each cluster with the rank of its associated principal component and then sort the products by increasing cluster label order. We observe higher correlation among clusters formed by the top principal components and, in general, more clustered products. We also find that 137 principal components are sufficient for capturing 95 percent of the variance in the cost data across 1,000 random products, as shown in Figure 5.16.



(a) Sorted by category

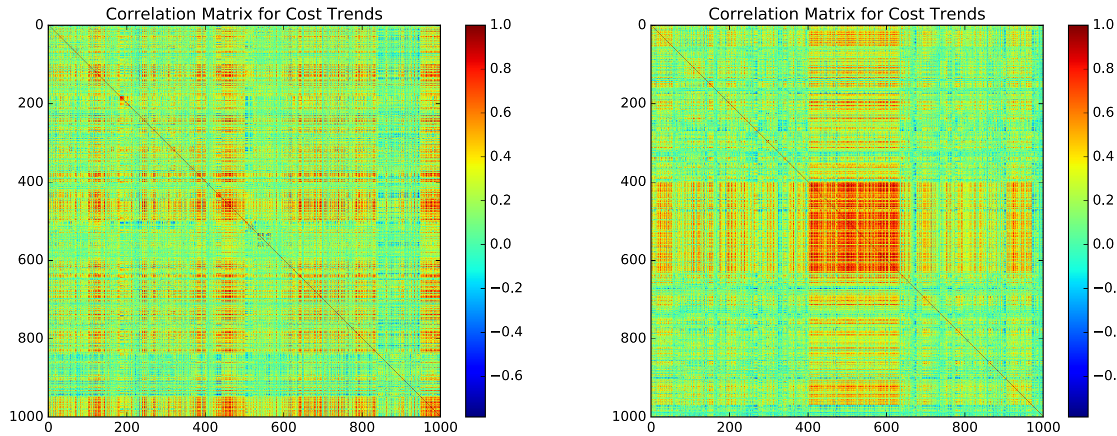


(b) Sorted by principal component

Figure 5.15: Matrix of correlations between the price trends for pairs of products from a set of 1,000 random products, first organized by category and then by principal component. Sorting by principal component identifies similar products better than sorting by product category.

5.5.2 Clustered by Price Stickiness

Figure 5.17 displays two clusters that are identified when we use price stickiness measures to characterize products. We observe a clear distinction in the magnitudes



(a) Sorted by category **(b)** Sorted by principal component

Figure 5.16: Matrix of correlations between the cost trends for pairs of products from a set of 1,000 random products, first organized by category and then by principal component. Sorting by principal component identifies similar products better than sorting by product category.

of the products' ℓ_2-norms of entropy rates for different price setting periods—namely, products that are colored red in Figure 5.17 have ℓ_2-norm values of higher than 6, while products corresponding to the blue cluster have ℓ_2-norms that are less than 6.

Figure 5.18 plots two examples, one drawn from the red cluster and the second drawn from the blue cluster. Both of these trends have a small $\|\Delta\|_1$-measure for the reference price; however, the ℓ_2-norm of entropy rates is drastically contrasted.

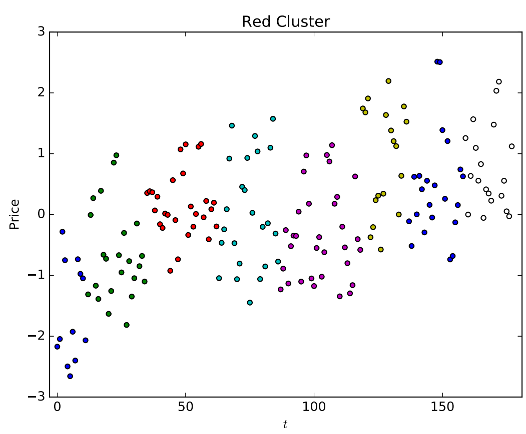
It is important to keep in mind that the ℓ_2-norm measure is more of an internal metric to characterize price distributions within PSPs than the $\|\Delta\|_1$-measures for reference prices. Thus, we are able to capture both internal and cross-PSP expressions of stickiness. Interestingly, a few products have high reference price stickinesses but low ℓ_2-norms of entropy rates. This result suggests that a distinction exists between measures of stickiness applied to entire trends and measures based on internal dynamics of PSPs.

We find that the $\|\Delta\|_1$-measure for reference prices is highly sensitive to outliers that sometimes occur in trends. For instance, there are trends that remain at a single price through all weeks except a small number of weeks. Since we standardize our data, these outliers are amplified, and the distances between these outliers and the mode price become large. These examples generally show up in the right half of 5.17. This problem is mitigated by the fact that we, nevertheless, see that most products are highly concentrated around small values for $\|\Delta\|_1$-measure for reference prices.

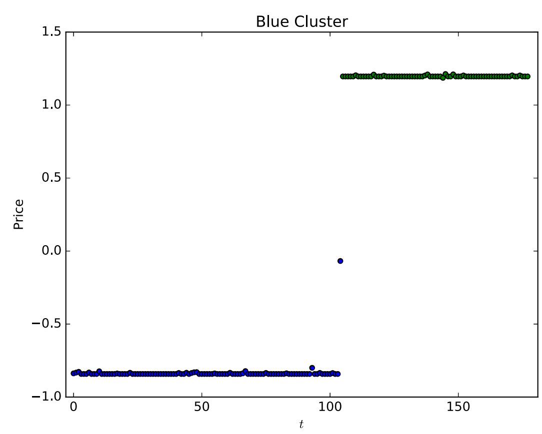
Importantly, even though most of the mass in the distribution of $\|\Delta\|_1$-measures for reference prices is uniformly concentrated, roughly between 0 and 10, there is an easily detectable split between the two clusters, identifiable using a threshold for the entropy rate. Furthermore, the two clusters that are identified have roughly equal mass. This result confirms that prices are sticky indeed, but critically: our methodology helps us distinguish products with highly sticky prices from products with less sticky prices by looking at internal dynamics of price setting periods.

Both the $\|\Delta\|_1$-measure for reference prices and the ℓ_2-norm of enropy rates for PSPs inform us of the stickiness of products. Nevertheless, the ℓ_2-norm of entropy rates is a stronger measure: we can distinguish very sticky products from less sticky products simply by examining this one metric, which captures the internal dynamics of price setting periods.

Figure 5.17: Product features are defined as ($\|\Delta\|_1$-measure for reference prices, ℓ_2-norm of entropy rates for PSPs). We identify clusters of products by thresholding the entropy rate feature.



(a) Example from the red cluster



(b) Example from the blue cluster

Figure 5.18: Example trends drawn from the clusters identified using price stickiness measures. Both trends have a low $\|\Delta\|_1$-measure for reference prices, but trend (a) has a significantly higher ℓ_2-norm of entropy rates over price setting periods.

Chapter 6

Conclusion

6.1 Implications

Our paper has focused on algorithmically identifying price setting periods for any given product and characterizing a set of various regular prices within price setting periods, given only the values of the price time series. Our work departs from the existing literature in three important ways. For each product,

1. We flexibly identify price setting periods without making any assumptions on the length of the periods or the number of periods.

2. We do not simply study one regular, or reference, price. Instead, we identify several "modal" prices from the data that correspond to the peaks of the price distributions for the identified price setting periods.

3. We develop highly clusterable and interpretable metrics for price stickiness, which we demonstrate with an easily separable clustering of the products into two clean groups of highly sticky and non-sticky products.

Thus, we succeed in answering the questions that we laid out in Chapter 4.

6.2 Future Work

There are many possible extensions and continuations of our work.

6.2.1 Further Analysis of Product and Price-Setting-Period Clusters

Future research in this area can extend our work by attempting to identify well-known macroeconomic trends such as price seasonality by examining price setting periods. Furthermore, products that behave seasonally could also potentially be uncovered using our methodology as a starting point. We could further analyze the set of interesting prices that we learn for each PSP and try to identify which of those prices are sales; we can generally attempt to explain the source of each of the interesting prices we discover with our Dirichlet process Gaussian mixture model.

We could also attempt to characterize products and price setting periods by including costs in the equation, thus using both price and cost trends in the modeling.

6.2.2 Fine-Tuning Our Probabilistic Methods

Our Bayesian models work quite well at identifying an interesting set of regular prices found over PSPs. However, we can tweak our setup to potentially improve performance even more.

We essentially use two separate probabilistic models over the course of this thesis: an event-segmentation hidden Markov model and a Dirichlet process Gaussian mixture model. We use the models for two separate tasks—identifying price setting periods and finding regular prices. Yet, there is no reason to do these tasks separately; We could instead build a probabilistic model, where the event-segmentation HMM is another layer on top of the DPGMM. Essentially, we could condition on the state of the event-segmentation HMM before outputting a price from the DPGMM model and have one complete model to encapsulate both parts of our analysis.

Another modification that we might consider adding to our probabilistic model is a layer that allows us to share information across products; namely, our model could additionally condition on the product that it came from. Since we empirically know that products cluster well, we could insert assumptions into the model that products come from different clusters, and we could potentially implement something similar to a Chinese Restaurant Franchise: a hierarchical extension of the DPGMM (also known as a Chinese Restaurant Process) that we use in this thesis. This step would allow us to introduce more information into the model, while potentially making it more complicated to fit.

6.2.3 Connections to Macroeconomic Theory

In the vein of most papers in the literature, we could try to integrate our probabilistic models more thoroughly in an established macroeconomic price model framework such as the menu-cost, Calvo, flexible price, or multisector model. We could then compare the predictions of our theoretical model with actual macroeconomic price trends to see how well we perform at predicting true aggregate price movements. One potential difficulty would be that our dataset covers a relatively uninteresting and short period of time, macroeconomically speaking (2004—2007). Thus, performing this kind of analysis on a larger dataset of similar flavor could potentially produce more interesting intuition into the macroeconomic trends underlying price dynamics.

Appendix A

Statistical Methods

A.1 Hidden Markov Model

A hidden Markov model (HMM) is a model of a probabilistic process that evolves over discrete time steps. At any given point in time, an HMM outputs a value conditioned upon the current unobserved "hidden" state that it is in. Learning the output distribution of values given the hidden states and learning the probabilistic transition matrix between hidden states fully characterizes the model.

In the simpler Markov models, such as the Markov chain generated by Eichenbaum et al. (2011), the states are directly visible—{price = reference price, price $\neq$ reference price}; thus, the state transition probabilities are the only parameters to learn. Note that we cannot directly encode any information about sales in this type of model. However, in a hidden Markov model, the states are not directly visible, though the output states, which depend on the states, are visible. For example, we cannot tell directly from the data that at a given time, a given product is on sale, but we are able to learn this unobserved information with an HMM.

Figure A.1 demonstrates an example hidden Markov model, where outputs are weekly prices. T denotes the transition matrix and O denotes the output probability matrix. In this diagram, there are four hidden states: A, B, C, and D.

Generally, $T \in \mathbb{R}^{m \times m}$ is the transition probability matrix, and $O \in \mathbb{R}^{m \times n}$ is the output probability matrix, where m is the number of hidden states and n is the number of output states. (In the toy example, $m = 4$ and $n = 4$.) For instance, our hidden states here could correspond to "being in a sale," "at a regular price," or "at a temporary price increase."

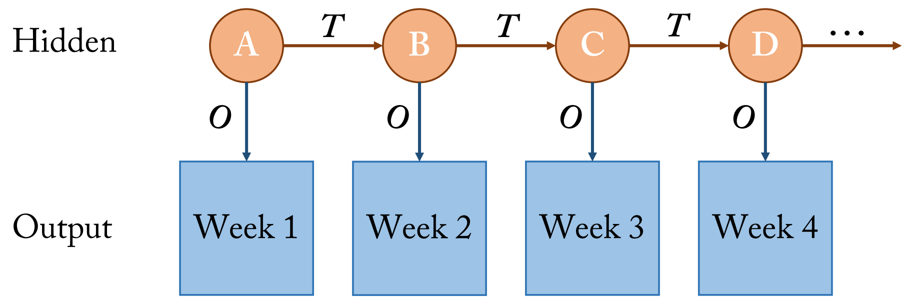


Figure A.1: An example hidden Markov model, where outputs are weekly prices. T denotes the transition matrix and O denotes the output probability matrix. In this diagram, there are four hidden states: A, B, C, and D.

A.2 Principal Component Analysis

We cover the dimensionality reduction technique of principal component analysis (Ramadge 2016). We ask: given data $\{x_j \in \mathbb{R}^n\}_{j=1}^p$, is there some subspace $\mathcal{U}$, with dimension $q < n$, such that if we orthogonally project each data point down to this subspace, then the errors we incur will be quite small? In other words, we want to find the "best" subspace $\mathcal{U}$ with dimension q with the minimum sum of squared norms of the error residuals over all possible subspaces of the vector space.

We form an orthornomal basis $\{u_1, \dots, u_q\}$ for $\mathcal{U}$ and form $U = \begin{bmatrix} u_1 & u_2 & \dots & u_q \end{bmatrix} \in \mathbb{R}^{n \times q}$ with $U^T U = I_q$. We have two cases:

1. $q = n$: $U \in \mathcal{O}_n$.

2. $q < n$: $U \in \mathcal{O}_{n \times q}$.

The projection of x_j, the j^{th} (centered) data point, onto $\mathcal{U}$ is given by $\hat{x}_j = UU^T x_j \in \mathbb{R}^n$; while this projection does not reduce the dimension of the data, it maps data points onto a q-dimensional subspace in $\mathbb{R}^n$. The error residual of this projection is given by $x_j = \hat{x}_j + r_j$. We aim to replace x_j with $\hat{x}_j$ and suffer the loss of information via r_j. Note that instead of working with $\hat{x}_j$, we can equivalently work with its coordinates $c_j \in \mathbb{R}^q$ with respect to the basis U, since $\hat{x}_j = Uc_j$. There is no loss of information when we work with c_j instead of $\hat{x}_j$ since $c_j = U^T \hat{x}_j = U^T x_j$. So now we have $\hat{x}_j = Uc_j, c_j \in \mathbb{R}^q$. To summarize:

$$x_j \in \mathbb{R}^n \xrightarrow{UU^T x_j} \hat{x}_j \in \mathbb{R}^n \xrightarrow{U^T \hat{x}_j} c_j \in \mathbb{R}^q.$$

$$x_j \xleftarrow{r_j + \hat{x}_j} \hat{x}_j \xleftarrow{Uc_j} c_j.$$

This technique is referred to as linear dimensionality reduction. Other methods exist for selecting the subspace $\mathcal{U}$ in different ways.

We now discuss how to go about choosing q and its corresponding subspace. Our first step is to center the data. We subtract the sample mean $\mu = \frac{1}{p}\sum_{i=1}^{p} x_j$ from each x_j to form $z_j = x_j - \mu$. In matrix form, we have, prior to centering,

$$X = \begin{bmatrix} x_1 & x_2 & \dots & x_p \end{bmatrix} \in \mathbb{R}^{n \times p}.$$

The sample mean is

$$\mu = \frac{1}{p}(X\mathbf{1}_p),$$

where $\mathbf{1}_p \in \mathbb{R}^p$ is a vector of all 1's. The new matrix of centered data points is

$$Z = \begin{bmatrix} z_1 & z_2 & \dots & z_p \end{bmatrix} = X - \mu\mathbf{1}_p^T = X\left(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p^T\right) = X(I_p - uu^T).$$

From this point forward, we assume that the data have been centered.

Now, we need to parametrize $\mathcal{U}$; in other words, we now fix q. Pick an orthornomal basis $U = \begin{bmatrix} u_1 & u_2 & \dots & u_q \end{bmatrix} \in \mathcal{O}_{n \times q}$. Note that this orthonormal basis representation is not unique, since there are infinitely many orthonormal bases for $\mathcal{U}$. Take $U_1, U_2 \in$

$\mathcal{O}_{n\times q}$, which both span $\mathcal{U}$. We can write $U_1 = U_2 Q$ since they span the same subspace, where $Q \in \mathbb{R}^{q\times q}$. Thus, $U_2^T U_1 = Q$.

$$Q^T Q = U_1^T (U_2 U_2^T) U_1 = I_q,$$

$$Q Q^T = U_2^T (U_1 U_1^T) U_2 = I_q,$$

since $U_2 U_2^T$ and $U_1 U_1^T$ are projection matrices, so $U_1 U_1^T = U_2 U_2^T$. Thus, $Q \in \mathcal{O}_q$, and any two orthonormal basis representations U_1, U_2 of $\mathcal{U}$ are related via a $q \times q$ orthogonal matrix Q: $U_1 = U_2 Q$ and $U_2 = U_1 Q^T$.

We examine how the data are spread around $\mathbf{0}$. Select a unit norm vector $u \in \mathbb{R}^n$ and project x_j onto the line through $\mathbf{0}$ in the direction u, yielding $\hat{x}_j = uu^T x_j,\ j = 1, \ldots, p$. Since the direction is fixed to be u, the projected data are specified by the set of scalars $u^T x_j$ with zero sample mean. The sample variance in direction u is

$$\begin{aligned}
\sigma^2(u) &= \frac{1}{p}\sum_{j=1}^{p}(u^T x_j)^2 \\
&= \frac{1}{p}\sum_{j=1}^{p} u^T x_j x_j^T u \\
&= u^T \left(\frac{1}{p}\sum_{j=1}^{p} x_j x_j^T\right) u \\
&= u^T R u,
\end{aligned}$$

where we define the sample covariance matrix as $R := \frac{1}{p}\sum_{j=1}^{p} x_j x_j^T$, a symmetric positive semidefinite $n \times n$ matrix. Thus, the variance of the data in direction u is

$$\sigma^2(u) = u^T R u.$$

It may be interesting to observe the directions of largest variance of the data, for they capture most of the variability in the data. The direction of maximum variance is

$$u_1 = \arg\max_{\|u\|=1} u^T R u,$$

where u_1 is a unit norm eigenvector corresponding to the largest eigenvalue σ_1 of R. Let us suppose that we want to find a second direction with the second largest variance. We do not want it to be arbitrarily close to u_1, so we can constrain the second direction to be orthogonal to the first. For q orthogonal directions, then, we want to find $U = \begin{bmatrix} u_1 & \dots & u_q \end{bmatrix} \in \mathcal{O}_{n \times q}$ to maximize $\sum_{j=1}^{q} u_j^T R u_j = \text{Tr}(U^T R U)$. The solution is attained by taking the q directions to be unit norm eigenvectors $u_1, \dots, u_q$ for the largest q eigenvalues of R. Generalizing, we can see that we obtain n orthonormal directions of maximum variance in the data: these directions $v_1, \dots, v_n$ and the corresponding variances $\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_n^2$ are eigenvectors and corresponding eigenvalues of R, with $R v_j = \sigma_j^2 v_j, \ j = 1, \dots, n$. The vectors v_j are called the principal components of the data, and this decomposition is called principal component analysis (PCA). Let V be the matrix with the v_j's as its columns and $\Sigma^2 = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$. PCA is an ordered eigendecomposition of the sample covariance matrix $R = V \Sigma^2 V^T$.

We can quickly see that this interpretation of PCA is the same as the previous approach of finding a subspace that minimizes the sum of squared norms of the residuals. We can write

$$R = \frac{1}{p} \sum_{j=1}^{p} x_j x_j^T = \frac{1}{p} X X^T = \frac{1}{p} U \Sigma V^T V \Sigma U^T = \frac{1}{p} U \Sigma^2 U^T.$$

So the principal components are simply the eigenvectors of XX^T listed in order of decreasing eigenvalues, and in particular, the first q principal components are the first q eigenvectors of XX^T, which define the orthonormal basis U^* and, thus, an optimal q-dimensional projection subspace $\mathcal{U}^*$. Thus, the leading q principal components give a particular orthonormal basis for an optimal q-dimensional projection subspace.

We can form $V_q = \begin{bmatrix} v_1 & \dots & v_q \end{bmatrix}$, which defines a subspace capturing the q directions of largest variance. We can project the data onto the span of the columns of V_q: $\hat{x}_j = V_q(V_q^T x_j)$, where $c_j = (V_q^T) x_j$ gives the coordinates of x_j with respect to V_q, and $V_q c_j$ synthesizes $\hat{x}_j$ using these coefficients to form the appropriate linear combination of the columns of V_q. Thus, PCA is essentially looking around to find orthogonal directions of maximum variance, assuming that maximum variance is important.

References

Anderson, Eric, Benjamin A. Malin, Emi Nakamura, Duncan Simester, and Jón Steinsson, "Informational Rigidities and the Stickiness of Temporary Sales," *Working Paper, Columbia University*, 2016.

Baldassano, Christopher, Janice Chen, Asieh Zadbood, Jonathan W. Pillow, Uri Hasson, and Kenneth A. Norman, "Discovering Event Structure in Continuous narrative perception and memory," *bioRxiv*, 2016.

Bhattarai, Saroj and Raphael Schoenle, "Multiproduct Firms and Price-Setting: Theory and Evidence from U.S. Producer Prices," *Journal of Monetary Economics*, 2014, *66* (C), 178–192.

Bils, Mark and Peter J. Klenow, "Some Evidence on the Importance of Sticky Prices," *Journal of Political Economy*, 2004, *112* (5), 947–985.

Blei, David M. and Michael I. Jordan, "Variational Inference for Dirichlet Process Mixtures," *International Society for Bayesian Analysis*, 2004, *1* (1).

Blinder, Alan S., *Asking About Prices: A New Approach to Understanding Price Stickiness*, Russell Sage Foundation, 1998.

Campbell, Jeffrey R. and Benjamin Eden, "Rigid Prices: Evidence from U.S. Scanner Data," *International Economic Review*, 2014, *55* (2), 423–442.

Carvalho, Carlos, "Heterogeneity in Price Stickiness and the Real Effects of Monetary Shocks," *Frontiers of Macroeconomics*, 2006, *2* (1).

Chahrour, Ryan A., "Sales and Price Spikes in Retail Scanner Data," *Quarterly Journal of Economics*, 2011, *110* (2), 143–146.

Chevalier, Judith A., Anil K Kashyap, and Peter E. Rossi, "Why Don't Prices Rise During Periods of Peak Demand? Evidence from Scanner Data," *American Economic Review*, 2003, *93* (1), 15–37.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans, "Monetary Policy Shocks: What Have We Learned and to What End?," *Handbook of Macroeconomics*, 1999, *1* (A), 65–148.

Dutta, Shantanu, Mark Bergenb, and Daniel Levy, "Price Flexibility in Channels of Distribution: Evidence from Scanner Data," *Journal of Economic Dynamics and Control*, 2002, *26* (11), 18451900.

Eichenbaum, Martin, Nir Jaimovich, and Sergio Rebelo, "Reference Prices, Costs, and Nominal Rigidities," *American Economic Review*, 2011, *101* (1), 234–262.

_ , _ , _ , and Josephine Smith, "How Frequent Are Small Price Changes?," *American Economic Journal: Macroeconomics*, 2014, *6* (2), 137–155.

Frigyik, Bela A., Amol Kapila, and Maya R. Gupta, "Introduction to the Dirichlet Distribution and Related Processes," *UWEE Technical Report*, 2010, (UWEETR-2010-0006).

Görür, Dilan and Carl Edward Rasmussen, "Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution," *Journal of Computer Science and Technology*, 2010, *25* (4), 615–626.

Guimaraes, Bernardo and Kevin D. Sheedy, "Sales and Monetary Policy," *American Economic Review*, 2011, *101* (2), 844–876.

Kehoe, Patrick and Virgiliu Midrigan, "Prices Are Sticky after All," *Journal of Monetary Economics*, 2015, pp. 35–53.

Klenow, Peter J. and Oleksiy Kryvtsov, "State-Dependent Or Time-Dependent Pricing: Does It Matter for Recent U.S. Inflation?," *Quarterly Journal of Economics*, 2008, *123* (3), 863–904.

Kumar, V. and Robert P. Leone, "Measuring the Effect of Retail Store Promotions on Brand and Store Substitution," *Journal of Marketing Research*, 1988, *25* (2), 178–185.

Midrigan, Virgiliu, "Menu Costs, Multiproduct Firms, and Aggregate Fluctuations," *Econometrica*, 2011, *79* (4), 11391180.

Nakamura, Emi and Dawit Zerom, "Accounting for Incomplete Pass-Through," *Review of Economic Studies*, 2009, *77*, 1192–1230.

_ **and Jón Steinsson**, "Five Facts about Prices: A Reevaluation of Menu Cost Models," *Quarterly Journal of Economics*, 2008, pp. 1415–1464.

_ **and** _ , "Price Rigidity: Microeconomic Evidence and Macroeconomic Implications," *Annual Review of Economics*, 2013, *5*, 133–163.

_ **and** _ , "Fiscal Stimulus in a Monetary Union: Evidence from U.S. Regions," *American Economic Review*, 2014, *104* (3), 753–792.

Ramadge, Peter J., "ELE 535 Lecture," *Princeton University*, 2016.

Ramey, Valerie A., "Can Government Purchases Stimulate the Economy?," *American Economic Review*, 2011, *49* (3), 673–685.

Romer, Christina D. and David H. Romer, "A New Measure of Monetary Shocks: Derivation and Implications," *American Economic Review*, 2004, *94* (4), 1055–1084.

Stratos, Karl, Michael Collins, and Daniel Hsu, "Unsupervised Part-Of-Speech Tagging with Anchor Hidden Markov Models," *Transactions of the Association for Computational Linguistics*, 2016, *4*, 245257.

Walters, Rockney G., "Assessing the Impact of Retail Price Promotions on Product Substitution, Complementary Purchase, and Interstore Sales Displacement," *Journal of Marketing*, 1991, *55* (2), 17–28.