Handelshøyskolen BI

GRA 19703 Master Thesis

Thesis Master of Science 100% - W

Predefinert informasjon

Startdato:	16-01-2022 09:00	**Termin:**	202210
Sluttdato:	01-07-2022 12:00	**Vurderingsform:**	Norsk 6-trinns skala (A-F)
Eksamensform:	T		
Flowkode:	202210||10936||IN00||W||T		
Intern sensor:	(Anonymisert)		

Deltaker

Navn:	Xinran Ji og Binh Duc Le

Informasjon fra deltaker

Tittel *:	Using Bayesian Linear Models and Deep Neural Networks for Decomposing the Performance Effect of Promotion
Navn på veileder *:	Auke Hunneman

Inneholder besvarelsen konfidensielt materiale?:	Nei	**Kan besvarelsen offentliggjøres?:**	Ja

Gruppe

Gruppenavn:	(Anonymisert)
Gruppenummer:	112
Andre medlemmer i gruppen:	

Master Thesis

Using Bayesian Linear Models and Deep Neural Networks for Decomposing the Performance Effect of Promotion

Xinran Ji, Binh Le

Supervisor: **Auke Hunneman**

Table of Contents

Abstract 3

I. Introduction 4

II. Theoretical Background 6

2.1 Unified Framework for sales promotion decomposition 6

2.2. Multiple Linear Regression 7

2.3. Bayesian linear regression 7

2.3.1 Bayesian Inference 8

2.3.2 Linear model: conjugate priors 9

2.3.3 MCMC methods 10

2.4. Deep Neural Networks 11

III. Model Specification 13

3.1 Standard Decomposition via multivariate linear regression 13

3.2 Standard Decomposition via Bayesian linear regression 16

3.3 Deep Neural Network Prediction 17

IV. Data 18

V. Result 19

5.1 Standard Decomposition via multivariate linear regression 19

5.2 Bayesian Standard Decomposition 21

5.2.1 Convergence 21

5.2.2 Effects Interpretation 23

5.3 Predictive Modelling 27

VI. Conclusion 28

Appendix 1 30

Reference 39

Abstract

Price promotions can drive up short-term sales substantially. To establish whether and how a business can truly benefit from a price promotion bump, previous research has proposed multivariate linear regression models decomposing the performance effect of promotion into three constituent parts: cross-brand effects, cross-period effects, and category expansion effects. However, under missing data conditions, the original models fail to perform well, with a large part of constituent effects unexplainable. In this study, we propose a system of models that possess both explanatory and predictive power and can directly work on an imperfect dataset. Results show that the Bayesian linear regression models are able to conduct standard decomposition by demonstrating uncertainty and that our proposed deep neural networks drive predictive performance up to 42.14%.

I. Introduction

Sales promotion usually comes with a sales bump for the promoted item (Neslin, 2002, p. XI). However, the sales increase does not always bode well. To illustrate, suppose one promoted brand experiences a large sales surge. The increase can come from other brands of the same product category dropping sales (cross-brand effect) or from bringing in sales from other time periods (e.g., through stockpiling). The retailer can hardly gain any significant benefit from these two effects unless there exists a possible difference in margins. In comparison, both retailers and manufacturers would benefit from price promotions with conspicuous category consumption increase (category-expansion effect). Thus, it is meaningful to decompose the effect of a price promotion into its elemental parts.

For this purpose, Van Heerde et al. (2004) developed a series of standard decomposition models based on multivariate linear regression. The models break down the own-brand sales effect into net cross-brand, cross-period, and category-expansion effects. However, in real life, data is usually messy and incomplete. It is often unfeasible for retailers or manufacturers to track and collect all the relevant data perfectly. The reasons can be sensor failure, data transformation errors or more commonly that costly data storage systems constrain the businesses to collect based on priority. In our case, in the event of missing regular price and display form data, the previously proposed standard decomposition models fail to perform well with the incapability to model uncertainty, model fits dropping and a large part of constituent effects unexplainable.

In this study, we use store-level scanner data and build a system of explanatory and predictive models based on Bayesian linear regression and deep neural networks. Directly working on less and incomplete data, the Bayesian linear regression models proposed by us are able to conduct standard decomposition through demonstrating uncertainty. Furthermore, using RMSE as the evaluation metric, our proposed deep neural networks drive predictive performance up to 42.14%.

The paper is organized as follows. In section §2, we present related theoretical framework. §3 discusses our scanner dataset, and §4 shows the specifications of previous models and models proposed by us. We provide empirical results in §5, and in the last section, §6 we present the conclusions, implications, and limitations for future research.

II. Theoretical Background

2.1 Unified Framework for sales promotion decomposition

Van Heerde et al. (2004) adopt a unit-based approach to decompose sales-promotion effects into its constituent parts. In their work, they define secondary demand effects as the net effect of a promotion on the sales of nonpromoted brands in the same week, category, and store, and primary demand effects as the net effect of a promotion on category sales within the same week in the same store.

Primary demand effects are further split into cross-period effects and category-expansion effects. The former stand for short-period sales shift caused by lead effects (Doyle & Saunders, 1985; Gonul & Srinivasan, 1996; Kalwani et al., 1990; Macé & Neslin, 2004; van Heerde et al., 2000) and lagged effects (e.g., van Heerde et al., 2000). The latter effects may consist of increased consumption (Ailawadi & Neslin, 1998; Assuncao & Meyer, 1993; Chandon & Wansink, 2002), deal-to-deal purchasing (Krishna, 1994), category switching (Walters, 1991), and store switching (Bucklin & Lattin, 1992).

Previous studies form a unified framework capable of defining sales-promotion effects mathematically. First, a series of criterion variables are specified for both the own-brand sales effect and its constituent effect. More specifically, variable OBS which stands for own-brand sales set as the criterion variable for the own-brand sales effect. Similarly, cross-brand sales, cross-period categorical sales, and total category sales are set as criterion variables for cross-brand effect, cross-period effect, and category-expansion effect, respectively. Then, they regress these criterion variables on the same set of independent variables constructed by promotion variables of interest such as price index, cross-brand price index, and regular price, and sensible covariates such as the lag terms of the price index. Across models, only the left-hand side of the equations differs. This stability makes it possible to achieve a decomposition of performance effect of promotion by comparing the parameters of interest across equations.

2.2. Multiple Linear Regression

The regression analysis aims to uncover the correlations between two or more variables with causal effects and make predictions for the topic using the relation. The regression models with one dependent variable and more than one independent variable are known as multivariate regression analysis (Büyüköztürk, 2002; Köksal, 1985; Tabachnick & Fidell, 1996).

In multivariate regression analysis, an attempt is made to account for the variation of the independent variables in the dependent variable synchronically (Ünver & Gamgam, 1999). The multivariate regression analysis model is formulated as:

$$Y = \beta_0 + \beta_1 x_1 + .. + \beta_n x_n + \varepsilon$$

where, Y = dependent variable

x_i = independent variable

β_i = parameter

ε = error term

The assumptions of multivariate regression analysis are normal distribution of errors, linearity, freedom from extreme values and having no multiple ties between independent variables (Büyüköztürk, 2002)

2.3. Bayesian linear regression

The way we estimate parameters of a multivariate linear regression model is to minimize its loss function Mean Squared Error (MSE) either analytically or through using gradient descent algorithms.

$$MSE_1 = \frac{1}{N}\sum_{i=1}^{n}(y_i - y'_i)^2$$

where $y_i$is the actual value and y'_i the predicted value derived from provided variables. In our case,

$$MSE_2 = \tfrac{1}{N}\textstyle\sum_{i=1}^{n}(y_i - y''_i)^2$$

where $y_i$is the same actual value with MSE_1 and y''_i the predicted value derived from approximated variables due to the failure of gathering complete data.

Thus, the missing regular price data inject uncertainty into parameter estimation compared to the case with complete data. However, multivariate linear regression adopts the frequentist point estimate method, which only summaries the posterior distribution. Interpreting models via tables of point estimates throws out the uncertainty of the parameters added due to incomplete data.

In contrast, from a Bayesian perspective, a parameter is a distribution or density displaying copious information regarding uncertainty (e.g., the most probable values, the degree of skewness, and credible interval). Therefore, adopting Bayesian inference allows us to decompose sales promotion effect through modelling both intrinsic uncertainty and extrinsic uncertainty added by the missing data.

We introduce the Bayesian linear regression models by illustrating the Bayesian perspective, prior distributions and MCMC estimation.

2.3.1 Bayesian Inference

Utilizing probabilistic models, the Bayesian perspective focuses on making inferences regarding target parameters of interest through using data and subjective analysis. The adoption of subjective analysis characterizes the Bayesian analysis approach. More specifically, one would first gather all the available information about the parameter of interest to derive a prior distribution accordingly, and the choice of the prior distribution can affect the final inference hinging on data availability.

Among the types of prior distributions, the conjugate priors are set after one makes assumptions based on mathematical or computational convenience and have the same parametric form of the likelihood function. As a result of conjugacy, one can calculate the posterior distribution using an analytical approach and produce posterior samples directly.

Mathematically speaking, let $\theta = (\theta_0, \ldots, \theta_{d-1})^T$ be a vector of parameter of interest, and $y = (y_1, \ldots, y_n)^T$ a vector of realization of the random variables with distribution $p(y_i|\theta)$. The likelihood function of y_i is as follows:

$$\mathcal{L}(\theta|y) = \prod_{i=1}^{n} p(yi|\theta) \tag{1}$$

Equation (1) contains all the information from the observations y_i given θ. According to the Bayes' theorem, θ is the joint posterior distribution:

$$p(\theta|y) = \frac{\mathcal{L}(\theta|y)p(\theta)}{\int_{\odot} \mathcal{L}(\theta|y)p(\theta)d\theta} \tag{2}$$

where ⊙ is the parametric space of θ and p(θ) stands for the prior distribution. $\int_{\odot} \mathcal{L}(\theta|y)p(\theta)d\theta$ is the marginal distribution of y and independent on θ, the equation (2) can thus be expressed as:

$$p(\theta|y) \propto Ꝉ(\theta|y)p(\theta) \quad (3)$$

The predictive posterior distribution which makes the prediction of unknown values of the dependent variable y' and the marginal distribution of y:

$$p(y`|y) = \int_{\odot} p(y`|\theta, y)p(\theta|y)d\theta, \qquad y` \sim p(y`|\theta, y)$$

$$p(y) = \int_{\odot} Ꝉ(\theta|y)p(\theta)d\theta.$$

Bayesian inference takes all the available information into account. The prior information is included via the prior distribution and combined with the likelihood function that represents the data. The inference is conducted based on the posterior distribution.

2.3.2 Linear model: conjugate priors

A Bayesian linear model in the form

$$\boldsymbol{y} = X\beta + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_d)$$

Where $\sigma^2 > 0$, $\mathbf{I}_d$ stands for identity matrix, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{d-1})^{\mathrm{T}}$ a $d \times 1$ vector, $\mathbf{X}$ an $n \times d$ design matrix, and we assume that $\varepsilon_i's$ are independent. The likelihood function is also:

$$f_y(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\}$$

From Bayesian perspective, the inference process centers around data and prior information. Thus, we assume a $\boldsymbol{N}_d(\boldsymbol{m}, \sigma^2\mathbf{V})$, which is a conjugate prior distribution for $\boldsymbol{\beta}|\sigma^2$ as follows

$$f(\boldsymbol{\beta}|\sigma^2, \mathbf{m}, \mathbf{V}) = (2\pi\sigma^2)^{-d/2}|\mathbf{V}|^{-d/2}\exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \mathbf{m})^{\mathrm{T}}\mathbf{V}^{-1}(\boldsymbol{\beta} - \mathbf{m})\}$$

For σ^2, we also set a conjugate prior distribution given by an Inverse Gamma denoted by IG(a,b) in the form of

$$f(\sigma^2|a, b) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{a-1}\exp\{-\frac{b}{\sigma^2}\}$$

where $a > 0$ and $b > 0$. Since we have the likelihood function and the proper priors, we can then find the posterior distribution to make inference on the parameters $\boldsymbol{\beta}$ and σ^2. Using the Bayes' theorem, we have

$$f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) = \frac{f_y(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)f(\boldsymbol{\beta}|\sigma^2, \mathbf{m}, \mathbf{V})f(\sigma^2|a, b)}{f_y(\boldsymbol{y})}$$

$$\propto\ f_y(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2)f(\boldsymbol{\beta}|\sigma^2,\mathbf{m},\mathbf{V})f(\sigma^2|a,b),$$

$$\propto\ (\sigma^2)^{-n/2-d/2+a-1}(\exp\left\{-\frac{A}{2\sigma^2}\right\},$$

Where,

$$A=\ (\mathbf{y}-\ \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\ \mathbf{X}\boldsymbol{\beta})+\ (\boldsymbol{\beta}-\mathbf{m})^{\mathrm{T}}\mathbf{V}^{-1}(\boldsymbol{\beta}-\mathbf{m})+2b,$$

$$=\ \mathbf{y}^{\mathrm{T}}\mathbf{y}-\mathbf{y}\mathbf{X}\boldsymbol{\beta}-\ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}+\ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}+\ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{\beta}-\ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{m}-\mathbf{m}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{\beta}+\ \mathbf{m}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{m}+\ 2b,$$

$$=\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\ \mathbf{V}^{-1})\boldsymbol{\beta}+\ \boldsymbol{\beta}^{T}(\mathbf{X}^{\mathrm{T}}\mathbf{y}+\mathbf{V}^{-1}\mathbf{m})+(\mathbf{m}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{m}+2b+\mathbf{y}^{\mathrm{T}}\mathbf{y})-\ (\mathbf{y}^{\mathrm{T}}\mathbf{X}+\ \mathbf{m}^{\mathrm{T}}\mathbf{V}^{-1})\boldsymbol{\beta}.$$

For convenience, let $\boldsymbol{\Lambda}\ =\ (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\ \mathbf{V}^{-1})^{-1}$ a $d\ \times\ d$ matrix and $\boldsymbol{\mu}\ =(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{V}^{-1})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y}+\ \mathbf{V}^{-1}\mathbf{m})$ a $d \times 1$ vector. Therefore,

$$A=\ \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta}-\ \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu}-\ \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta}+\ \mathbf{m}^{\mathrm{T}}\mathbf{V}^{-1}\mathrm{m}+2b+\ \mathbf{y}^{\mathrm{T}}\mathbf{y},$$

$$=(\boldsymbol{\beta}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})-\ \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu}+\ \mathbf{m}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{m}+2b+\ \mathbf{y}^{\mathrm{T}}\mathbf{y}.$$

Finally, the joint posterior distribution for $\boldsymbol{\beta}$ and σ^2 is given by

$$f(\boldsymbol{\beta},\sigma^2|\mathrm{y},\ \mathbf{X})\propto f_y(\boldsymbol{y}|\ \mathbf{X},\boldsymbol{\beta},\sigma^2)f(\boldsymbol{\beta}|\sigma^2,\mathbf{m},\mathbf{V})f(\sigma^2|a,\mathrm{b}),$$

$$\propto\sigma^{2^{-\frac{d}{2}}}\exp\left\{-\frac{(\boldsymbol{\beta}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}}{2\sigma^2}\right\}$$

$$\times(\sigma^2)^{-\frac{n}{2}+a-1}\exp\left\{-\frac{\mathbf{m}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{m}-\ \boldsymbol{\mu}^{T}\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu}+2b+\ \mathbf{y}^{\mathrm{T}}\mathbf{y}}{2\sigma^2}\right\}. \tag{4}$$

Therefore, the equation (4) shows that the posterior distribution $f(\boldsymbol{\beta},\sigma^2\,|\mathbf{y},\ \mathbf{X})$ is proportional to the multiplication of kernels of the $\boldsymbol{N}_n(\boldsymbol{\mu},\sigma^2\boldsymbol{\Lambda})$ and IG $(a^*=\ -\frac{n}{2}+a, b^*=b+\ \frac{\mathbf{m}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{m}-\ \boldsymbol{\mu}^{T}\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu}+\ \mathbf{y}^{\mathrm{T}}\mathbf{y}}{2})$.

2.3.3 MCMC methods

This section introduces the MCMC algorithm used in our study: Metropolis-Hastings. When it is unfeasible to find an analytical solution to compute a finite integral

$$\int g(\theta)p(\theta)d\theta, \tag{5}$$

Where g(*) is an integrable function and p(*) is a probability density function. Independent and identical distributed samples are drawn from p(*) and estimated at g(*) and then averaged. Under affluent samples, the Strong Law of Large Numbers states that this average value eventually converges to (5). When directly sampling from p(*) is impossible, Metropolis-Hastings is an alternative to draw samples from p(*) following a Markovian dependence structure.

Suppose we want to draw a sample from the posterior distribution $p(\theta|y)$. The Metropolis-Hastings proposes distribution $q(\theta^{(t-1)})$ that generates candidate values θ^* that are accepted as values from $p(\theta|y)$ with a certain probability (Ahmed Ali et al., 2014)

The Metropolis algorithm is presented below:

Algorithm Metropolis-Hastings

1. Initialize t = 1 and define the initial values ${\theta_1}^{(0)}, {\theta_2}^{(0)}, \dots, {\theta_d}^{(0)}$ for the vector $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_{d-1})$;
2. Sample $\boldsymbol{\theta}^*$ from the proposal distribution $q(\boldsymbol{\theta}^{(t-1)})$;
   a. Compute
   $$\alpha\left(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*\right) = \min\{1, \frac{p(\boldsymbol{\theta}^*|\mathbf{y})q(\boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)}|y)q(\boldsymbol{\theta}^{(*)})}\},$$
   b. Compute $u \sim U[0,1]$. If $u < \alpha(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*)$, then $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$, otherwise, $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$;
3. Take $t = t+1$ and return to step 2 until the desired posterior sample has been obtained.

2.4. Deep Neural Networks

Previous research has explored the implementation of neural network on predicting online product sales via promotion strategies (Chong et al., 2016). The result suggests that the neural network model shows reliability with low average relative error. Theoretically, when the linear restriction of the model form is relaxed, the possible number of non-linear structures that can be used to predict sales is numerous. Artificial neural networks (ANNs) are flexible computing frameworks for modeling a broad range of non-linear problems. One significant advantage of the ANN models over other classes of non-linear models is that ANNs are universal approximators which can approximate a large class of functions with a high degree of accuracy (Zhang et al., 2001). Drawing on the practical and theoretical advantages of neural networks, we explore the predictive performance of deep neural network models in our case.

We provide an introduction to ANNs and deep neural networks (DNNs) and how to implement them for regression. In essence, a DNN is an ANN with many hidden layers. Both of them are represented by a network of nodes. Each node first sums its inputs and then applies an activation function onto the result to achieve a non-linear transformation.

Let x_i be the ith input to the node of a neural network, w_i be the weight of the ith input, b_i be the bias of the ith input, n the total number of inputs, o the output of the node and σ the activation function. For each node:

$$o = \sigma\left(\sum_{i=1}^{n} w_i x_i + b_i\right),$$

In our study, we use RELU as an activation function:

$$\sigma(x) = \max(0, x)$$

For regression problems, the output layer is a layer of a single node with a linear function as its activation function:

$$o = \sum_{i=1}^{n} w_i x_i + b_i$$

Fig 1 below presents the structure artificial neural network. The arrows show how the output of one node in a certain layer gets fed forward as the input to the nodes in the next layer. In our study, each node is connected with all the nodes in the next layer, thus constructing a fully-connected neural network. The input layer is symbolled with V, the output layer with O, and the hidden layers are labelled with $H_{j,k}$, where j represents the layer number and k the node number.

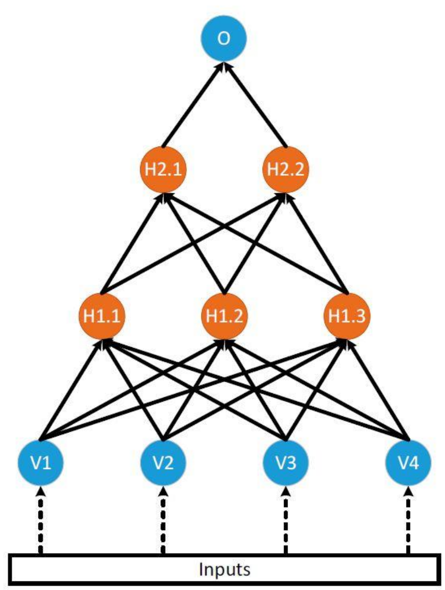


Figure 1. A feedforward neural network with two hidden layers

The training process of neural networks aims to minimize the loss function by updating weights and biases using the backpropagation algorithm. Each time the backpropagation runs over the entire training data, it forms an epoch. Typically, many epochs need to perform until the algorithm reaches a termination criterion, such as a fixed number of epochs or the error falling below a threshold.

III. Model Specification

3.1 Standard Decomposition via multivariate linear regression

The standard decomposition divides the own-brand sales bump into three sources: cross-brand effect, cross-period effect, and category-expansion effect. In calculation, cross-brand effects are the shift from other brands' sales to the promoted brand's sales in this week when the price promotion of the focal brand exists. Cross-period effects represent the shift of both own-brand and cross-brand sales from pre- and post-promotion period to the price promotion period. Since the cross-period effect is defined at the category level, it contains accelerated brand switching. Lastly, category-expansion effects are expansion effects shown in own-brand sales that cannot be ascribed to other brands or periods.

In total, the soda category contains J brands. When calculating cross-period and category-expansion effects, we define the time window as [t-T*, t+T] and estimate the models pooled across stores (i=1, … l). Following the work of van Heerde et al. (2004), in order to derive equal effects across stores with different category sales volumes, we calculate the divisions of all criterion variables by the average category sales per store (CS_i).

Let S_{ijt} be the unit sales of brand j in store i in week t, the criterion variables for the standard decomposition are calculated as follows:

OBS_{ijt}= - $\frac{S_{ijt}}{CS_i}$, CBS_{ijt}= $\sum_{\substack{k=1 \\ k\neq j}}^{J} \frac{S_{ikt}}{CS_i}$,

$PPCS_{it}$=$\sum_{\substack{s=-T* \\ s\neq 0}}^{T} \sum_{k=1}^{J} \frac{S_{ikt+s}}{CS_i}$, TCS_{it}=$-\sum_{s=-T*}^{T} \sum_{k=1}^{J} \frac{S_{ikt+s}}{CS_i}$

And in the end,

$$OBS_{ijt} = CBS_{ijt} + PPCS_{it} + TCS_{it} \tag{6}$$

where, minus Own-Brand Sales (OBS) stands for the own-brand effect, Cross-Brand Sales (CBS) represents cross-brand effect, Pre- and Post-Promotion Category Sales (PPCS) stands for the cross-period effect and Total Category Sales (TCS) stands for the category-expansion effect.

In this study, we adopt a price index variable also used in ACNielsen's Scan∗Pro model (Wittink et al., 1988) to distinguish promotional prices from regular prices. The price index variable (PI) for each item equals the actual unit price divided by its regular unit price. In the case of the changes in the regular unit price, we update the price index according to the changed regular unit price. In the end, the price index only reflects temporary price promotion.

Standard decomposition linear regression model specification:

$$OBS_{ijt} = \alpha_{1j} + \beta_{ob,j} PI_{ijt} + \gamma_{11j} CPI_{ijt} + \gamma_{21j} RP_{ijt} + \gamma_{31j} CRP_{ijt} + \sum_{\tau=T+T*+1}^{T_{max}-T-T*} \gamma_{41,\tau j} W_t + \sum_{\tau=1}^{T+T*} \gamma_{51,\tau j} PI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{61,\tau j} PI_{ijl-\tau} + \sum_{\tau=1}^{T+T*} \gamma_{71,\tau j} CPI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{81,\tau j} CPI_{ijl-\tau} + \mu_{1ijt} \quad (7)$$

$$CBS_{ijt} = \alpha_{2j} + \beta_{cb,j} PI_{ijt} + \gamma_{12j} CPI_{ijt} + \gamma_{22j} RP_{ijt} + \gamma_{32j} CRP_{ijt} + \sum_{\tau=T+T*+1}^{T_{max}-T-T*} \gamma_{42,\tau j} W_t + \sum_{\tau=1}^{T+T*} \gamma_{52,\tau j} PI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{62,\tau j} PI_{ijl-\tau} + \sum_{\tau=1}^{T+T*} \gamma_{72,\tau j} CPI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{82,\tau j} CPI_{ijl-\tau} + \mu_{2ijt} \quad (8)$$

$$PPCS_{ijt} = \alpha_{3j} + \beta_{cp,j} PI_{ijt} + \gamma_{13j} CPI_{ijt} + \gamma_{23j} RP_{ijt} + \gamma_{33j} CRP_{ijt} + \sum_{\tau=T+T*+1}^{T_{max}-T-T*} \gamma_{43,\tau j} W_t + \sum_{\tau=1}^{T+T*} \gamma_{53,\tau j} PI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{63,\tau j} PI_{ijl-\tau} + \sum_{\tau=1}^{T+T*} \gamma_{73,\tau j} CPI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{83,\tau j} CPI_{ijl-\tau} + \mu_{3ijt} \quad (9)$$

$$TCS_{ijt} = \alpha_{4j} + \beta_{ce,j} PI_{ijt} + \gamma_{14j} CPI_{ijt} + \gamma_{24j} RP_{ijt} + \gamma_{34j} CRP_{ijt} + \sum_{\tau=T+T*+1}^{T_{max}-T-T*} \gamma_{44,\tau j} W_t + \sum_{\tau=1}^{T+T*} \gamma_{54,\tau j} PI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{64,\tau j} PI_{ijl-\tau} + \sum_{\tau=1}^{T+T*} \gamma_{74,\tau j} CPI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{84,\tau j} CPI_{ijl-\tau} + \mu_{4ijt} \quad (10)$$

for i= 1, …, I (stores), j= 1, …, J(brands), and t= T+ T*+1, …, T_{max} – T-T* (weeks), where,

PI_{ijt} = price index for brand j in store i in week t; PI_{ijt} equals 1−d/100 if there is a d percent discount for brand j with support l in week t in store i, and 1 otherwise.

CPI_{ijt} = average price index across brands k, k=1, …, J, kj.

RP_{ijt} = regular price for brand j in store i in week t.

CRP_{ijt} = average regular price across brands k, k=1, …, J, kj.

W_t = week dummy: 1 for week t, 0 otherwise.

μ_{1ijt}, μ_{2ijt}, μ_{3ijt}, μ_{4ijt} = disturbance terms for brand j in store i in week t for equations

T* is the number of leads, T is the number of lags, and T_{max} is the total number of weeks

$\beta_{ob,j}$ = for brand j, the effect on minus own-brand sales of the price index

$\beta_{cb,j}$ = for brand j, the cross-brand effect of the price index

$\beta_{cp,j}$ = for brand j, the cross-period effect of the price index

$\beta_{ce,j}$ = for brand j, the category-expansion effect of the price index

For equations (7)-(10), each dependent variable is regressed on the same set of independent variables. Thus, we derive the following equation from equation (6):

$$\beta_{ob,j} = \beta_{cb,j} + \beta_{cp,j} + \beta_{ce,j} \tag{11}$$

Put in words, equation (11) shows that for each brand, the own-brand effect equals the sum of its cross-brand effect, cross-period effect, and category-expansion effect.

van Heerde et al. (2003, 2004) define the fraction of each effect as follows:

fraction cross-brand effect= $\frac{\beta cb,j}{\beta ob,j}$

fraction category-expansion effect=$\frac{\beta ce,j}{\beta ob,j}$

fraction cross-period effect=$\frac{\beta cp,j}{\beta ob,j}$

RP_{ijt} and CRP_{ijt} serve as control for regular price effects. Given the lack of regular price data, we approach each brand's regular price in store i at week t by calculating the mode value. Similarly, CPI_{ijt} controls for cross-brand price-promotion effects. Besides, weekly indicator variables W_t accounts for the seasonal effect in sales. We also put in the same set of lead and lagged variables for PI and CPI to achieve mathematical consistency.

The time window T for post-promotion period and T* for pre-promotion period should be as smallest as possible to prevent the diminishing of the number of degrees of freedom due to the rapid increase of the number of independent variables and decrease of the sample size. In our study, we use T = T*= 6, based on van Heerde et al. (2000), Macé & Neslin (2004), Nijs et al. (2001), and Pauwels et al. (2002).

3.2 Standard Decomposition via Bayesian linear regression

We use the same equations (Equation (7) - (10)) when conducting Bayesian linear regression. The residuals follow Gaussian distributions with means of 0 and unknown standard deviations.

$$e_{1ijt} \sim N(0, \rho_1) \tag{12}$$

$$e_{2ijt} \sim N(0, \rho_2) \tag{13}$$

$$e_{3ijt} \sim N(0, \rho_3) \tag{14}$$

$$e_{4ijt} \sim N(0, \rho_4) \tag{15}$$

Next, we must define the likelihood for the data and priors for all parameters in the model. To keep the models most consistent with least-squares regression, we use the most common choice-Gaussian distribution:

$$OBS_{ijt} \sim N(u_1, \rho_1) \tag{16}$$

$$CBS_{ijt} \sim N(\mu_2, \rho_2) \tag{17}$$

$$PPCS_{ijt} \sim N(\mu_3, \rho_3) \tag{18}$$

$$TCS_{ijt} \sim N(\mu_4, \rho_4) \tag{19}$$

which says that each dependent variable follows a Gaussian (normal) distribution with a mean= μ and a standard deviation. The mean is equal to the right side of equations (7)- (10), and the standard deviation is the same standard deviation as in equation (12)- (15). Thus, equations (16)- (19) can be rewritten into:

$$OBS_{ijt} \sim N(\alpha_{1j} + \beta_{ob,j} PI_{ijt} + \gamma_{11j} CPI_{ijt} + \gamma_{21j} RP_{ijt} + \gamma_{31j} CRP_{ijt} + \sum_{\tau=T+T*+1}^{T_{max}-T-T*} \gamma_{41,\tau j} W_t + \sum_{\tau=1}^{T+T*} \gamma_{51,\tau j} PI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{61,\tau j} PI_{ijl-\tau} + \sum_{\tau=1}^{T+T*} \gamma_{71,\tau j} CPI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{81,\tau j} CPI_{ijl-\tau} + e_{1ijt}, \rho_1)$$

$$CBS_{ijt} \sim N(\alpha_{2j} + \beta_{cb,j} PI_{ijt} + \gamma_{12j} CPI_{ijt} + \gamma_{22j} RP_{ijt} + \gamma_{32j} CRP_{ijt} + \sum_{\tau=T+T*+1}^{T_{max}-T-T*} \gamma_{42,\tau j} W_t + \sum_{\tau=1}^{T+T*} \gamma_{52,\tau j} PI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{62,\tau j} PI_{ijl-\tau} + \sum_{\tau=1}^{T+T*} \gamma_{72,\tau j} CPI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{82,\tau j} CPI_{ijl-\tau} + e_{2ijt}, \rho_2)$$

$$PPCS_{ijt} \sim N(\alpha_{3j} + \beta_{cp,j} PI_{ijt} + \gamma_{13j} CPI_{ijt} + \gamma_{23j} RP_{ijt} + \gamma_{33j} CRP_{ijt} + \sum_{\tau=T+T*+1}^{T_{max}-T-T*} \gamma_{43,\tau j} W_t + \sum_{\tau=1}^{T+T*} \gamma_{53,\tau j} PI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{63,\tau j} PI_{ijl-\tau} + \sum_{\tau=1}^{T+T*} \gamma_{73,\tau j} CPI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{83,\tau j} CPI_{ijl-\tau} + e_{3ijt}, \rho_3)$$

$$TCS_{ijt} \sim N(\alpha_{4j} + \beta_{ce,j}PI_{ijt} + \gamma_{14j}CPI_{ijt} + \gamma_{24j}RP_{ijt} + \gamma_{34j}CRP_{ijt} + \sum_{\tau=T+T*+1}^{T_{max}-T-T*} \gamma_{44,\tau j}W_t + \sum_{\tau=1}^{T+T*} \gamma_{54,\tau j}PI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{64,\tau j}PI_{ijl-\tau} + \sum_{\tau=1}^{T+T*} \gamma_{74,\tau j}CPI_{ijl+\tau} + \sum_{\tau=1}^{T+T*} \gamma_{84,\tau j}CPI_{ijl-\tau} + e_{4ijt}, \rho_4)$$

The bambi package in Python will intelligently generate priors for all model terms and standard deviations by loosely scaling them to the observed data. The generated priors for parameters of PI and standard deviation are (see Appendix 1 for the priors of all parameters):

$$\beta_{ob,j} \sim N(0,2.5972)$$

$$\rho_1 \sim HalfStudentT(4,1.0123)$$

$$\beta_{cb,j} \sim N(0,2.5808)$$

$$\rho_2 \sim HalfStudentT(4,1.006)$$

$$\beta_{cp,j} \sim N(0,2.5653)$$

$$\rho_3 \sim HalfStudentT(4,0.9999)$$

$$\beta_{ce,j} \sim N(0,2.5640)$$

$$\rho_4 \sim HalfStudentT(4,0.9994)$$

3.3 Deep Neural Network Prediction

In order to achieve high-performance predictive modeling results, we construct deep neural networks for each equation (7)-(10). However, due to its nonparametric and black-box nature, it becomes unable to decompose the own-brand sales effect.

The unavoidable randomness emerging from the results when training any type of neural network puts a single developed and trained model with good performance into question: Is the model performing well due to fine-tuned parameters or randomness? This study follows Hansen & Salamon's method (1990) and tackles this problem using grid-search cross-validation and an ensemble of similar neural networks.

Table 1. Hyperparameter tuning result for each model

	Optimizer	Batch size	Epochs
OBS	‘adam’	2000	70
CBS	'rmsprop'	500	50
PPCS	'rmsprop'	3000	50
TCS	'rmsprop'	500	50

Large deep neural networks with four hidden layers of 20, 20, 40, and 30 neurons, respectively, are constructed. We set the activation function as 'RELU' for each hidden layer to avoid vanishing gradient or exploding gradient problem and 'linear' for the output layer to tackle regression problems.

IV. Data

We have collected one-year worth of weekly, store-level scanner data from a local supermarket chain, starting in the 1st week of 2021 and ending at the week 53 of 2021. The dataset contains week number, store id (475 stores in total), sales and quantity for each transaction but fails to include regular price data.

We select the soda category as our main target category due to the convenience of extracting brands. The original dataset does not contain brand information, which demands us to extract brands from the Sales description, a text file with brands randomly embedded. The brands in the soda category show an aligned pattern and thus can be collected through algorithms. Also, we limit our analysis to seven brands sold in the category due to data completeness.

To tackle on the problem of missing regular price data in the original dataset, we approximate regular price by calculating the mode value. If the item's unit price is less than 5% of the regular price, it is considered under a promotion and set with a promotional tag. There is no information about the supermarket's feature advertising or display activity provided during the chosen period.

The descriptive statistics are presented below.

Table 2. Descriptive Statistics for Dataset

Brands	Price promotion (%)	Total sales (million kroner)	Total quantity (l)
Coca-cola	81.28	31.65	2,786,584.25
Cola	82.79	14.63	1,935,838.5
Pepsi	1.83	0.76	53,877
Solo	33.87	0.99	60,444
Sprite	4.48	9.20	592,830
Fanta	5.04	7.21	567,017.75
Hansa	5.56	1.06	78,756.55

For deep neural network prediction and standard decomposition using multiple linear regression, we conduct train-test split with train: test ratio = 8:2 before feeding into models. Each train set contains 104,034 rows. For each Bayesian linear regression model, we randomly sample 3,000 rows from the corresponding train set after train-test split as above.

V. Result

In this section, we obtain the results of our explanatory and predictive models after conducting a series of model estimations and evaluations. For the explanatory modeling part, we present results for the "standard decomposition via multiple linear regression" in §5.1 and results for the "Bayesian standard decomposition" in §5.2. For the predictive modeling part, we present evaluation metrics of the deep neural networks in §5.3.

5.1 Standard Decomposition via multivariate linear regression

We deploy multiple linear regression models and decompose the own-brand sales effect into cross-brand, cross-period, and category expansion effects in Table 3.

Table 3. Average Decomposition of Constant Price Effects

Category: soda	Own-brand effect $\hat{\beta}_{ob}$	Cross-brand effect $\hat{\beta}_{cb/}\hat{\beta}_{ob}$	Cross-period effect $\hat{\beta}_{sp/}\hat{\beta}_{ob}$	Category-expansion effect $\hat{\beta}_{ce/}\hat{\beta}_{ob}$
Price Index	0.1234 (0.004)	11.02%	3.08%	2.51%
R^2 – train	0.460	0.590	0.814	0.815
Percentage significant (two-sided, p <0.05)	99%	99%	89%	81%

The last row in Table 3 points out that for all the brands in the soda category, 99 % of all-own brand effect and cross-brand effect, 89% of all cross-period effect and 81% of the category-expansion effect are statistically significant (two-tailed, $p < 0.05$). A previous study has indicated that the power in models of purchase quantity and interpurchase time gradually reduces (Neslin et al., 1985, fig. 2). Moreover, Bell et al. (1999) reported a low signal-to-noise ratio for the quantity portion of primary demand effects in the elasticity decomposition based on household data. In our study, the gradually decreased pattern in the frequency of significant primary demand effect is aligned with all previous studies mentioned above.

The results in Table 3 are the mean value of brand-level estimates. The first column stands for the mean own-brand effect across all brands corresponding to equation (7). Column 2 to column 4 indicate the fraction cross-brand effect, fraction cross-period effect and fraction category-expansion effect in percentage corresponding to equation (8)-(10). All mean coefficient estimates show positive signs as expected.

The coefficient estimates indicate the magnitude of influence of the price index exerted on the criterion variables. In this study, since we have standardized all dependent variables and independent variables before feeding them into the models, the interpretation of the coefficient estimates has changed accordingly. To illustrate, the value of the price index variable equals 1 when there is no price promotion and changes to 0.7 when there is a 30% discount. The coefficient of 0.1234 for the own-brand effect for soda implies that 1 standard deviation change in the price index variable results in 0.1234 standard deviation change in the own-brand sales.

In the absence of regular price data, R^2 value only reaches 0.460 for the regression model with own-brand sales as the dependent variable in the training dataset, which suggests an undesirable model fit. The underperformed model fit reflects that the approximated independent variables have biases against their actual values. As illustrated in section 2.3, using point estimate, multivariate linear regression models choose to average out these biases or uncertainties rather than model them out. The results are shown in Table 3, on average, the own-brand sales effect is decomposed into 11.02% of the cross-brand effect, 3.08% of the cross-period effect and 2.51% of the category-expansion effect. 83.39% stays unexplainable. Thus, the multiple regression models fail to fulfil as explanatory decomposition models.

5.2 Bayesian Standard Decomposition

We fit the Bayesian linear regression models as illustrated in section 3 using Bambi package in Python. In each model, we set up 2 chains to sample for each parameter, and each chain draws 1,500 samples from the posterior.

5.2.1 Convergence

First, we examine the 2 chains of each model to make sure the convergence of MCMC sampling, which indicates that the samples are indeed drawn from the posterior. The right-hand side of Fig 2 shows trace plots for $\beta_{ob,j}, \beta_{cb,j}, \beta_{cp,j}, \beta_{ce,j}$, where the x-axis and y-axis display the iteration number and the sampled parameter value, respectively.

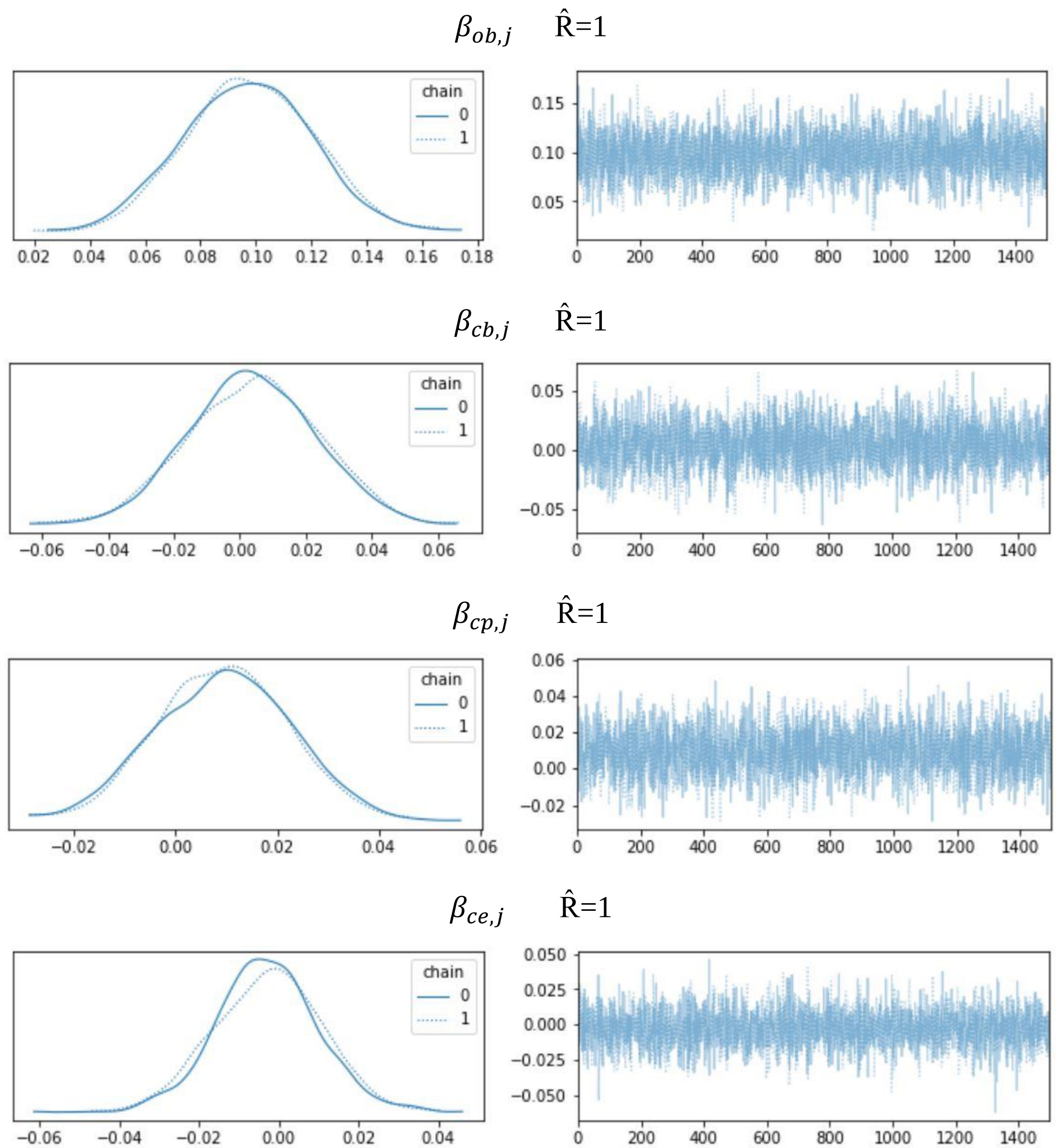


Fig 2. Traceplot and posterior density for PI

McElreath, (2016) states that trace plots should possess two attributes: stationary and good mixing. For each trace plot in Fig 2, different chains and iterations fluctuate within the identical parameter space. Besides, each trace plot bounces up and down around the posterior. Consequently, in each chain, sequentially related samples are uncorrelated, and the chain is drawing samples from all parts of the posterior. Thus, Fig 2 shows that all the trace plots are stationary and have good mixing.

To inspect the convergence of chains, we use the criteria provided by the Gelman-Rubin statistic ($\hat{R}$) (Gelman & Hill, 2007). Fig 2 indicates that all $\hat{R}$ equals to 1. For each model, the variability is the same between chains and within chains. Thus, we can conclude that all the chains have reached convergence.

5.2.2 Effects Interpretation

After establishing convergence, next, we use the simulations to make inferences from the model. The left-hand side of Fig 2 shows the posterior distribution for each chain. Since we set priors as weakly informative normal distribution, the posterior distribution of each chain also follows a normal distribution, which is confirmed through our result. Table 4 summarizes the mean and standard deviation of parameters of interest ($\beta_{ob,j}, \beta_{cb,j}, \beta_{cp,j}, \beta_{ce,j}$) and R^2 for each model.

Table 4. Posterior summary

	R^2-train	mean	std
$\beta_{ob,j}$	0.474	.097	.023
$\beta_{cb,j}$,	0.600	.004	.019
$\beta_{cp,j}$	0.818	.009	.013
$\beta_{ce,j}$	0.810	-.003	.013

Sample size N: 3000

The first column indicates that training through a much smaller dataset with 3000 samples, Bayesian linear regression models have shown a better fit to the training dataset except for the last model ($\beta_{ce,j}$) than the multivariate linear regression models. From the Bayesian perspective, the cross-brand effect, cross-period effect, and category-expansion effect are all ratios between two normal variables, which would follow a ratio distribution (Leal et al., 2014). To pinpoint the type, first, we need to check the correlation for each effect.

Table 5. Correlation between parameters

	$\beta_{ob,j}$	$\beta_{cb,j}$	$\beta_{cp,j}$	$\beta_{ce,j}$
$\beta_{ob,j}$	1	0.0085	-0.022	-0.040

Table 5 shows the correlation between each parameter. The highest absolute value 0.040 is still negligible when it comes to correlation. Thus, random variables $\beta_{cb,j}, \beta_{cp,j}, \beta_{ce,j}$, $\beta_{cb,j}, \beta_{cp,j}, \beta_{ce,j}$ are uncorrelated and the probability density functions of each ratio between $\beta_{cb,j}, \beta_{cp,j}, \beta_{ce,j}$ *and* $\beta_{ob,j}$ follow the formula of uncorrelated normal ratio distribution:

$$fz(w) = \frac{(\beta w\sigma^2 + 1)e^{-\frac{\sigma^2(w-\beta)^2}{2\delta_y{}^2(w^2\sigma^2+1)}} Erf(\frac{\sqrt{\frac{1}{2w^2\sigma^2 + 2}}(\beta w\sigma^2 + 1)}{\delta_y}}{\sqrt{2\pi}\sqrt{w^2 + \frac{1}{\sigma^2}}(\delta_y + \delta_y w^2\sigma^2)}$$

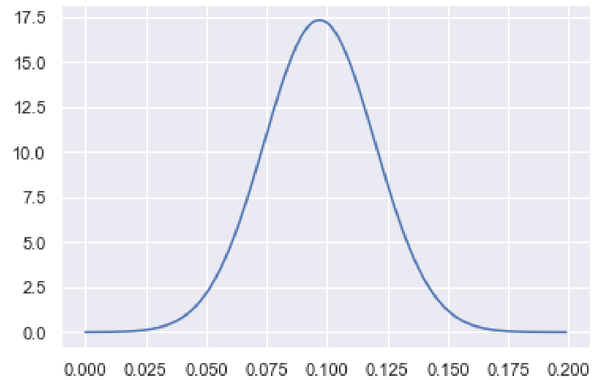
where, Erf is error function- a complex function of a complex variable defined as:

$$\operatorname{erf} z = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\,dt$$

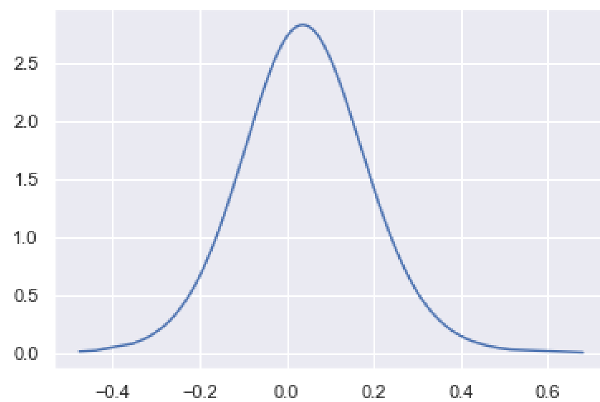
Fig 3 below is the probability density distribution for each effect, and Table 6 shows the corresponding statistics summary. The mean indicates the most probable estimate for each effect. In our case, 1 standard deviation change in the price index variable most probably results in 0.097 change in the own-brand sales in standard deviation. Own-brand sales effect can be most probably decomposed into 3.71% cross-brand effect, 9.16% cross-period effect, and -2.81% category-expansion effect. A previous study (Simonson et al., 1994) finds that when customers are uncertain about the values of products and about their preferences, such features and premiums provide reasons against buying the product and are seen as susceptible to criticism. In our case, if customers who perceive the promotion for a brand as offering little or no value or who become uncertain about their preferences after the discount outnumbers the customers who find the feature attractive, the promotion can reduce a product's overall choice probability and thus results in negative category-expansion effect.

Besides the mean estimates, Bayesian linear regression models present uncertainty regarding the parameters to avoid put excessive confidence on a noisy estimate through Bayesian credible intervals. A 95% credible interval indicates the upper and lower bound for the middle 9% of the total distribution (Kruschke, 2015). Compared to the frequentist viewpoint where each effect is a point estimate representing how brands respond to price promotions on average, the Bayesian perspective deals with the uncertainty by counting how much specific brands are expected to vary around that average. For example, when it comes to the cross-period effect for the soda category, most brands have 3.71% effect. However, some specific brands respond well off the average in specific stores at specific weeks. Given that 95% of the posterior distribution falls between -26.30% and 35.90%, 95% of the cross-period effect is between -26.30% and 35.90% across the brands, stores and weeks.

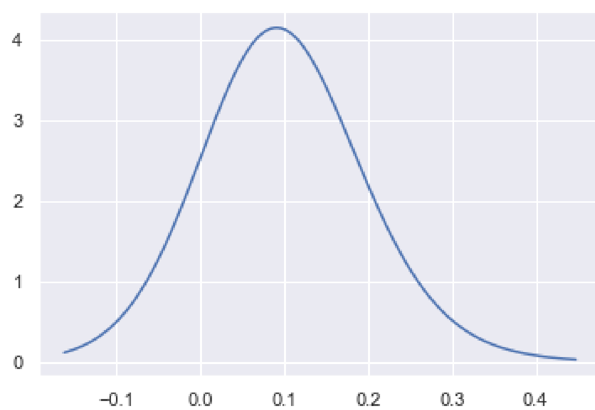
Own-brand effect



Cross-brand effect



Cross-period effect



Category-expansion effect

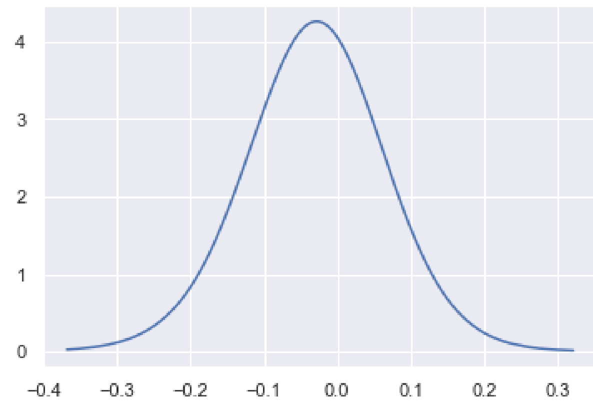


Fig 3. Ratio distribution for each effect

Table 6. Summary for each effect

	mean	95% CI
OBS	0.097	0.055; 0.139
CBS	3.71%	-26.30%; 35.90%
PPCS	9.16%	-11.22%; 33.46%
TCS	-2.81%	-22.50%; 20.54%

5.3 Predictive Modelling

The predictive performance of the regression models employed in this study is assessed with R^2 along with the regression evaluation metric RMSE. RMSE (Root Mean Squared Error) is the most frequently used metric in practice for the purpose of gauging accuracy for continuous variables and regression analysis. The square root standardized the scale of the errors to the same scale of targets. The equation of RMSE is presented in the formula below, where for a testing vector of length n, actual value O and predicted value P. Smaller values of RMSE indicate better performance.

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(P_i-O_i)^2}{n}}$$

The R^2 and RMSE of the performance of each model on the test dataset are shown in Table 7. The deep neural network regression models that we designed have the lowest RMSE and highest R2 compared to the multivariate linear regression and Bayesian linear regression models. Given an imperfect dataset without regular price data, our proposed deep neural network models can still make accurate predictions on own-brand sales, cross-brand sales, cross-period sales, and category-expansion effect for price promotion.

Table 7. Model Performance on test dataset

	Linear Regression		Bayesian Linear Regression		Deep Neural Network	
	R^2-test	RMSE-test	R^2-test	RMSE-test	R^2-test	RMSE-test
OBS	0.456	0.738	0.442	0.747	0.817	0.427
CBS	0.593	0.638	0.585	0.644	0.700	0.548
PPCS	0.812	0.434	0.809	0.437	0.817	0.428
TCS	0.813	0.432	0.810	0.436	0.825	0.418

VI. Conclusion

Working with imperfect data, we construct a system of models for unit-sales-based decomposition and prediction for store-level price promotion. In a real-world setting, gathering relevant regular price data of each brand in each store for each week is often unfeasible. Approximating through various approaches, in our case through calculating mode values, bias is still unavoidable. Under this uncertainty, the original multivariate linear regression models for standard decomposition proposed by Van Heerde et al. (2004) fail to explain the decomposition of own-brand sales effect for price promotion, with a large part of it (83.39%) staying unexplained.

In an attempt to conduct decomposition under data feature deficiency, we build a series of Bayesian linear regression models, which clarify uncertainty in parameter estimates. Training with only 3000 samples, the Bayesian linear regression models show an overall better fit than multivariate linear regression models, and each chain reaches convergence when conducting MCMC sampling. Subsequently, each effect is presented via a ratio distribution between two normal random variables (coefficient of price index for each model). In this way, we manage to decompose the own-brand sales effect from a Bayesian perspective. It deals with uncertainty not only representing mean value that stands for how brands respond to price promotions on average but also counting how much specific brands are expected to vary around that average.

Furthermore, we lift the restriction on parametric estimation and construct deep neural network for regression models. The metrics indicate that the proposed deep neural networks have the lowest RMSE and highest R^2 and thus the best prediction performance compared to multivariate linear regression and Bayesian linear regression models.

Our study has several managerial implications. First, when it is difficult for retailers to collect and store regular price data for all brands each week in all stores, managers can still not only understand how own-brand sales effect decompose for price promotion but also how specific brands will react to the price promotion compared to the average. From a retailer perspective, category-expansion effects are the most sought-after because cross-brand effects cannibalize other brands' sales, and cross-period effect hinders future sales. From a manufacturer's perspective, cross-period effects are also not beneficial except that they diminish sales opportunities for other brands. Consequently, a conservative estimate for the net effect for the manufacturer is the sum of cross-brand and category-expansion effects. Overall, from a category management perspective jointly pursued by manufacturers and retailers, category-expansion effects are potentially the most favorable among the three standard decomposition effects. Second, our designed deep neural network regression models can give managers accurate predictions on own-brand, cross-brand, cross-period and cross-category sales. Managers will be able to adjust the extent of price promotion and whether or not to start promotion campaigns based on the prediction results.

Our study has several limitations: first, our computation power is very limited for model training and hyperparameter tuning. With more capacity, MCMC sampling and deep neural network hyperparameter settings can be largely improved. Second, the proposed system of models is still highly restricted by the lacking of features such as regular price and display form. The performance of our models will further improve with the arrival of more complete data. Third, in this study, each constituent effect is still decomposable. For example, the category-expansion effect can be further split into cross-category effects. A further limitation is that we study in a stable environment where promotional magnitude stays the same for a given period. Future direction can thus focus on exploring decomposition methods in an evolving environment where promotional intensity constantly changes.

Appendix 1

a. OBS

	mu	sigma
Intercept	-0.0057	3.61
regular_price	0	2.5574
PI	0	2.5972
CPI	0	2.5461
crp	0	2.5604
week7	0	16.8832
week8	0	15.9038
week9	0	16.765
week10	0	14.9166
week11	0	15.0819
week12	0	15.8055
week13	0	16.2104
week14	0	16.1061
week15	0	16.2104
week16	0	14.9166
week17	0	14.3798
week18	0	15.4305
week19	0	14.4528
week20	0	14.6024
week21	0	15.0819
week22	0	14.099
week23	0	14.679
week24	0	14.0314
week25	0	15.2531
week26	0	14.527
week27	0	13.7711
week28	0	14.8361
week29	0	15.341
week30	0	15.7091

week31	0	14.7569
week32	0	14.099
week33	0	15.341
week34	0	15.1668
week35	0	16.2104
week36	0	16.3168
week37	0	15.0819
week38	0	17.5154
week39	0	17.3831
week40	0	17.004
week41	0	16.8832
week42	0	17.3831
PI1	0	2.5386
PI2	0	2.4804
PI3	0	2.5197
PI4	0	2.4212
PI5	0	2.5154
PI6	0	2.4529
PI11	0	2.6063
PI22	0	2.5982
PI33	0	2.5636
PI44	0	2.518
PI55	0	2.6195
PI66	0	2.5387
CPI1	0	2.5682
CPI2	0	2.5255
CPI3	0	2.6283
CPI4	0	2.5584
CPI5	0	2.55
CPI6	0	2.5876
CPI11	0	2.4952
CPI22	0	2.5705

CPI33	0	2.5235
CPI44	0	2.5154
CPI55	0	2.4918
CPI66	0	2.5901

Auxiliary parameters

sigma ~ HalfStudentT(nu: 4, sigma: 1.0123)

b. CBS

	mu	sigma
Intercept	-0.0046	3.5874
regular_price	0	2.5413
PI	0	2.5808
CPI	0	2.5301
crp	0	2.5444
week7	0	16.7771
week8	0	15.8038
week9	0	16.6597
week10	0	14.8229
week11	0	14.9872
week12	0	15.7062
week13	0	16.1086
week14	0	16.005
week15	0	16.1086
week16	0	14.8229
week17	0	14.2895
week18	0	15.3336
week19	0	14.362
week20	0	14.5106
week21	0	14.9872
week22	0	14.0104
week23	0	14.5868

week24	0	13.9433
week25	0	15.1573
week26	0	14.4357
week27	0	13.6846
week28	0	14.7429
week29	0	15.2446
week30	0	15.6104
week31	0	14.6642
week32	0	14.0104
week33	0	15.2446
week34	0	15.0715
week35	0	16.1086
week36	0	16.2143
week37	0	14.9872
week38	0	17.4053
week39	0	17.2739
week40	0	16.8972
week41	0	16.7771
week42	0	17.2739
PI1	0	2.5227
PI2	0	2.4648
PI3	0	2.5039
PI4	0	2.4059
PI5	0	2.4996
PI6	0	2.4375
PI11	0	2.5899
PI22	0	2.5819
PI33	0	2.5475
PI44	0	2.5022
PI55	0	2.603
PI66	0	2.5227
CPI1	0	2.5521

CPI2	0	2.5096
CPI3	0	2.6118
CPI4	0	2.5423
CPI5	0	2.534
CPI6	0	2.5713
CPI11	0	2.4796
CPI22	0	2.5544
CPI33	0	2.5076
CPI44	0	2.4996
CPI55	0	2.4761
CPI66	0	2.5739

Auxiliary parameters

sigma ~ HalfStudentT(nu: 4, sigma: 1.006)

c. PPCS

	mu	sigma
Intercept	0.0015	3.5657
regular_price	0	2.526
PI	0	2.5653
CPI	0	2.5149
crp	0	2.529
week7	0	16.6758
week8	0	15.7084
week9	0	16.5591
week10	0	14.7334
week11	0	14.8966
week12	0	15.6113
week13	0	16.0112
week14	0	15.9083
week15	0	16.0112
week16	0	14.7334

week17	0	14.2031
week18	0	15.2409
week19	0	14.2752
week20	0	14.423
week21	0	14.8966
week22	0	13.9258
week23	0	14.4986
week24	0	13.8591
week25	0	15.0657
week26	0	14.3485
week27	0	13.6019
week28	0	14.6538
week29	0	15.1525
week30	0	15.5161
week31	0	14.5756
week32	0	13.9258
week33	0	15.1525
week34	0	14.9804
week35	0	16.0112
week36	0	16.1163
week37	0	14.8966
week38	0	17.3002
week39	0	17.1695
week40	0	16.7951
week41	0	16.6758
week42	0	17.1695
PI1	0	2.5074
PI2	0	2.4499
PI3	0	2.4888
PI4	0	2.3914
PI5	0	2.4845
PI6	0	2.4227

PI11	0	2.5742
PI22	0	2.5663
PI33	0	2.5322
PI44	0	2.4871
PI55	0	2.5873
PI66	0	2.5075
CPI1	0	2.5367
CPI2	0	2.4944
CPI3	0	2.596
CPI4	0	2.5269
CPI5	0	2.5187
CPI6	0	2.5558
CPI11	0	2.4646
CPI22	0	2.5389
CPI33	0	2.4925
CPI44	0	2.4845
CPI55	0	2.4611
CPI66	0	2.5583

Auxiliary parameters

sigma ~ HalfStudentT(nu: 4, sigma: 0.9999)

d. TCS

	mu	sigma
Intercept	-0.0012	3.5639
regular_price	0	2.5247
PI	0	2.564
CPI	0	2.5136
crp	0	2.5278
week7	0	16.6676
week8	0	15.7007
week9	0	16.551

week10	0	14.7262
week11	0	14.8894
week12	0	15.6037
week13	0	16.0034
week14	0	15.9005
week15	0	16.0034
week16	0	14.7262
week17	0	14.1962
week18	0	15.2335
week19	0	14.2683
week20	0	14.4159
week21	0	14.8894
week22	0	13.919
week23	0	14.4916
week24	0	13.8523
week25	0	15.0584
week26	0	14.3415
week27	0	13.5953
week28	0	14.6467
week29	0	15.1451
week30	0	15.5085
week31	0	14.5685
week32	0	13.919
week33	0	15.1451
week34	0	14.9731
week35	0	16.0034
week36	0	16.1084
week37	0	14.8894
week38	0	17.2917
week39	0	17.1611
week40	0	16.7869
week41	0	16.6676

week42	0	17.1611
PI1	0	2.5062
PI2	0	2.4487
PI3	0	2.4875
PI4	0	2.3902
PI5	0	2.4833
PI6	0	2.4215
PI11	0	2.573
PI22	0	2.565
PI33	0	2.5309
PI44	0	2.4859
PI55	0	2.586
PI66	0	2.5062
CPI1	0	2.5355
CPI2	0	2.4932
CPI3	0	2.5947
CPI4	0	2.5257
CPI5	0	2.5174
CPI6	0	2.5546
CPI11	0	2.4634
CPI22	0	2.5377
CPI33	0	2.4913
CPI44	0	2.4832
CPI55	0	2.4599
CPI66	0	2.5571

Auxiliary parameters

sigma ~ HalfStudentT(nu: 4, sigma: 0.9994)

Reference

Ailawadi, K. L., & Neslin, S. A. (1998). The Effect of Promotion on Consumption: Buying More and Consuming It Faster. *Journal of Marketing Research*, *35*(3), 390. https://doi.org/10.2307/3152036

Assuncao, J. L., & Meyer, R. J. (1993). The Rational Effect of Price Promotions on Sales and Consumption. *Management Science*, *39*(5), 517–535. https://doi.org/10.1287/mnsc.39.5.517

Bell, D. R., Chiang, J., & Padmanabhan, V. (1999). The Decomposition of Promotional Response: An Empirical Generalization. *Marketing Science (Providence, R.I.)*, *18*(4), 504–526. https://doi.org/10.1287/mksc.18.4.504

Bucklin, R. E., & Lattin, J. M. (1992). A model of product category competition among grocery retailers. *Journal of Retailing*, *68*(3), 271.

Büyüköztürk, S. (2002). Sosyal bilimler için veri analizi el kitabı. *Ankara: Pegem Yayıncılık.*

Chandon, P., & Wansink, B. (2002). When Are Stockpiled Products Consumed Faster? A Convenience-Salience Framework of Postpurchase Consumption Incidence and Quantity. *Journal of Marketing Research*, *39*(3), 321–335. https://doi.org/10.1509/jmkr.39.3.321.19111

Chong, A. Y. L., Li, B., Ngai, E. W. T., Ch'ng, E., & Lee, F. (2016). Predicting online product sales via online reviews, sentiments, and promotion strategies. *International Journal of Operations & Production Management*, *36*(4), 358–383. https://doi.org/10.1108/IJOPM-03-2015-0151

Doyle, P., & Saunders, J. (1985). The Lead Effect of Marketing Decisions. *Journal of Marketing Research*, *22*(1), 54. https://doi.org/10.2307/3151550

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gonul, F., & Srinivasan, K. (1996). Estimating the Impact of Consumer Expectations of Coupons on Purchase Behavior: A Dynamic Structural Model: Abstract. *Marketing Science (Providence, R.I.)*, *15*(3), 262.

Hansen, L. K., & Salamon, P. (1990). Neural Network Ensembles. *Pattern Analysis and Machine Intelligence*, *IEEE Transactions On*, *12*, 993–1001. https://doi.org/10.1109/34.58871

Kalwani, M. U., Yim, C. K., Rinne, H. J., & Sugita, Y. (1990). A Price Expectations Model of Customer Brand Choice. *Journal of Marketing Research*, *27*(3), 251. https://doi.org/10.2307/3172584

Köksal, B. . (1985). İstatistik Analiz Metodları. *Çağlayan Kitabevi, İstanbul*.

Krishna, A. (1994). The Impact of Dealing Patterns on Purchase Behavior. *Marketing Science (Providence, R.I.)*, *13*(4), 351–373. https://doi.org/10.1287/mksc.13.4.351

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and stan*. Academic Press.

Leal, C., Oliveira, T. A., & Oliveira, A. (2014). RSM e convergência algorıtmica: revisao metodológica. *BOOK OF ABSTRACTS and PROGRAM*, 37.

Macé, S., & Neslin, S. A. (2004). The Determinants of Pre- and Postpromotion Dips in Sales of Frequently Purchased Goods. *Journal of Marketing Research*, *41*(3), 339–350. https://doi.org/10.1509/jmkr.41.3.339.35992

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (1st ed., Vol. 122). Boca Raton: CRC Press. https://doi.org/10.1201/9781315372495

Neslin, S. A. (2002). Sales Promotion. In *Relevant Knowledge Series*. Marketing Science Institute, Cambridge, MA.

Neslin, S. A., Henderson, C., & Quelch, J. (1985). Consumer Promotions and the Acceleration of Product Purchases. *Marketing Science (Providence, R.I.)*, *4*(2), 147–165. https://doi.org/10.1287/mksc.4.2.147

Nijs, V., Dekimpe, M., Steenkamp, J.-B., & Hanssens, D. (2001). The Category Demand Effects of Price Promotions. *Marketing Science*, *20*, 1–22.

Pauwels, K., Hanssens, D. M., & Siddarth, S. (2002). The Long-Term Effects of Price Promotions on Category Incidence, Brand Choice, and Purchase Quantity. *Journal of Marketing Research, 39*(4), 421–439. http://www.jstor.org/stable/1558555

Simonson, I., Carmon, Z., & Fogel, S. (1994). Experimental Evidence on the Negative Effect of Product Features and Sales Promotions on Brand Choice. *Marketing Science*, *13*, 23–40. https://doi.org/10.1287/mksc.13.1.23

Tabachnick, B. G., & Fidell, S. L. (1996). *Using multivariate statistics* (3 Edition).

Harper Collins College Publishers.

Ünver, Ö., & Gamgam, H. (1999). Uygulamalı İstatistik Yöntemleri. *Siyasal Kitabevi. Ankara*.

van Heerde, H. J., Gupta, S., & Wittink, D. R. (2003). Is 75% of the Sales Promotion Bump Due to Brand Switching? No, only 33% Is. *Journal of Marketing Research*, *40*(4), 481–491. https://doi.org/10.1509/jmkr.40.4.481.19386

van Heerde, H. J., Leeflang, P. S. H., & Wittink, D. R. (2000). The estimation of pre- and postpromotion dips with store-level scanner data. *Journal of Marketing Research*, *37*(3), 383–395. https://doi.org/10.1509/jmkr.37.3.383.18782

van Heerde, H. J., Leeflang, P. S. H., & Wittink, D. R. (2004). Decomposing the sales promotion bump with store data. *Marketing Science*, *23*(3), 317–334. https://doi.org/10.1287/mksc.1040.0061

Walters, R. G. (1991). Assessing the Impact of Retail Price Promotions on Product Substitution, Complementary Purchase, and Interstore Sales Displacement. *Journal of Marketing*, *55*(2), 17. https://doi.org/10.2307/1252234

Wittink, D. R., Addona, M. J., Hawkes, W. J., & Porter, J. C. (1988). SCAN* PRO: The estimation, validation and use of promotional effects based on scanner data. *Internal Paper, Cornell University*.

Zhang, G. P., Patuwo, B. E., & Hu, M. Y. (2001). A simulation study of artificial neural networks for nonlinear time-series forecasting. *Computers & Operations Research*, *28*(4), 381–396. https://doi.org/https://doi.org/10.1016/S0305-0548(99)00123-9