# Estimating price elasticity at scale using Bayesian methods

Jay Li, Filip Ryzner

## I. Introduction

The importance of the effect of price change on product or service demand levels, the so called price elasticity of demand (price elasticity), has been considered to be of a mixed importance for real-world business decision making by academicians during the first half of the 20th century. While [Markham, 1951] argued that the ability of modern companies to understand sensitivity of their product demand to changes in their prices can yield significant benefits, [Clodius and Mueller, 1961] did not even consider it to be a factor worth investigating. Next, [Johnson and Helmberger, 1967] argued based on elementary data analysis that the price elasticity of demand is of significant importance for the market structure from the economic point of view. Moreover, [Johnson and Helmberger, 1967] argued that the price elasticity of demand should not be expected to be the same across various industries, thus directly pointing to its heterogeneity at least at the industry level.

The importance of price elasticity has been proven empirically by the quantity of academic literature investigating it across various sectors of the economy. For example, [Andreyeva et al., 2010] investigated the effect of food prices on their consumption, [Jawad et al., 2018] has investigated the price elasticity of non-cigarette tobacco products, [Pendzialek et al., 2016], estimated the price elasticity of demand for health insurance, and many other examples of price elasticity estimation can be easily found. Finally, [Chindarkar and Goyal, 2019] has demonstrated the price elasticity of electricity to be significantly heterogeneous with respect to geographical regions of India. Therefore, [Chindarkar and Goyal, 2019] findings suggest that price elasticity is not only elastic across industries, but various factors can also cause within-industry heterogeneity of the price elasticity of demand.

Therefore, possessing a strong understanding of product or service price elasticity is an important factor for business decision making process with regards to both internal factors, such as inflation and competition, as well as external factors, such as revenue targets and marketing campaign planning. The ability to reliably predict demand response to price changes can make the overall business planning much easier and may potentially increase the business stability.

Finally, while reviewing the academic literature focused on the price elasticity of demand, one can notice that while it relies on the traditional econometric tools. These methods do not attempt in any rigorous way to account for the uncertainty in the estimates. This uncertainty can be based on inherent data noise, and can also spring from the oftentimes major simplifications induced by economic models (as outlined in [Ng, 2016]).

In this project we have decided to probabilistically model price elasticity of products from 10 different stores, which we sourced from the Walmart dataset provided for the M5 forecasting competition [Wal, ]. Our approach was inspired by [Weber and Steiner, 2021] and compares the effectiveness of approaches based on Hierarchical Linear Regression and Empirical Bayes Methods. Our implementation relies on the STAN [Carpenter et al., 2022] probabilistic programming language and also the Facebook Prophet modelling package introduced in [Taylor and Letham, 2018]. We have found that our own implementation of Empirical Bayes Method which increases the informativeness of the priors of Prophet modelling package significantly over-performs a simple Hierarchical Regression model implemented using the Stan modelling package in terms of both accuracy and speed of training.

The paper is structured in the following way. In Chapter 2 we provide a review of the relevant literature for our work, In Chapter 3 we discuss the dataset used for our analysis, then in Chapeter 4 we introduce and discuss all modeling approaches used in this paper, then in . Finally, in the Chapter 5 we provide summary of our exploratory analysis, whose importance we elaborate on in the Chapter 6. Lastly, we conclude our paper in the Chapter 7.

## II. Literature Review

The core paper which motivated this project is [Weber and Steiner, 2021]. The paper considers sales of Orange Juice with a hierarchical structure to model the heterogeneous effects between stores. Hierarchical models are mainstays of Bayesian inference, with group level biases allowing for better estimates of individual parameters. In a frequentist setting these methods are sometimes known as random coefficient models. [Greg M. Allenby, 1999]. The field of marketing is an area where the latent coefficients towards prices have been well studied. When individual response data is sparse, group level relationships can better estimate overall effects.

$$y_{t,i} = \beta_{0,i} + \beta_{1,i} \cdot price_t + \epsilon \tag{1}$$

Here, $y_{t,i}$ would denote individual i's demand at time $t$ for the product. This is often a binary 1 or 0 if the customer purchases that product. With few data points at an individual's level i estimating such parameters with accuracy has been difficult. Bayesian hierarchical models have been used to estimate the behaviour of specific respondents and also the distribution of all $\beta_t$ of all the respondents.

In [Weber and Steiner, 2021], authors were interested in the price elasticity of products, defined as the change in quantities sold per change in price in log space. This is also the coefficient attached to a simple linear model between $Q$

| | Hobbies | Household | Food |
|---|---|---|---|
| **Total Products** | **565** | **1047** | **1437** |
| Sub-Category 1 | 416 | 523 | 216 |
| Sub-Category 2 | 149 | 515 | 398 |
| Sub-Category 3 | N/A | N/A | 823 |

**TABLE I:** Walmart dataset composition overview

quantities sold and $P$ price.

$$\ln Q = \beta \cdot \ln(P) \tag{2}$$

Since the elasticity coefficient $\beta$ is considered an important latent variable, it is worthwhile to formulate a Bayesian treatment and examine the most credible $\beta$ variables as a distribution. The formulation of three different models led to $\beta$ being treated in three separate ways:

*i)* To be uniform across all stores (homogeneous model)
*ii)* To be different and independent across all stores (independent model)
*iii)* To be different but related in a hierarchical manner (hierarchical model)

The paper finds that the hierarchical model, specified with Bayesian priors performed best in future predictions on a withheld test set. However data source in the paper was limited to 67 weeks and only orange juice. We hope to extend this analysis to other products and over a longer time scale. For this, we considered the Walmart sales dataset used in the recent M5 Forecasting competition. This dataset contains over 30,000 times series of sales and prices across 10 different locations in 3 different U.S. states over a recent 5 year period.

Once we extended the data to a 5 year period, it was clear that strong seasonality behaviours existed in certain products. We needed a more robust implementation of seasonality features and used Facebook Prophet for their Fourier series model of seasonality and Bayesian estimate of trend and trend change-points [Taylor and Letham, 2018]. Prophet also has useful analyst-in-the-loop features and built-in functions to deal with holidays and outliers.

The challenge with Prophet is the inability to fit multiple time series and $\beta$s in a hierarchical structure (fitting a homogeneous and independent version is straightforward). We would like each product to have it's own seasonality and trend behaviours, but for the $\beta$ coefficients to influenced by other similar products or other products from the same state. For this, we consider an Empirical Bayes variation where anticipate some regression to the mean for the latent variable $\beta$, and try to improve the estimates for each product by taking information from all products [Casella, 1985].

## III. Dataset and Exploratory analysis

The M5 competition Walmart dataset is obtained from Kaggle public repository [Wal, ]. The dataset contains store-level sales data for 3,049 unique Walmart products, which are distributed across the following three categories food (47%), hobbies (19%) and household (34%) products, whose composition we detail in the Table I.

Moreover, the sales data were collected across three U.S. states, specifically California, Texas and Wisconsin, with each of these states collecting data across 4, 3 and 3 local stores respectively. We note that every Walmart product in the dataset is sold in every store in the dataset. Therefore, data are easy to aggregate at various levels such as category or state. In our analysis, we decided to focus on predicting item store-level sales via infering price elasticity of demand as was done in the [Weber and Steiner, 2021].

The whole dataset contains observations for a period of 4.5 years on a daily basis. Price data for products were provided at a weekly rate so we aggregate to the weekly level for our models. Grouping the data over the weekly period helps to contain the daily variation as during our data analysis we observed that sales data are much more stable over weekly periods compared to the daily granularity. Therefore, our whole dataset consists of 218 weekly observations per product and per store. For our analysis we split the dataset into a training and testing, where the training data consists of the first 175 weeks and the model is validated on the next 43 weeks.

Before starting with the modelling, we have carried out an exploratory data analysis, which has uncovered the following underlying trends in the data:

- The same product is often priced differently across the various stores and changes price at different times, which is probably due to different promotions and different factors determining the sales strategy across the states. Lastly, product prices for stores in the same state are either same or extremely similar.
- Most products do not undergo many price changes. This is likely due to Walmart's low everyday price model compared to other retailers that may choose to run many promotions. This is demonstrated in Figure 1, which depicts the histogram of total price changes over the period across all products.
- We have noticed unexpected periods of 0 sales for product with high average turnover. We assume that this is because occasionally items can run out of stock. This may bias the elasticity estimates as there are sudden drops in sales despite the price remaining constant, we have have to deal with these stock-outs as outliers.


**Fig. 1:** Histogram of number of price changes - demonstrating that most of the products had up to 4 price changes over the period in question

Lastly we note that the dataset is fully anonymized, thus it does not contain any product specific details except for the category product belongs into and and a unique numerical ID. Therefore, we were unable to determine competitors or related product relationships in the dataset. We considered making

the assumption that products with a similar price and quantity sold are direct competitors. However, this would have been a very strong simplification not guaranteeing that there are no significant underlying differences in the products chosen as competitors, which may mislead the modelling process.

## IV. Methodology

Inspired by the [Weber and Steiner, 2021], we have used the STAN probabilistic programming language [Carpenter et al., 2022] (STAN) to build several benchmark hierarchical linear models, which we compare to the Empirical Bayes implementation using the Prophet modeling package [Taylor and Letham, 2018] (Prophet). In the following section we detail all the approaches used in our analysis.

We note that while [Weber and Steiner, 2021] always works with natural logarithms of price and demand, $\ln P, \ln Q$, however the store level quantities we are working with are at a much lower magnitude and we have generally obtained better results by working with the non-logged values of price and demand. This is reflected in our model equations.

### A. Bayesian hierarchical linear regression models

*1)* ***Independent Model****:* The Independent Linear regression model is the simplest estimating the price elasticity of demand $\beta$ separately for every store independently. It can be formulated as follows:

$$Q_t = \alpha + \beta \cdot P_t + \sum_{q=2}^{4} \delta_q \cdot E_{q,t} + \epsilon_t \tag{3}$$

Where $Q_t$ is the sales quantity of a product sold at time $t \in 1 \dots T$, $P_t$ represents the product price at time $t$ with the corresponding $\beta$ price sensitivity of demand coefficient. Finally, $E_{q,t}$ is the quarterly seasonality dummy for quarter $q$ at time $t$ and corresponding demand sensitivity coefficient $\delta_q$.

While this model is trivial and fast to fit, it cannot benefit from any information sharing across stores as all stores are sampled and is prone to over-fitting due to sometimes very occasional price changes across individual stores.

*2)* ***Homogeneous model****:* The Homogeneous model fits to data for a product across all stores and assumes that the price elasticity of demand is not heterogeneous for products. It infers one price elasticity coefficient $\beta$ for each product, ignoring any potential clustering or differences based on geographical region or stores themselves. The model equation can be described as follows:

$$Q_{m,i,t} = \alpha_m + \beta_m \cdot P_{m,i,t} + \sum_{q=2}^{4} \delta_{m,q} \cdot E_{q,t} + \epsilon_{m,i,t} \tag{4}$$

The meaning of coefficients and variables is the same as in the Simple hierarchical model above with the addition of the store index $m \in 1 \dots J$.

Compared to the model introduced in [Brezger and Steiner, 2008], [Steiner et al., 2007] and subsequently used in [Weber and Steiner, 2021], we had to drop the cross-product price effects as well as the dummy variable signalling whether product was on shelved or not. However, these simplifications should not significantly weaken our model as [Hanssens et al., 2003] have demonstrated that the cross-item price effects are oftentimes significantly weaker compared to the effect of changes in price of the product itself. Regarding the dummy shelve variable, we assume that all products were shelved at all times as we cannot infer from data whether period of 0 sales indicated no sales or no stock of the product.

The homogeneous model remains simple, fast to sample, and unlike the above defined Independent regression model can benefit from information sharing across stores, however it makes the critical assumption that consumers across regions and stores react in the same way to a unit price change in the product price. This seems like a very strong assumption, especially given our preliminary data analysis, which highlighted the fact that the same product is very often priced differently across different stores. Thus, making the same unit change in price across regions with different product prices necessarily results in different % change in the product price, which has been shown to affect customers behaviour more than level/absolute changes in the price [Cornell, 1981].

*3)* ***Hierarchical model****:* To address the above mentioned simplification ignoring potential heterogeneity in the price sensitivity of demand, we introduce the Heterogeneous hierarchical model, which requires that data collected at each store are tied to one and only one segment from $K$ predefined segments, where the highest possible $K$ corresponds to the number of stores in the dataset. The model is defined by the following equation:

$$Q_{m,i,k,t} = \alpha_{m,k} + \beta_{m,k} \cdot P_{m,i,t} + \sum_{q=2}^{4} \delta_{m,q} \cdot E_{q,t} + \epsilon_{m,i,t} \tag{5}$$

where the meaning of all variables and coefficients gain remains the same, however we add the variable $k \in 1 \dots K$ indicating store segment membership. We assume that the segment membership is given by the user and not inferred by the model. Therefore, we now sample $K$ different segment price sensitivities of demand $\beta_k$ alongside with $K$ different intercepts $\alpha_k$.

We observe that for $K = 1$ the Heterogeneous model collapses into the Homogeneous model. Moreover, for $K$ equal to the number of stores in the dataset, the model approaches the Simple hierarchical linear regression model. However, a key difference comes from the fact that the model now samples the quarterly seasonality coefficients as well as the error term jointly for all stores, thus sharing information.

*4)* ***Priors****:* For setting the model priors we followed an approach introduced in [Otter et al., 2004] and used in [Weber and Steiner, 2021], which uses a fixed effect model estimated by OLS to obtain prior means for $\alpha$, shared coefficients $\delta_q$ and price elasticity coefficient $\beta$. Our priors were set to following values:

- Intercepts: $\alpha_m \sim Normal(mu_\alpha, 50)$
- Price sensitivity coefficients: $\beta_{m,k} \sim Normal(\mu_\beta, \tau_k)$, where $tau_k \sim InverseGamma(1, 1)$
- Quarterly effect coefficients: $\delta_{m,q} \sim Normal(\mu_q, 50)$
- Error term: $\delta m, i, t \sim InverseGamma(1, 1)$

Where $mu_\alpha, \mu_\beta$ and $\delta_{m,q}$ are taken from the fixed effect OLS estimates model. We note that we have found that our trace plots have improved when we replaced 0 means with OLS estimated means. This points to the fact that the OLS means can help achieve faster and better convergence, which seems to be a reasonable assumption as for example for some high turnover products tend to have relatively high intercept, this setting the intercept mean to 0 while sampling from normal distribution may slow the convergence. Finally, our priors are set to be still relatively uninformative given the high variance of 50 when sampling from the normal distribution, which is a result of the high variability in the dataset.

All sampling in STAN is done using the No-U-Turn sampler (NUTS) introduced in [Hoffman et al., 2014]. Moreover, we set the number of Markov Chains in STAN to 4 and we take 2000 samples per chain per iteration out of which 1000 is discarted as warm-up.

### B. Prophet time series model

There are a few key advantages when applying the Prophet implementation of time series regression over the linear bayesian model specified above. Prophet uses a decomposable time series model shown in Equation 6.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{6}$$

Where $g(t), s(t)$ and $h(t)$ represent the trend, seasonality and holiday effects respectively. The model is also specified probabilistically using Stan allowing for uncertainty interval estimates. Thus we can directly compare samples from Prophet to those generated via Bayesian linear regression earlier. For the problem of finding elasticity in sales data the main advantages of using Prophet are:

- Dealing with seasonality using Fourier series, this is a major improvement from using a seasonal effect feature as it allows for a cyclic connection from the first month to the last.
- Allowing for potential changes in trend, is a unique feature which allows for changing trends where neccessary over time
- Outliers and holidays are built in, which can explain peak sales at particular weeks, and occasional stock-outs of 0 sales in the training data
- Scaling and standardisation of the data, as well as vectorized Stan code, speeding up inference

To consider the effects of the price $P_{i,t}$ we use the add regressor functionality built into Prophet, which allows for additional regressors in the time series models. We treat the weekly price of the product as an additional feature which is added linearly Equation 6. This allows us to easily fit an independent model to every item every store.

The disadvantage to using Prophet is that we are unable to fit multiple time series in a hierarchical manner. For the same product, we would want our elasticity estimates from one store to influence another. We expect that customers from different stores to behave similarly toward this product as prices change. For a single store with a few price changes, we may occasionally find large positive elasticity values which we should be cautions of.

One way we are able to share information between models is to use an informative prior. Empirical Bayes [Casella, 1985] allow us to find such a prior with data. For a given product we accomplish this in a three step process:

*i)* Fit to each individual store independently, using the non-informative prior which are prophet defaults
*ii)* Given all the distribution of elasticity for each store, fit a normal distribution which describes the overall elasticity (expected to be a negative value)
*iii)* Use the fitted normal distribution as the prior for elasticity, refitting to every store again. Report the final elasticity and predictions

In order to specify the priors on this additional regressor, we had to introduce minor modifications to the Prophet source code itself, which includes:

1) Changing the add_regressors() function to include a prior mean argument. 2) Rewriting the Stan code to accept the prior mean, keeping as 0 for all other features except for the added regressor. These changes allow us to implement a Shared prophet model, which will hopefully improve on the Independent version without any shared information.

## V. Results

In this section we evaluate all our models for accuracy and the Prophet model for inference.

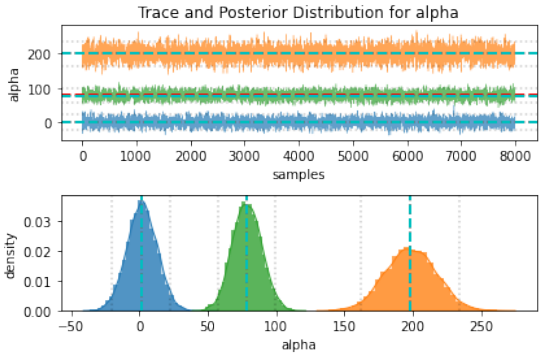
*i)* **Accuracy:** The quality of predictions for future data in the test set (43 weeks after training on 175 weeks)
*ii)* **Inference:** The quality of inference of the elasticity coefficients from the training set. Taken directly from the fitted Prophet, which estimates it as a latent variable

For our tests we work with 15 products, with 5 from each of the major categories - Household, Hobbies, and Food. All product labels were sampled uniformly at random from all the products in the category that registered at least one price change over the training period.

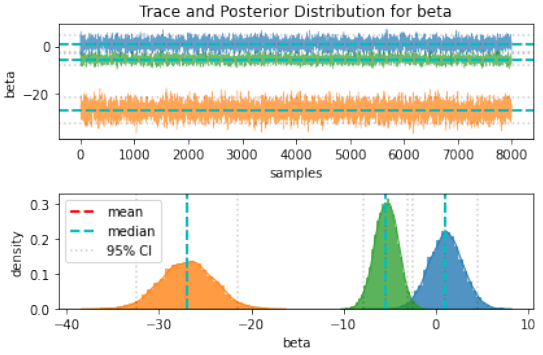
### A. Markov Chain diagnostics

We have used the traditional Markov Chain Monte Carlo diagnostics tools to check the quality of the samples generated via the NUTS algorithm. While we have checked all our models for the 15 examples used in this section. We provide one example of our diagnostics tools for illustration. Figure 2 demonstrates the trace plots for relevant parameters in the Heterogeneous model for ITEM 'FOODS_1_096' where the number of segments/clusters $K = 3$ and stores are segmented based on the region they are located in.

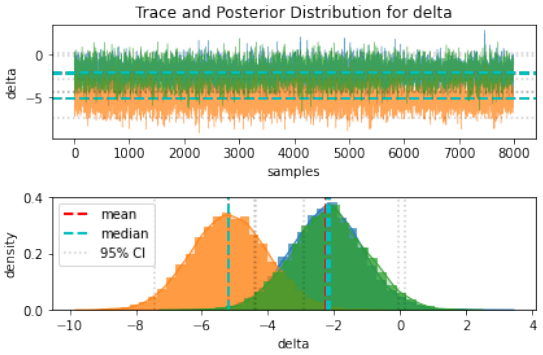
We observe that all trace plots for all variables across segments $k$ appear to be without any visible anomalies as they all resemble a random noise centered around stable mean without any visible trends or correlations. The density plots also appear to be as expected.



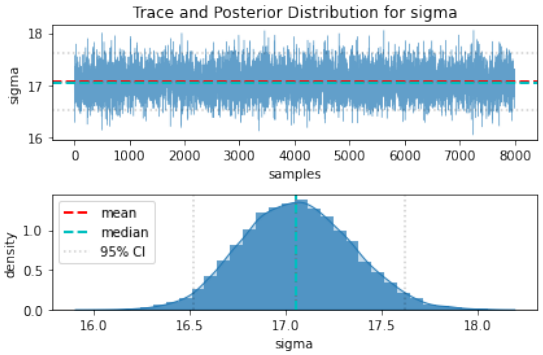

**(a)** Diagnostic plots for $\alpha_k$




**(b)** Diagnostic plots for $\beta_k$




**(c)** Diagnostic plots for $\delta_q$




**(d)** Diagnostic plots for error $\sigma$

**Fig. 2:** MCMC Diagnosis for 'FOODS_1_096'

### *B. Error metric for accuracy*

We chose the symmetric Mean Absolute Percentage Error (sMAPE), which is an adjusted version of the Mean Absolute Percentage Error, to asses the accuracy of our models. The main reason for choosing sMAPE over MAPE was due to the fact that it addresses the MAPEs problem with producing extreme values for points, which have the actual value close to 0 as its value is upper bounded by 200% and lower bounded by 0%. This decision was taken as we were simultaneously running all the test on logged data, thus we needed a metric that is able to produce comparable results for both logged and non-logged data. The sMAPE metric is defined as follows:

$$\text{sMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2} \tag{7}$$

where $F_t, A_t$ represent the forecast and actual value at time $t$ respectively. The result is in $\%$ and the maximum value of the metric is 200%. For Deeper discussion of the metric please refer to [Kreinovich et al., 2014].

Moreover, we have implemented the metric in a way that when both forecast and actual value are 0, the computation does not fail. This was implemented as especially for product with common zero sales it is possible that both forecasted and actual value are 0. To obtain our sMAPE values, we fit each item sales data across all 10 stores across three states. The reported sMAPE is the average error across all 10 stores.

### *C. sMAPE evaluation across models*

Throughout this section we will be referring to the Table II, which contains all our results for 15 randomly selected products. Moreover, to save space, we have labelled the models using shortcuts, which are as follows: Homog Linear - Homogeneous Model; Indep. Linear - Independent Model; Hierarchical Region - Hierarchical model with $K = 3$ based on store regions; Hierarchical Separate - Hierarchical model with $K = 10$ thus each store in its own segment; Independent Prophet - Individually fitting stores using the Prophet; Shared Prophet - Prophet extensions using Empirical Bayes.

When comparing the sMAPE accuracy scores across our models, we observe the following.

First of all, we see that all the models implemented in STAN are clearly out-performed by the models implemented in the Prophet. It appears Prophets additional treatment of trend and elasticity fits better to independent time series.

In terms of the models implemented in STAN the results are in line with our expectations. The Homogeneous model is clearly the worst out of the STAN models. We remind the reader that the upper bound for the sMAPE metric is 200%, thus the Homogeneous model error probably exceeds the upper bound as most of the sMAPE values are 200%. This is expected due to the high variability the prices and sales for one product across the regions. Therefore, having the same inferred coefficient will be a large compromise.

The Independent model seems to be a slight improvement on the Homogeneous model. Again, this confirms our intuition as the Independent model is more flexible to the coefficients of each store. However it is prone to over-fitting as some regions

change price less often than others. Finally, we observe that both Heterogeneous models are clearly superior to the Homogeneous and Independent models. Again, in line with our intuition the Hierarchical model, which segments stores into $K = 3$ segments based on their regions clearly outperforms the Independent model, which segments all stores independently using $K = 10$. This is again due to the fact that too granular segmentation makes the model prone to over-fitting a store over the training period.

For the models implemented in Prophet we observe that while in some instances the Independent prophet model is slightly better than the Shared model, the overall trend is clearly in favor of the Shared model using the Empirical Bayes implementation. This highlights that incorporating the Empirical Bayes approach into the Prophet forecasting adds value.

Overall, the accuracy results seem to confirm that there is indeed a heterogeneity in demand as the Homogeneous model fails to provide reasonable predictions. Moreover, it also appears that grouping stores into segments makes the model more robust, thus less prone to over-fitting in certain scenarios.

### *D. Inferred price elasticity values*

From here, we investigate the elasticity coefficients as inference within the model. Since Bayesian linear regression fails to capture the seasonality and outliers in the data, we are sceptical of the elasticity estimates from this method. Instead, we focus on the inference of elasticities from Prophet, which are also shown in Table II

With Prophet, price effects and elasticity are used to explain the leftover sales after accounting for trends and seasonality, we can see the exact magnitude of price change effects in the Prophet's decomposition plots. This is shown in Figure 4

Even with trend and seasonality removed, the residuals remaining will still fluctuate greatly despite prices remaining the same. This should be accounted for as random noise in the Prophet model, but can sometimes be mistakenly taken as signal when positively correlated to the price feature. In these situations, the elasticity estimate could potentially be positive, suggesting that the sales of a product would increase given a price increase. We plot the estimates of elasticity for one product in Figure 6. While this could be possible with luxury goods or market conditions we do not expect it to be the case with products at Walmart.

After an initial heterogeneous fit we can easily identify the outlying stores with extremely positive or negative elasticity. We use Empirical Bayes to estimate a informative prior distribution based on every store, which is then used as the prior for the individual stores. This is telling Prophet that the most likely values for elasticities are expected to be similar to the elasticities we identified independently at other stores. This should pull the outlying positive elasticity closer to the other stores, just as like having shrinkage in hierarchical modelling, as shown in Figure 6.

We see that the the majority of the stores report an elasticity close to 0, but specific stores in Wisconsin report some samples of large betas greater than 20 which is very unlikely. After applying empirical Bayes and setting an informative prior, the large positive values are no longer credible, and our elasticity estimates become more reasonable at the bottom of Figure 6.

During fitting, we still used trace plots and diagnostics to ensure our model has converged. Nonetheless, we can never be certain that our model and our latent elasticity estimates have converged. The confidence in our model comes from the accuracy of the prediction it generates, show in Table II however this is not a guarantee.

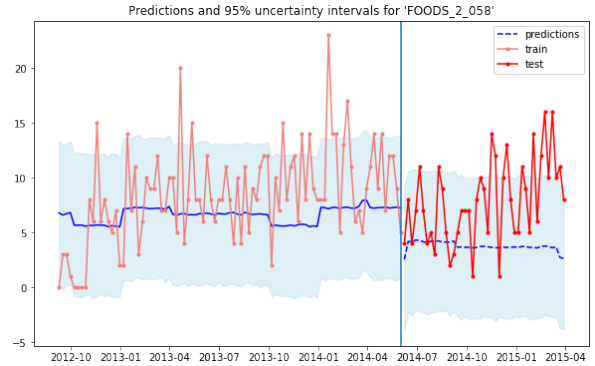



**Fig. 3:** Hierarchical model predicted sales versus actual sales for FOODS_2_058 in CA_1 over the testing period

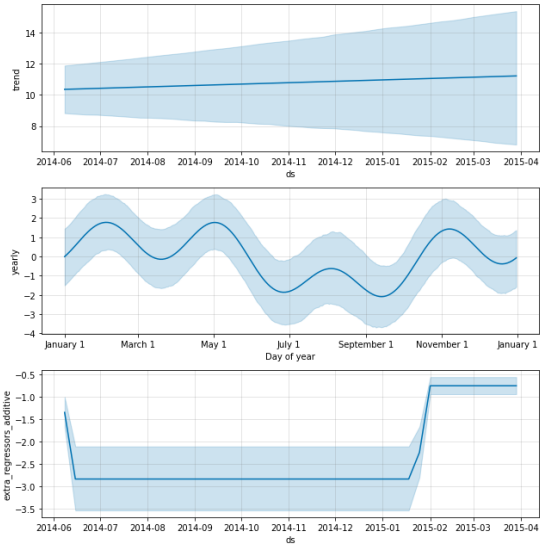



**Fig. 4:** Prophet trend, seasonality, and extra regressor(price) for FOODS_2_058 in store CA_1 over the testing period

| | sMAPE | | | | | | Mean Elasticity (STD) | |
|---|---|---|---|---|---|---|---|---|
| Item Id | Homog Linear | Indep Linear | Hierarchical Separate | Hierarchical Linear | Independent Prophet | Shared Prophet | Independent Prophet | Shared Prophet |
| FOODS_1_093 | 200% | 200% | 96.2% | 156.6% | **70.0**% | 74.7% | 2.202 (26.397) | -0.686 (3.006) |
| FOODS_1_096 | 95.4% | 176% | 127.1% | 103.6% | 41.0% | **40.6**% | -5.929 (15.040) | -4.570 (5.095) |
| FOODS_1_199 | 200% | 198% | 87.9% | 180.7% | 54.6% | **52.9**% | 1.985 (13.034) | 0.120 (0.674) |
| FOODS_2_058 | 76.2% | 118% | 80.4% | 101.1% | 84.0% | **59.6**% | -1.765 (16.010) | -2.834 (3.909) |
| FOODS_2_298 | 200% | 157% | 104.6% | 159.4% | 104% | **97.0**% | -0.057 (2.356) | 0.072 (0.565) |
| HOBBIES_1_157 | 200% | 171% | 65.29% | 173.2% | **56.8**% | 57.4% | -0.807 (2.715) | -0.628 (0.435) |
| HOBBIES_1_312 | 148.2% | 148% | 137.7% | 143.4% | 66.5% | **64.9**% | -114.472 (174.127) | -116.822 (41.911) |
| HOBBIES_1_381 | 200% | 129% | 68.6% | 60.1% | 54.1% | **54.1**% | -87.098 (190.304) | -55.950 (31.839) |
| HOBBIES_2_048 | 137.4% | 151% | 136.6% | 170.5% | **79.9**% | 83.7% | 3.046 (9.037) | 4.336 (1.979) |
| HOBBIES_2_144 | 158.9% | 161% | 161.9% | 156.9% | **138**% | 160% | 1.871 (33.579) | 2.552 (45.231) |
| HOUSEHOLD_2_155 | 200% | 137% | 134.3% | 165.9% | 47.5% | **47.4**% | -1.130 (2.537) | -1.396 (0.643) |
| HOUSEHOLD_2_159 | 200% | 174% | 69.4% | 85.5% | 48.6% | **48.0**% | 0.377 (21.997) | 1.984 (1.652) |
| HOUSEHOLD_2_235 | 200% | 183% | 120.1% | 172.8% | 53.8% | **49.9**% | -0.351 (2.759) | 0.341 (0.263) |
| HOUSEHOLD_2_296 | 200% | 104% | 81.9% | 132.6% | 54.8% | **53.8**% | 0.159 (1.670) | 0.049 (0.073) |
| HOUSEHOLD_2_376 | 200% | 198.6% | 68.6% | 121.8% | **51.0**% | 51.3% | -1.162 (1.093) | -1.111 (0.368) |
| Average value | 176% | 176% | 102% | 138% | 66.9% | **66.4**% | | |
| Median | 200% | 178% | 93% | 150% | 56% | **56**% | | |

**TABLE II:** Predictive errors and inferred elasticities from 15 products

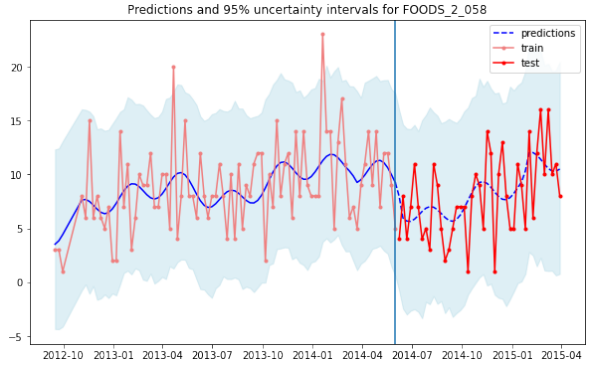



**Fig. 5:** Prophet predicted sales versus actual for FOODS_2_058 in store CA_1, over the training and testing period

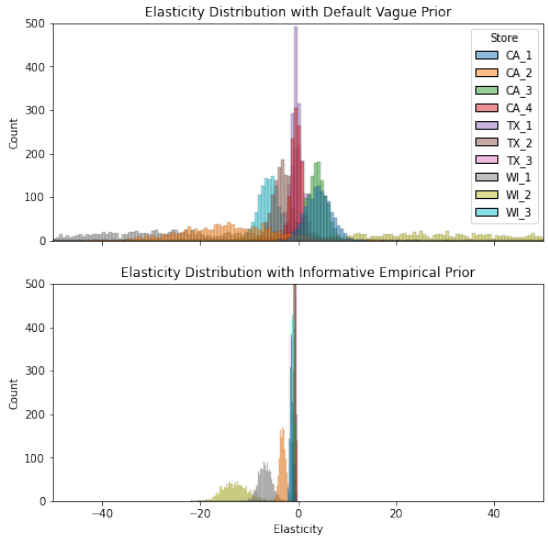



**Fig. 6:** Distribution of Elasticity Estimates with default vs empirical prior for FOODS_2_058, all 10 stores

## VI. Discussion

We reflect on our findings by comparing homogeneity and heterogeneity, hierarchical linear regression and empirical Prophet, as well as investigating why predictions for certain items were more challenging than others.

### A. Homogeneity vs Heterogeneity

Early work in this field tells us that we should not treat all stores as homogeneous and that modelling store elasticities independently would lead to a better fit. However sales data can be quite noisy with price changes quite sparse. The inference of elasticities would benefit from shared information across all stores but still having flexibility to adjust to a specific store. We find that a hierarchical approach achieves this and leads to the best model accuracy. Even when using Prophet, informative priors which hold information from other stores leads to more reasonable inferences and a slight increase in accuracy.

### B. Hierarchical vs Prophet implementation

A hierarchical model allows the group to share information by fitting all the time series co-currently with priors set on the group elasticity means and variances. However, to leverage Prophet's trend and seasonality tools each time series must be fit independently which would not allow for information exchange. Empirical Bayes allows us to take the groups elasticities estimates and use it as an informative prior before refitting to the data. Using this new starting point, the shared Prophet elasticities have significantly less variance, and many of the extreme values are no longer credible per Figure 6. Overall the inferred latent variables become more trustworthy and the predictive performance increased slightly.

### C. Variability of errors across products

Without exact knowledge of which products refer to which item IDs, we are unable to fully investigate why predictive performance for some products were significantly worse than

others. However it seems that the specific products with extremely high errors and most unreasonable inferred elasticities had some key distinguishing characteristics.

The highest error reported was from HOBBIES_2_144, we see that from Fig 7 that the item had zero sales until 2013, this is possibly a newer product and thus training data was very limited, leading to the high errors.

For HOBBIES_1_312, the predictions and sMAPE were reasonable, but the inferred elasticities are in the range of -114, well outside the other products. We see from Fig 7 there was a period of very little sales in the summer of 2013. This could potentially have been a stock-out, and the model mistakenly attributing it to pricing effects. At the bottom of Fig 7 we see a HOUSEHOLD product for which the predictions were significantly better, it does not have any sudden dips in sales and was thus easier to predict.

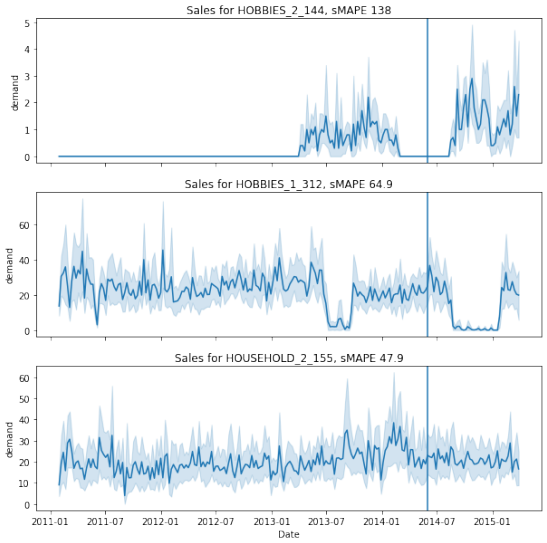



**Fig. 7:** Sales of three specific products used in testing, mean and 1 standard deviation of all 10 stores

## VII. Conclusion

In this paper we have successfully applied modern demand forecasting techniques using the STAN probabilistic programming language and Prophet modelling package on the M5 forecasting competition dataset containing Walmart. The main goal of our work was to use several approaches to infer the latent price elasticities for various products alongside with uncertainty intervals and compare their performance.

To achieve our goal, we have directly compared several Bayesian methods some of which assumed that the price elasticity of demand is homogeneous and others assuming that it is heterogeneous. We have demonstrated that the Hierarchical model assuming heterogeneity of demand clearly outperforms other linear models as it is more robust due to information sharing. Moreover, we have demonstrated that grouping observations together into segments, where the number of segments is lower than the number of observation types (i.e. stores locations), is beneficial in accuracy of the predictive outcome. Furthermore, all linear models were clearly outperformed by approaches using the Prophet. Finally, we have shown that a modified version of Prophet that allows for prior specification using Empirical Bayes on for latent variables, can result in further performance enhancement of the Prophet predictive powers in these problems.

We note that the trend and seasonality fit from Prophet has been especially useful in forecasting selected time series. To use this in conjunction with the hierarchical clustering from [Weber and Steiner, 2021] would require all stores to be fitted simultaneously. While we were able to replicate this model with a linear model in STAN, and approximate this method with an modified version of Prophet, fitting all stores simultaneously in Prophet would require an extensive overhaul of the source code. Our contributions were made on a small section of the code but there is still room for future improvements if the entire Prophet model could be fit in a hierarchical manner.

## REFERENCES


[Wal, ] Walmart Dataset md5 competition. https://www.kaggle.com/c/m5-forecasting-accuracy/data. Accessed: 2022-04-27.

[Andreyeva et al., 2010] Andreyeva, T., Long, M. W., and Brownell, K. D. (2010). The impact of food prices on consumption: A systematic review of research on the price elasticity of demand for food. *American Journal of Public Health*, 100(2):216–222. PMID: 20019319.

[Brezger and Steiner, 2008] Brezger, A. and Steiner, W. J. (2008). Monotonic regression based on bayesian p–splines: An application to estimating price response functions from store-level scanner data. *Journal of business & economic statistics*, 26(1):90–104.

[Carpenter et al., 2022] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2022). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

[Casella, 1985] Casella (1985). An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–87.

[Chindarkar and Goyal, 2019] Chindarkar, N. and Goyal, N. (2019). One price doesn't fit all: An examination of heterogeneity in price elasticity of residential electricity in india. *Energy Economics*, 81:765–778.

[Clodius and Mueller, 1961] Clodius, R. L. and Mueller, W. F. (1961). Market structure analysis as an orientation for research in agricultural economics. *American Journal of Agricultural Economics*, 43(3):515–553.

[Cornell, 1981] Cornell, B. (1981). Relative vs. absolute price changes: An empirical study. *Economic Inquiry*, 19(3):506.

[Greg M. Allenby, 1999] Greg M. Allenby, P. E. R. (1999). Marketing models of consumer heterogeneity. *Journal of Econometrics*, 39(2):89.

[Hanssens et al., 2003] Hanssens, D. M., Parsons, L. J., and Schultz, R. L. (2003). *Market response models: Econometric and time series analysis*, volume 2. Springer Science & Business Media.

[Hoffman et al., 2014] Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

[Jawad et al., 2018] Jawad, M., Lee, J. T., Glantz, S., and Millett, C. (2018). Price elasticity of demand of non-cigarette tobacco products: a systematic review and meta-analysis. *Tobacco control*, 27(6):689–695.

[Johnson and Helmberger, 1967] Johnson, A. and Helmberger, P. (1967). Price elasticity of demand as an element of market structure. *The American Economic Review*, 57(5):1218–1221.

[Kreinovich et al., 2014] Kreinovich, V., Nguyen, H. T., and Ouncharoen, R. (2014). How to estimate forecasting quality: A system-motivated derivation of symmetric mean absolute percentage error (smape) and other similar characteristics.

[Markham, 1951] Markham, J. W. (1951). The nature and significance of price leadership. *The American Economic Review*, 41(5):891–905.

[Ng, 2016] Ng, Y.-K. (2016). Are unrealistic assumptions/simplifications acceptable? some methodological issues in economics. *Pacific Economic Review*, 21(2):180–201.

[Otter et al., 2004] Otter, T., Tüchler, R., and Frühwirth-Schnatter, S. (2004). Capturing consumer heterogeneity in metric conjoint analysis using bayesian mixture models. *International Journal of Research in Marketing*, 21(3):285–297.

[Pendzialek et al., 2016] Pendzialek, J. B., Simic, D., and Stock, S. (2016). Differences in price elasticities of demand for health insurance: a systematic review. *The European Journal of Health Economics*, 17(1):5–21.

[Steiner et al., 2007] Steiner, W. J., Brezger, A., and Belitz, C. (2007). Flexible estimation of price response functions using retail scanner data. *Journal of retailing and consumer services*, 14(6):383–393.

[Taylor and Letham, 2018] Taylor, S. J. and Letham, B. (2018). Forecasting at Scale. *The American Statistician*, 72(1):37–45.

[Weber and Steiner, 2021] Weber, A. and Steiner, W. J. (2021). Modeling price response from retail sales: An empirical comparison of models with different representations of heterogeneity. *European Journal of Operational Research*, 294(3):843–859.