Consumer Heterogeneity Modeling by Hierarchical Gaussian Process

Xiaoting Sun[1], Alan Montgomery[1, 2], and Zachary C. Lipton[1,2]

[1] *Machine Learning Department, Carnegie Mellon University*
[2] *Tepper Business School, Carnegie Mellon University*

December 16, 2018

1 Introduction

Consumer's behaviors vary a lot across individuals: for instance, people tend to have different price sensitivities to the same good and tend respond differently to promotions. It is called consumer heterogeneity in the field of marketing and plays a central role in many marketing activities. For example, pricing and product design decisions are based on an understanding of the differences among consumers in price sensitivity and valuation of product attributes(Allenby & Rossi, 1998; Montgomery, 1997). Therefore, to appropriately model the consumer heterogeneity is of great importance. However, understanding the relationship between consumer behaviors and marketing variables is challenging: first of all, the individual-level data are sparse thus to model each consumer's behavior separately tends to overfit. Although nowadays we can collect and keep large collections of consumer behaviors' data, the relevant data associated with a specific study are limited. Secondly, the form of the relationship itself can be complicated. For example, even we know that demand decreases as price increases, the exact form of this relationship can be involved. The effect of price in demand might be weak unless the price increases to a certain threshold. Such complicated interactions would pose a challenge of specifying the form of the model.

To deal with these two challenges, we propose a hierarchical Gaussian process method. Although consumer behaviors vary a lot, people have similar demographic features such as gender, and income may share some similarities in their purchasing behaviors. Utilizing the hierarchical structure would allow to share information across different individuals. To address the second problem, we introduce the Gaussian process. It defines prior distributions for flexible models in which the regression or class probability functions are not limited to simple parametric forms. Such flexibility would enable us to capture the relationship between consumer behaviors and marketing variables automatically.

In this paper, we will demonstrate the effectiveness of the proposed Hierarchical Gaussian Process approach on store-level scanner data for the refrigerated orange juice category from a Chicago Super Market

2 Data Description

The refrigerated orange juice dataset comes from a major regional supermarket chain in the Chicago area. It contains weekly level scanner data for 11 products in 83 stores. For each product, it includes the price and sale information recorded by the log of the number of units sold. Apart from this, for each store, we also have 11 demographic information that comes from a marketing research company. Seven of the demographic variables relates to consumers: the percentage of the population over age 60, percentage of the population that has a college degree, the percentage of black and Hispanic persons, and the percentage of households with five or more members. The other demographic variables are the log of median income, the percentage of homes with a value higher than $150,000$, and the percentage of women who work. The other four variables measure the competitive environment of the store's trading area: distance to the nearest warehouse store, ratio of sales of this store to the nearest warehouse store, and average distance in miles to the nearest 5 supermarkets, the ratio of sales of this store to the average of the nearest five stores

2.1 Data Preprocess

The raw data contains $106,139$ records in total. Each contains 83 stores, and 11 products, to study the relationship between the log of movement and price at the store-level for a single product, we use the sale and price data of first product Tropicana Premium 64 oz, and it has 9649 records in total. For each store, we will use 70% of the data as the training data, and use the rest 30% as test data set to evaluate the performance.

2.2 Exploratory Analysis

Table 2 gives the summary statistics of the data set

Table 1: Descriptive Statistics

Variable	Mean	SD	Min	Max
price per ounce	0.04	0.01	0.02	0.06
sale (log based)	9.11	0.85	5.26	12.57
percentage of the population that is aged 60 or older	0.17	0.06	0.06	0.31
percentage of the population that has a college degree	0.22	0.11	0.04	0.52
percent of the population that is black or Hispanic	0.15	0.18	0.02	0.99
median income (log based)	10.618	0.28	9.86	11.24
percentage of households with 5 or more persons	0.12	0.03	0.01	0.21
percentage of women with full-time jobs	0.36	0.05	0.24	0.47
percentage of households worth more than $150,000	0.34	0.24	0.00	0.96
distance to the nearest warehouse store	5.10	3.48	0.13	17.85
ratio of sales of this store to the nearest warehouse store	1.21	0.53	0.40	2.57
average distance in miles to the nearest 5 supermarkets	1.21	0.73	0.88	4.10
ratio of sales of this store to the average of the nearest five stores	2.12	0.22	0.09	1.14

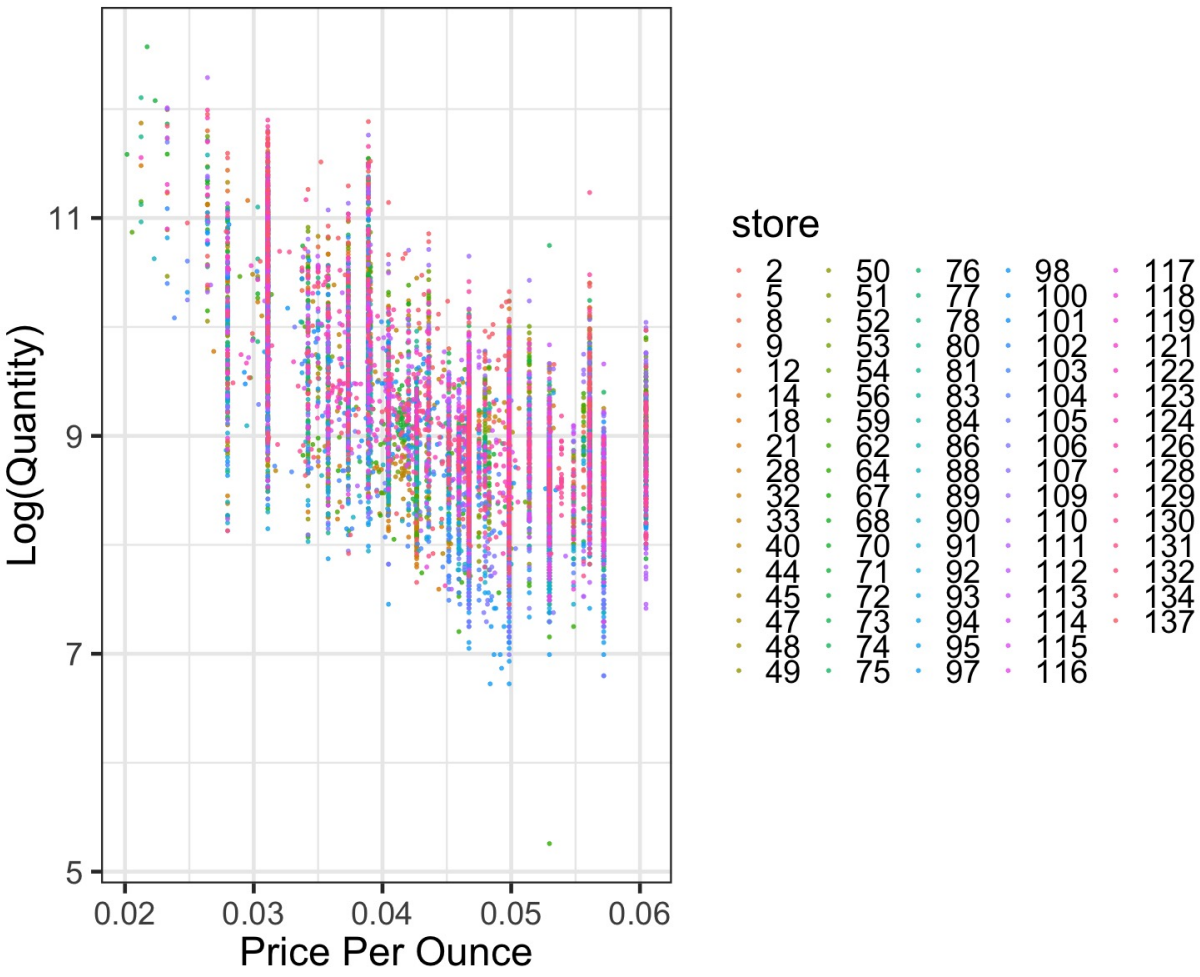


Figure 1: Scatter plot of log movement and price

To begin with, we use a scatter plot to detect the presence of the relationship between demand and price of these 83 stores. Figure 1 shows that there exists a linear trend between the log of movement and price at an aggregate level. Here, stores are encoded by different color. If we take a closer look at the store level, we can observe the difference of price sensitivities across stores. To better illustrate the heterogeneity, we select two typical ones, store 33 and store 111. Figure 2 shows a steeper slope for store 111. In other words, consumers purchasing at store 111 are more sensitive to the price change.

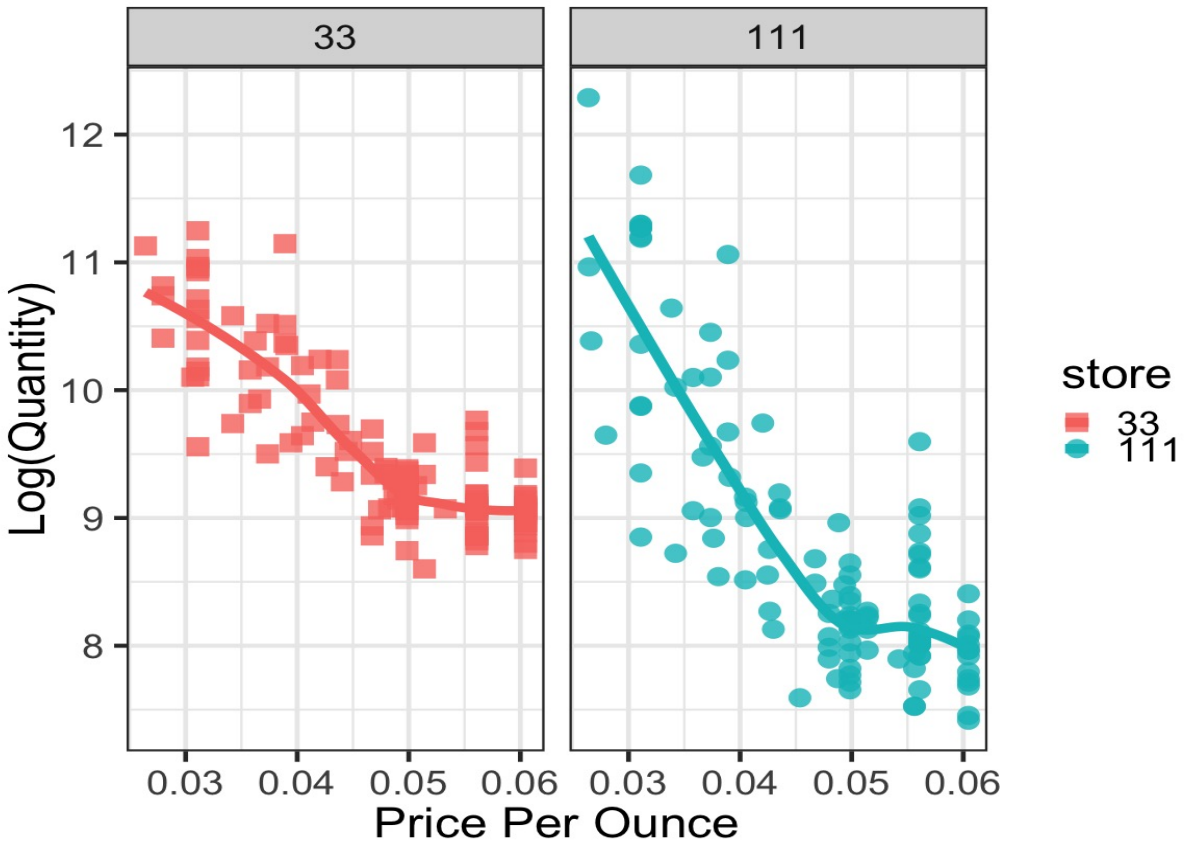


Figure 2: Locally Weighted Scatterplot Smoothing

In Figure 2, we use locally estimated scatter plot smoothing to fit all the data points for store 33 and store 111: when the price increases to a certain level, around 5 cents per ounce, consumers would become less sensitive to the price change. There exists a nonlinear relationship between price and demand. The consumer heterogeneity is, in nature, related to the demographic features. For instance, for highly educated people, due to the higher opportunity cost, they are less likely to seek for lower prices. Therefore their demands of the orange juice are less responsive to the price change. For stores located in an isolated area with few competitors nearby, due to the limited choices, the price decrements can effectively stimulate the consumption. Our dataset also reflects this: store 33 is less sensitive to the price change (see Figure2, and it locates in a more affluent and competitive area (see Table 2).

Table 2: Demographic Information for Store 33 and Store 111

STORE	33	111
AGE60	0.13	0.21
EDUC	0.42	0.10
ETHNIC	0.13	1.00
INCOME	10.35	10.14
HHLARGE	0.01	0.16
WORKWOM	0.47	0.29
HVAL150	0.86	0.01
SSTRDIST	3.13	12.19
SSTRVOL	2.42	1.89
CPDIST5	0.87	1.47
CPWVOL5	0.75	0.29

3 Modeling Framework

3.1 Hierarchical Model

If our model is at an aggregate level, it means that we would fit these points by lines with the same slope, and would fail to capture the heterogeneity across stores. To sum up, estimating the relationship with an aggregate model will lead to an underfitting problem.

A straightforward way to fix this is to model each store separately, but it has the risk to overfit: first, the relevant data for each store are limited. In our dataset, although we have about $10,000$ records in total, there are only about 100 records per store. Second, the relationship between demand and price is nonlinear. For each store, applying a complex model to capture the nonlinearity on such a small dataset is likely to overfit.

To address the overfitting problem, we consider the hierarchical structure, which utilize the demographic features to explain the variation of the price sensitivities. By modeling the price sensitivities in terms of demographic features, we force the demand-price response curves to be alike for stores with the similar features.

To sum up, our hierarchical framework combines within-store and across-store models. For store i with sale $\mathbf{y}_i$, price $\mathbf{x}_i$ and demographic features $\mathbf{z}_i$, the hierarchical model can be written in the form of

$$\text{First Stage}: \quad \mathbf{y}_i = f(\mathbf{x}_i; \boldsymbol{\beta_i}) + \epsilon_y \tag{1}$$

$$\text{Second Stage}: \boldsymbol{\beta_i} = g(\mathbf{z}_i; \boldsymbol{\theta}) \tag{2}$$

The core of our framework is to *a*) model the purchasing behaviors at the store-level: store i has the local parameters $\boldsymbol{\beta}_i$ to capture the heterogeneity; *b*) assume that demographic factors affect the price sensitivities globally through parameter $\boldsymbol{\theta}$. This hierarchical structure allows to share information across stores and further prevent overfitting. As for $f(\cdot)$ and $g(\cdot)$ in these two stages, due to the observed nonlinear relationships in first stage, the demand and price, and in second stage, the complicated interactions between demographics and

price sensitivites, we use the nonparametric approach - Gaussian process to model these relationships, allowing us to automatically accommodate different purchasing patterns that may underlie a given customer base.

3.2 Gaussian Process

We begin by describing the Gaussian process:a Gaussian process is a prior over functions f. It can be defined as a collection of random variables, any finite number of which have a joint Gaussian distribution (Rasmussen, 2004). It is specified by the mean function $m(\mathbf{x})$and covariance function $k(\mathbf{x}, \mathbf{x}')$. Usually, we will take the mean function to be zero, and it can be written as

$$f(\mathbf{x}) \sim \mathrm{GP}(0, k(\mathbf{x}, \mathbf{x}')) \tag{3}$$

In this paper, our input feature $\mathbf{x}$ is multi-dimensional, thus the ARD SE covariance function is employed. Assume $\mathbf{x} \in \mathbb{R}^p$, the ARD SE kernel can be written as

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp(-0.5 \sum_{i=1}^{p} \frac{(\mathbf{x}_i - \mathbf{x}'_i)^2}{l_i^2}) \tag{4}$$

For the i^{th} input dimension, we will learn an individual length scale parameter l_i. The length scale parameter here determines the relevancy of the i^{th} input feature to the regression. If l_i is very large, then the feature is irrelevant(Snelson, n.d.)

Suppose we observe $\{(\mathbf{x}_i, y_i)|i = 1, \cdots, n\}$, if we want to make prediction on the test set $\mathbf{x}_*$, when observations are noise free, we have

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{y}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K & K'_* \\ K_* & K_{**} \end{pmatrix}\right), \tag{5}$$

where $K = K(\mathbf{x}, \mathbf{x}')$ is the covariance matrix for the observed points, $K_* = K(\mathbf{x}_*, \mathbf{x}')$ is the covariance matrix between the test point and the observed points, and $K_{**} = K(\mathbf{x}_*, \mathbf{x}'_*)$ is the covariance matrix for the test points. Then,

$$p(y_*|y) \sim \mathcal{N}(K_* K^{-1} y, K_{**} - K_* K^{-1} K'_*). \tag{6}$$

Random Fourier Features Gaussian processes and kernel methods require computational cost as $O(N^3)$ in the number of observations N, due to the need to compute, store and invert the $N \times N$ kernel matrix in Equation 6, which becomes intractable the data set is at a large scale(Sejdinovic, 2017). To reduce the computational cost, we approximate the shifted invariant kernel $k(\mathbf{x}, \mathbf{y})$ by the inner product of two low-dimensional vectors(Rahimi & Recht, 2008). The underlying principle of the approach is a consequence of Bochner's theorem, which states that a continuous shift-invariant kernel can be represented as the Fourier transform of a positive finite measure. It can be written as

$$k(\mathbf{x}, \mathbf{y}) = \int_{R^d} p(\omega) \exp(j\omega'(\mathbf{x} - \mathbf{y}))d\omega \tag{7}$$

and we can obtain a real-valued mapping that satisfies the condition $E[v_\omega(\mathbf{x})v_\omega(\mathbf{y})] = k(x, y)$ by setting $v_\omega(\mathbf{x}) = \sqrt{2}\cos(\omega'\mathbf{x}+b)$ and ω is drawn from the distribution $p(\omega)$ proportional to its Fourier transform, and b is drawn from Uniform$(0, 2\pi)$. For ARD kernel, $\omega \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a diagonal matrix in which its diagonal elements are $l_1, \cdots, l_p$

Using a Monte Carlo estimation to approximate the kernel function, we have

$$k(\mathbf{x}, \mathbf{y}) \approx \frac{1}{D}\sum_{j=1}^{D} v_{\omega_j}(\mathbf{x})' v_{\omega_j}(\mathbf{y}) \tag{8}$$

With Random Fourier Features, we can approximate the proposed hierarchical Gaussian process, and it can be written in a parametric form as

$$y_i = \mathbf{v}_{\omega_x}(\mathbf{x}_i)'\boldsymbol{\beta}_i + \epsilon_y, \quad \epsilon_y \sim \mathcal{N}(\mathbf{0}, \sigma_y^2) \tag{9}$$

$$\boldsymbol{\beta}_i = \mathbf{v}_{\omega_z}(\mathbf{z}_i)'\boldsymbol{\theta} \tag{10}$$

4 Results and Analysis

In marketing, hierarchical linear regression is widely used to capture the heterogeneity but with the limitation induced by strong assumption of the linearity; Multilayer Perceptron(MLP) is a powerful machine learning technique to capture the nonlinear relationship on an aggregate level. In this section, we compare these two models with our proposed hierarchical Gaussian Process.

4.1 Benchmark Model

Hierarchical Linear Regression

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_i + \epsilon_y, \quad \epsilon_y \sim \mathcal{N}(\mathbf{0}, \sigma_y^2)$$
$$\boldsymbol{\beta}_i = \mathbf{z}_i'\theta + \epsilon_z, \quad \epsilon_y \sim \mathcal{N}(\mathbf{0}, \sigma_z^2)$$

4.2 Alternative Model

Multilayer Perception A multilayer perceptron (MLP) is a class of feedforward artificial neural network. It can be understand as a series of functional transformation

$$\phi(\mathbf{W}\mathbf{x} + \mathbf{b})$$

where $\mathbf{x}$ is the input, $\mathbf{W}$ is the weight matrix, and ϕ is the nonlinear activation function. The nonlinear activation functions give us much more power to approximate arbitrary functions, and can be used to model the relationship between demand and price.

An MLP consists of, at least, three layers of nodes: an input layer, a hidden layer, and an output layer. The first layer is the input layer, and it take the values of the input features. The last layer is the output layer, and it has a single output for prediction. All the layers in between are known as hidden layers (Bishop et al., 1995).

In our application, we consider a MLP with 2 hidden layer and use tanh as the activation function. Since it is an aggregate level model, for each store, we concatenate its demographics to its weekly prices and movements records. That is, the input $\tilde{\mathbf{x}} = [\mathbf{x} \quad \mathbf{z}]$. Our MLP can be written as

$$\begin{aligned}
\mathbf{h}^{(1)} &= \phi^{(1)}(\mathbf{W}^{(1)}\tilde{\mathbf{x}} + \mathbf{b}^{(1)}) \\
\mathbf{h}^{(2)} &= \phi^{(2)}(\mathbf{W}^{(2)}\mathbf{h}^{(1)} + \mathbf{b}^{(2)}) \\
\mathbf{y} &= \phi^{(3)}(\mathbf{W}^{(3)}\mathbf{h}^{(2)} + \mathbf{b}^{(3)})
\end{aligned}$$

4.3 Results

To evaluate the performance of all these models, we compare their Mean Squared Error(MSE) on the test dataset, and the results are summarized in Table 3. The Approximated Hierarchical GP is superior in all cases, and the accuracy improves 27.80% comparing to the linear regression, 3.4% comparing to the MLP model.

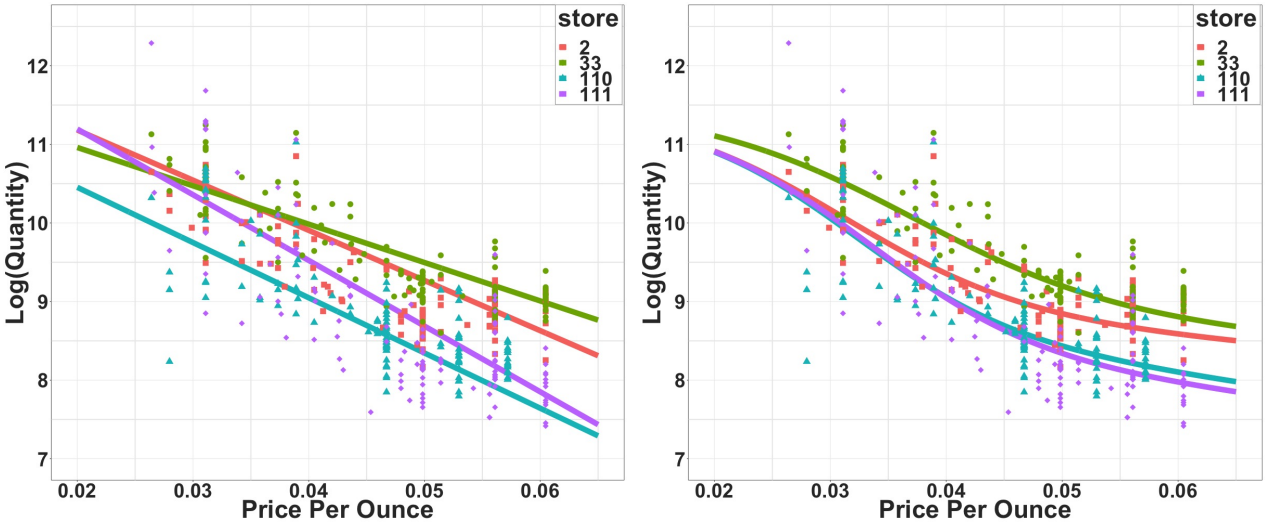
Table 3: Summary of MSE values on the Test dataset

Method	MSE
Hierarchical Linear Regression	0.277
MLP	0.207
Approximated Hierarchical GP	0.194

4.4 Analysis of First Stage

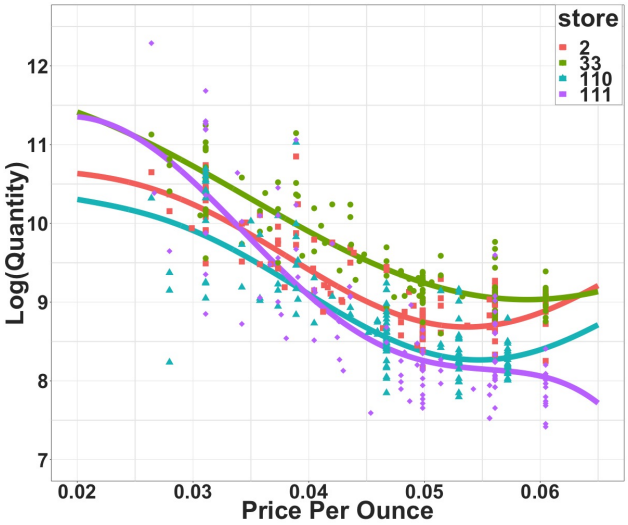
We select 4 stores to visualize the fitted results on the test dataset, and help us to analyze the differences between these three models. Store 2, 33, 110, 111 are selected based on the quantile of the price sensitivity coefficient of hierarchical linear regression.

Nonlinearity MLP allows all these explanatory variables to interact nonlinearly with each other, which leads to a huge gain in terms of the fit, especially when the price is above 5 cents per ounce (see Figure 3b). MLP is modeled at an aggregate level and we append the demographic features to the corresponding price and movement records, that is, there exists only one stage. The aggregate level model assumes that price, movement and demographic features interact in the same way across all the 83 stores.



(a) Hierarchical Linear Regression

(b) Multi-Layer Perceptron



(c) Hierarchical Gaussian Process

Figure 3: Price and Log of Quantity on the Test Dataset. The simulated results are based on store 2, 33, 110, 111 are selected based on the quantile of the price sensitivity coefficient of hierarchical linear regression.

Heterogeneity Due to the restriction of the aggregate model, MLP has a similar trend of the demand curve of stores, resulting in a limited capacity in modeling the heterogeneity. Allowing different parameters for each demand curves, the hierarchical Gaussian Process can better capture the different price sensitivities across stores. The scatter points in Figure 3 are the test data points, and store 111, the purple one, is more sensitive to the price change and differs from the other three stores a lot. From the plot, we can notice that there are improvements of the fits for points on the upper left corner comparing to the aggregate model.

4.5 Analysis of Second Stage

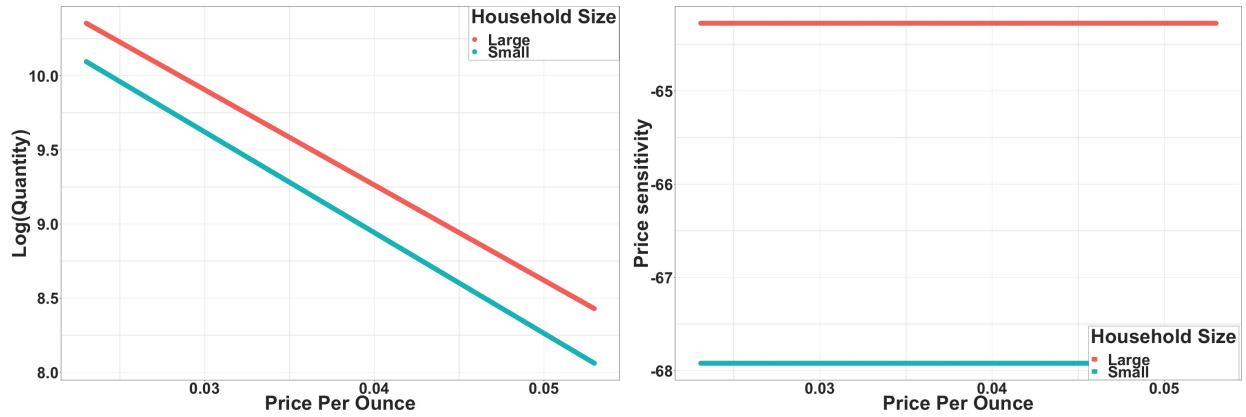
Another essential part is to model how the demographics features affect price sensitivities. For MLP, it only has one stage: the interactions between price and demographics directly affect the movements of the orange juice at an aggregate level; for the other two hierarchical models, the demographics determine the store level price sensitivities first and further impact the movements.

Here, we use HHLarge variable, which measures the percentage of households with 5 or more persons, to demonstrate the differences between these three models. We plot the marginal effect of HHLarge on the left column of Figure 4. Stores with the 1^{st} percentile and the 9^{th} percentile of the HHLarge are selected to visualize the effect. The chosen ones are store 119 and 134, with 8.94% and 15.49% households with 5 or more family members.

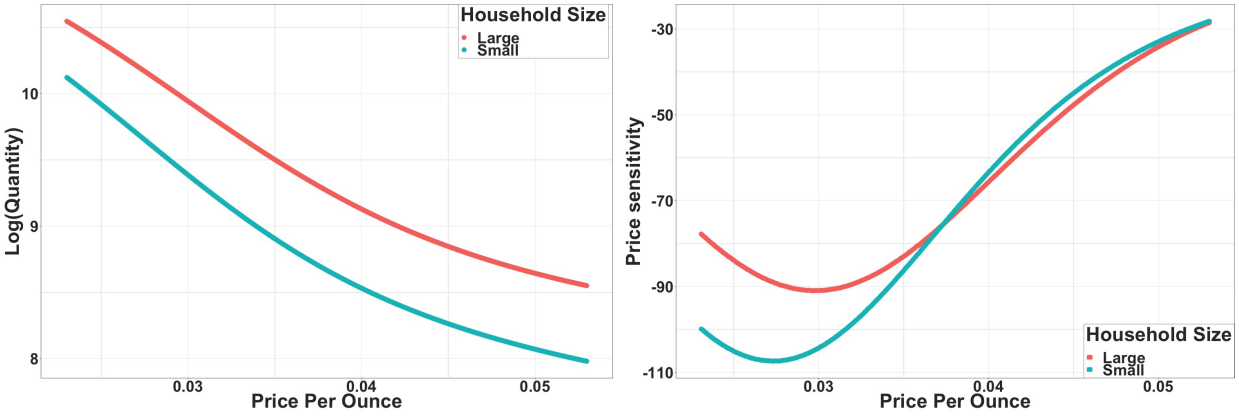
All of these three models show that stores that locate in the area with a higher proportion of large households consume more orange juice, however, when it comes to how price sensitivities, i.e. the derivatives of the curve, change regarding the price, these three models becomes different. To better understand these changes, we plot the derivatives of the demand curve on the right columns.

Nonlinearity Instead of modeling the movement decreasing with the price at a constant change rate, MLP allows nonlinear interactions between demographics and price sensitivities. The right column of Figure 4 demonstrate that derivatives of these two demands curve become more positive as the price increases, in other word, consumers tend to become less sensitive as the price increase and the rate of decrement become slower.

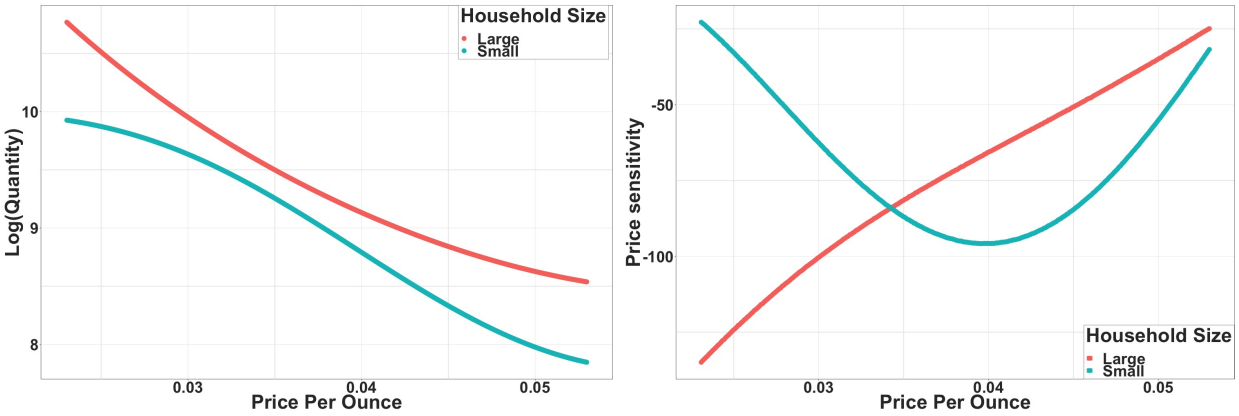
Heterogeneity However, since the MLP is an aggregate level model, the demand curves of these two stores look very similar to each other. When it comes to the hierarchical Gaussian Process, each store has its own parameters, allowing us to model these decreasing patterns flexibly: for stores locating in the area with a higher proportion of large households, the most negative value of derivative is achieved at the lowest price, around 2 cents per ounce, that is, consumers in this area will response to the decrement of the price most when price is at its lowest level. For the store locating in an area with fewer large households, things are different. The most radical change happens at the moderate price level, around 4 cents per ounce. It can be explained by the limited demands of the small household: when price further decreasing, their demands tend to be saturated and they will be less responsive to the purchasing.



(a) Hierarchical Linear Regression



(b) Multi-Layer Perceptron



(c) Hierarchical Gaussian Process

Figure 4: The Effect of Household Size in Price Sensitivities. The left column shows the price sensitivities of two stores locating in the areas that have a large/small percentage of households with 5 or more persons, and the right column shows the derivatives of the curvatures on the left figures. The simulated results are based on store 119 and store 134, which have the 1st percentile and the 9th percentile of the HHLarge, and are represented by the blue line and red line respectively.

5 Conclusion

In this paper, we propose and apply a hierarchical Gaussian Process model to a refrigerated orange juice data. Our proposed framework benefits from the flexible nonparametric Gaussian Process and the hierarchical structures, successfully modeling the nonlinear relationship between features and the heterogeneity across the stores, outperforming the existing models: the hierarchical linear regression and MLP.

6 Limitation and Future Work

In this paper, one limitation is on the second stage model: among the demographics features, many variables are highly correlated, such as income, household value, education level. Thus, further research could focus on a better representation to reflect the consumer heterogeneity. That is, for store i, the demographics variable $\mathbf{z}_i \in \mathbb{R}^p$ where p is the number of demographic features, we assume that there is a lower dimension latent variables $\mathbf{u}_i \in \mathbb{R}^q$ associated with the observed data that can capture the dominant correlation. In Bayesian Principle Analysis (Bishop, 1999), each variables is distributed with a prior by the standard normal distribution,

$$\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$$

The conditional distribution of the observed variable $\mathbf{z}_i$ on latent variable $\mathbf{u}_i$ is

$$\mathbf{z}_i|\mathbf{u}_i \sim \mathcal{N}(\mathbf{W}\mathbf{u} + \mathbf{b}, \sigma^2\mathbf{I}_p) \tag{11}$$

in which the mean of $\mathbf{z}$ is a general linear function of $\mathbf{u}$ and $\mathbf{b}$.

However, the linear mapping from the latent space to the observed space could be too restrictive. Replacing this mapping with Gaussian Processes gives us the Gaussian Process Latent Variable Model (GPLVM), allowing a more flexible relationship between the latent V and observed W(Titsias & Lawrence, 2010). That is,

$$\mathbf{z}_j|\mathbf{u} \sim \mathcal{N}(\mathbf{0}, K + \sigma^2) \tag{12}$$

where $\mathbf{z}_j \in \mathbb{R}^n$ is the j^{th} demographics feature and $\mathbf{u} \in \mathbb{R}^{n\times q}$ is the latent variable. Both contain information about all n stores.

This will force the lower dimension latent $\mathbf{u}$ to capture dominant variance among the demographics features, which can be used to build a more interpretable hierarchical model:

$$y_i = \mathbf{v}_{\omega_x}(\mathbf{x}_i)'\boldsymbol{\beta}_i + \epsilon_y \tag{13}$$
$$\boldsymbol{\beta}_i = \mathbf{v}_{\omega_z}(\mathbf{u}_i)'\boldsymbol{\theta} \tag{14}$$
$$\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q) \tag{15}$$
$$\mathbf{z}_j|\mathbf{u} \sim \mathcal{N}(\mathbf{0}, K + \sigma^2\mathbf{I}_n) \tag{16}$$

where $\epsilon_y \sim \mathcal{N}(\mathbf{0}, \sigma_y^2)$ and the log likelihood can be written as

$$p(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) = \prod_{i=1}^{n} \int p(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{u}, \boldsymbol{\theta}) p(\mathbf{u}) p(\mathbf{z}|\mathbf{u}) d\mathbf{u} \tag{17}$$

The integral is intractable as $\mathbf{u}$ appears nonlinearly inside the inverse of the kernel. Instead, we apply the variational inference framework to approximate it by taking a variational distribution

$$q(\mathbf{u}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{u}_i | \boldsymbol{\mu}_n, \mathbf{S}_n) \tag{18}$$

where the μ_n and S_n are variational parameters.

Then the loglikelihood is lower bounded by:

$$\begin{aligned} F(q) &= \sum_{i=1}^{n} \int q(\mathbf{u}) \log \frac{p(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{u}, \boldsymbol{\theta}) p(\mathbf{u}) p(\mathbf{z}|\mathbf{u})}{q(\mathbf{u})} d\mathbf{u} \\ &= \int q(\mathbf{u}) \log p(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{u}, \boldsymbol{\theta}) p(\mathbf{z}|\mathbf{u}) d\mathbf{u} - \mathbf{KL}(q||p) \end{aligned}$$

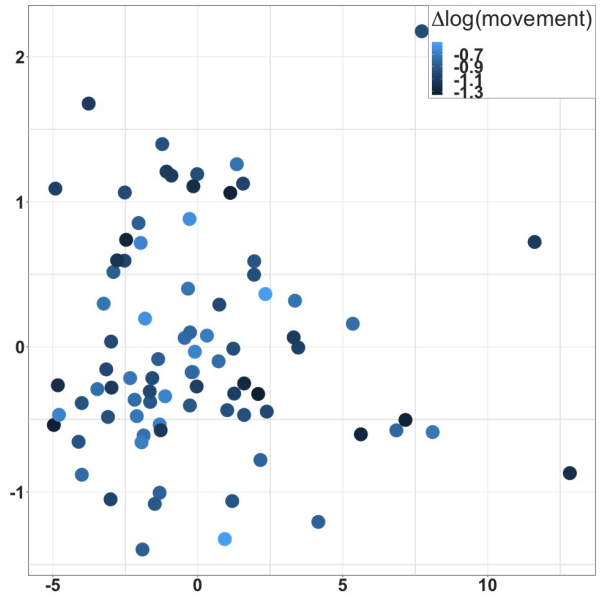
6.1 Preliminary Results and Analysis

In our dataset, we have 11 demographics features, and these features are mapped to a five dimension latent space by GPLVM. We initialize the latent variable based on PCA and achieve the MSE of 0.194 on the test dataset, which is the same as the hierarchical Gaussian Process. Yet the lower dimensional representation of the demographics offers a more interpretable model.

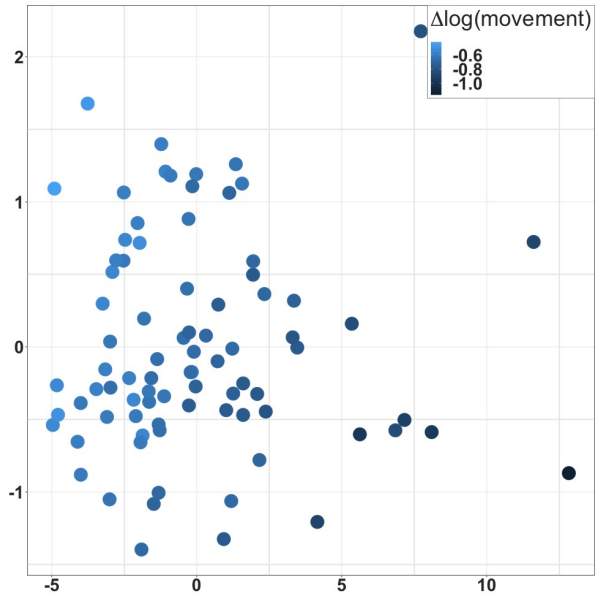
Here, we use the low dimensional representation to demonstrate the result of our model. For our store-level model, we consider a 2D visualization of the 83 stores by their two-dimensional representation of the demographics features. First, we need to decide which dimensions to drop: we select two latent directions based on the kernel length scale since it implies the "relevance" of the corresponding dimension. Dimensions with small length scales imply large variance along the dimensions of the function being modeled. Thus, we keep dimensions with the two smallest length scale. Such idea is similar to the dimension selection in Principle Component Analysis (PCA), where we choose eigenvectors with the highest eigenvalues since the large eigenvalues correspond to the large variance along the eigenvectors' direction, and contain the most information about the distribution of the data.

Figure 5 visualizes all 83 stores. Each store is represented by a point whose location is determined by the dimensions with the smallest two length scale. In each subfigure, stores are colored by different features to help us understand the mechanisim of the learned model : first row is about the price sensitives: Figure 5a and 5b colored by the change of movement when the price increases from 3 cents per ounce to 4 cents per ounce and from 4 cents per ounce to 5 cents per ounce respectively (since most prices are within the range of 3 cents per ounce and 5 cents per ounce). The darker the point is, the more sensitive the store is to the price increment; for the second row, we choose demographics feature "log of the median of consumer's income" to color the points. From the plots, we observe that when the price is at a relatively low level, as it increases from 3 to 4 cents per ounce, the movement change is highly correlated with the income: consumers with relatively low income are more sensitive to the price change. As the price increase to a high level,from 4 cents per ounce to 5 cents per ounce, their price sensitivities are highly correlated with the latent dimension instead of the explicitly demographics features: stores locate in the right part of the latent space are more sensitive to the price change.

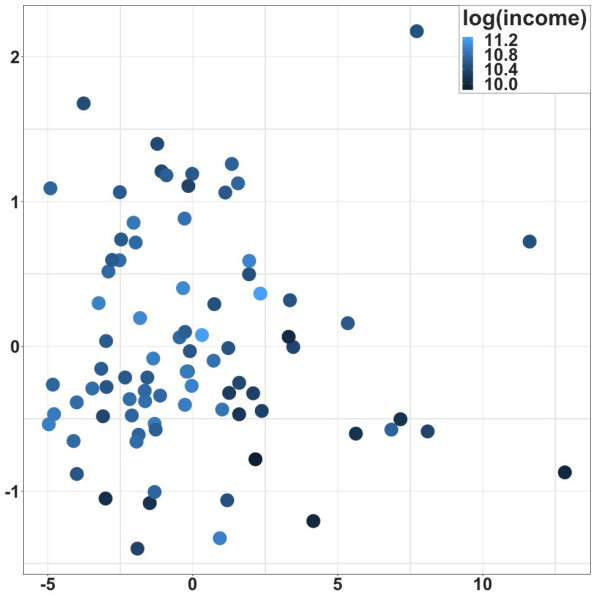
Figure 5: 2D Visualization of 83 Stores. The coordinate of the points are attained by the latent dimensions which have the largest inverse lengthscale value



(a) Scatter plot colored by change of Log(movement) when price increases from 3 cents per ounce to 4 cents per ounce



(b) Scatter plot colored by change of Log(movement) when price increases from 4 cents per ounce to 5 cents per ounce



(c) Scatter plot colored by log(income)

References

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal*

of econometrics, *89*(1-2), 57–78.

Bishop, C. M. (1999). Bayesian pca. In *Advances in neural information processing systems* (pp. 382–388).

Bishop, C. M., et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Montgomery, A. L. (1997). Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing science*, *16*(4), 315–337.

Rahimi, A., & Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in neural information processing systems* (pp. 1177–1184).

Rasmussen, C. E. (2004). Gaussian processes in machine learning. In *Advanced lectures on machine learning* (pp. 63–71). Springer.

Sejdinovic, D. (2017). *Oxford SC4/SM4 Data Mining and Machine Learning, Lecture Notes: Gaussian Processes.* (URL: `http://www.stats.ox.ac.uk/ sejdinov/teaching/dmml/17`$_{n}$*otes*8*.pdf*.)

Snelson, E. (n.d.). Tutorial: Gaussian process models for machine learning.

Titsias, M., & Lawrence, N. D. (2010). Bayesian gaussian process latent variable model. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 844–851).