OBJECTIVE PLATFORM

ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

Master Thesis Econometrics & Management Science

A Hierarchical Bayes Approach to Modelling Media Advertising Elasticities

Sanne Bakker (450842)

University supervisor
Dennis Fok

Company supervisor
Thomas van Noort

Second assessor
Martina Zaharieva

April 25, 2021

The views stated in this thesis are those of the author and not necessarily those of Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

Advertising elasticities, which are used in marketing budget allocation, tend to fluctuate over time, making estimation based on only most recent data preferable. The downside of this, is that purely data-driven estimation techniques will not perform well. In this paper, we therefore investigate how online advertising elasticities can be reliably estimated in a log-log model using a hierarchical Bayes model. In particular, for each type of investment within a marketing channel we want to estimate an elasticity. These investments can be classified into a specific hierarchy, forming a tree structure. We examine how this tree structure can be incorporated into the hierarchical prior of the elasticities. In this way, the sparse data problem is tackled, as different marketing investments within the same channel share their information. Instead of total sales, we employ attribution as dependent variable in the elasticity model, which is an important contribution to literature. By means of a simulation study and predictions on advertising data from a telecommunications company, the performance of the model is assessed. We find that in case different kind of investments within the same channel have comparable elasticities, the Bayesian estimation outperforms the frequentist estimation. In addition, when an aggregated interaction term is added to the model, the performance of the frequentist model decreases while the performance of the Bayesian model improves.

Keywords: advertising elasticities, Bayesian estimation, hierarchical prior framework, log-log specification, attribution, simulation, aggregated interactions

Contents

1 Introduction

Marketing Mix Modelling (MMM) is concerned with estimating the impact of various marketing investments on sales. This concerns both offline marketing channels like radio or television and online marketing channels like display or paid social marketing. While the effect of the former can only be measured on an aggregated level, the effect of the latter can be measured on customer based level. Over the years, the available marketing data has become more detailed. People are influenced by multiple marketing channels during their "customer journey" before they eventually make a purchase, which is called converting. Especially online marketing has gained ground here, as a consequence of personalized advertising and cookie tracking, which tracks a customer's activities online. Accordingly, the commonly used MMM techniques are not sufficient anymore, since the business topics of interest have become more complex. Still, for a company the goal remains to allocate their marketing budget in an optimal way. Usually, this means in such a way that it generates most extra sales. Marketing consultants often face the challenge of giving advice on this subject. Ideally, this advice is as detailed as possible. For example, whether or not to invest more in a certain campaign.

In this paper, we consider this challenge as a two step procedure, which is one of our main contributions to existing literature. Step one is the attribution step, where it is determined which part of the realized sale should be assigned to each marketing channel because of its contribution to the purchase. This attribution is expressed in revenue or conversions. In step two, which is the step we concentrate on, we create a model where attribution is explained by the marketing investments. The resulting advertising coefficient is the elasticity, meaning the percentage with which revenue or the number of conversions changes when the investment is changed with a one percent. These elasticities can next be used to optimize a marketing budget.

Marketing channels often can be categorized into different classes, forming a treelike structure. For example, the display channel can be split up into prospecting and retargeting marketing. Retargeting can be split up into different types of campaigns, which are active for a certain period and focus on a specific theme. In this way the marketing budget can be divided over different "levels". Because of this tree structure, it applies that the deeper the level, the fewer data points there are, as not every kind of marketing investment is done every week. In addition, we do not want to use data too far in the past as elasticities tend to fluctuate a lot over time. As a result, it becomes hard to accurately estimate the corresponding elasticity. Moreover, if the estimates are inaccurate, the advice will not be optimal. Therefore, preference will be given to an estimation method which also performs well using few data points. Consequently, it is interesting to estimate these elasticities by means of a hierarchical Bayes model and incorporate the hierarchical structure of the channels into the priors. In this way, different kind of marketing investments share their information, making the estimated elasticities more accurate. That is what this paper will focus on. To be precise, the following research question is investigated: *How can online advertising elasticities for specific campaigns be estimated using a hierarchical Bayes model?*, with the sub-question *How can the tree structure of the marketing channels be incorporated into the prior of the elasticity?*.

This thesis is done in cooperation with Objective Platform (OP) as part of a graduate internship. OP is a marketing consultancy company situated in Amsterdam and focuses on media budget optimization. To achieve this, they track multi-channel media activities. All findings are presented in the Objective Platform, a flexible online tool especially designed for visualization and optimization. In this way, they can assist their clients in making data-driven decisions regarding their media spend.

Currently, OP estimates the elasticities using a standard frequentist approach on a logarithmic linear regression model. This model does not perform well in case there are few data points, as this approach is purely data-driven. Therefore, another sub-question is: *How does the Bayesian elasticity model perform in comparison to the currently used frequentist model?* The models will be evaluated by means of a simulation study as well as a forecast on revenue of real data. However, OP's main interest is a valuable advice to the client. To be more specific, they use the elasticities as input for an optimizer which allocates a given marketing budget, aiming for the highest return on investment possible. The findings in this study may contribute to OP revising parts of their currently used elasticity framework in order to improve their advice to the client.

The remainder of this research is structured as follows. Section 2 starts off with a review about the relevant literature on the subject, followed by a description of the data provided by OP in Section 3. Subsequently, the methods used to estimate the advertising elasticities are presented in Section 4. After that, Section 5 discusses the results and the conclusion is stated in Section 6.

2 Literature

This section gives a detailed outline of the research done so far concerning marketing elasticities. To start, Section 2.1 gives an overview of modelling advertising effectiveness in general and compares this to the model built in this thesis. Secondly, Section 2.2 addresses common functional forms, as well as interaction effects between channels in Section 2.2.1. Furthermore, Section 2.3 discusses (hierarchical) Bayesian estimation in marketing research.

2.1 Modelling advertising effectiveness

When it comes to measuring advertising effectiveness, most researchers employ a model which explains the total sales using advertising investments (Bass [1969]). The more advanced papers conduct this simultaneously for different marketing channels or themes in a multiple linear regression (Bass et al. [2007]). Furthermore, several studies state which requirements the model should meet in order to be reliable or complete (Quandt [1964], Doyle and Saunders [1990]). Henningsen et al. [2011] even developed a list of significant determinants of advertising elasticities by comparing over more than sixty studies on the subject. It is generally agreed that one has to take into account factors like price, carry over effects and dynamic effects in order to distinguish the advertising effects from other variables that affect sales. That is also one of the major difficulties that Bass [1969] points out in measuring the influence of advertising.

However, in this research we assume that, given the investment on advertisement, the attributed revenue or number of conversions is known for each marketing channel. OP combines multiple ensemble techniques and heuristics to determine these attribution values (hereafter called the attribution model). This includes a so-called “multi-touch” attribution model which is based on Dalessandro et al. [2012]. Usually a customer has multiple moments of contact with marketing advertisement before a sale is made. Together these touch points form the customer journey. The attribution model ensures that when a customer for example clicks on a display banner, but only converts several days later, the display banner still receives part of the attribution. In addition, OP controls for the other factors stated above. Moreover, they include correcting variables like the day of the week, the weather, holidays and even events like periods during which the website was down.

Using the estimated attribution as a dependent variable, we can directly build a model for the advertising elasticities, using a separate equation for each kind of advertisement. That is, how much the attribution changes when the budget is changed with a certain percentage. As our dependent variable is the already isolated attribution per channel instead of total sales, our initial specification is relatively simple. This is an advantage, since estimation becomes more difficult when the model contains multiple variables (Jagpal et al. [1982]). Instead, we can focus on other aspects which potentially need to be included. For example, interactions effects between channels, see section 2.2.1.

This two-step procedure, in which step one is the attribution model and step two is the elasticity model, is one of our main contributions to existing literature. The advantage here is that we get concrete attribution values for each channel. On the other hand, we need a separate model for the estimation of both the attributions and the elasticities. This means that when there is a misspecification in the attribution model, the elasticity model will suffer from this.

2.2 Functional forms

Multiple model specifications are employed and compared extensively in literature (Henningsen et al. [2011]), of which logarithmic specifications are most widely used. In general, media advertising investments are known to have diminishing returns, resulting in a concave curve. Next to that, the existence of a certain threshold beneath which there would be little to none response has been investigated (Vakratsas et al. [2004]). This results in an S-shaped curve. There are studies which show evidence in favour of this form as well as against it. Nonetheless, Jin et al. [2017] showed that this more complex specification does not lead to more accurate estimates in comparison to the more flexible logarithmic specifications.

Formerly, OP estimated the elasticities using a linear-log specification, which was also implemented by Doyle and Saunders [1990]. An advantage of this model is that after estimating the parameters, a simple allocation rule can be applied to divide the marketing budget (Carroll et al. [1979]). The drawback is that the curve does not always cross the origin, which is not realistic, as we assume the attribution is exactly zero in case the marketing channel is inactive. Therefore, they made a transition to the log-log model (Jagpal et al. [1982], Bass [1969]), which is multiplicative instead of additive.

2.2.1 Interactions between channels

It is important to include interactions between marketing channels in the model specification, as the client can exploit this information when deciding which channels to set active together. There are two types of those interactions, also called cross-effects. Firstly, we may encounter positive cross-effects, for example when a lot of display improves the effect of paid social campaigns. Conversely, negative cross-effects occur in case marketing channels decrease in efficacy when used together.

Several studies have investigated the importance of interaction effects. For example, Henningsen et al. [2011] concludes that the hypothesis of positive interaction effects can neither be supported nor rejected. However, he only investigates the hypothesis that interactions have a positive effect, hereby ignoring the negative case. Other papers actually emphasise the need for modelling interactions (Jagpal et al. [1982]).

In this research, we follow the method of Bass et al. [2007], who implemented interactions in an aggregated form. That is, the effect of one channel versus all other channels which were active at that moment. The advantage of this is that the model specification stays relatively simple and no interaction effects are neglected. Besides, we do not need to do a variable selection to get to the actual interactions. In addition, implementing all interactions separately would mean a huge increase in parameters to be estimated, which can become a challenge with respect to estimation when limited data is available. On the other hand, the downside of an aggregated interaction term is that it gives less actionable insights. Furthermore, Bass et al. [2007] omits the interaction of channels with other variables like price and promotions. While we acknowledge the importance of such interactions, we also do not include these in the model as it is beyond the scope of this paper. Krishnamurthi and Raj [1985] for example showed that more advertisement leads to a smaller price sensitivity.

2.3 Bayesian estimation in marketing

Many papers use the classical frequentist approach to estimate the advertising elasticities, either by minimizing the Residual Sum of Squares (Jagpal et al. [1982]) or by maximizing the likelihood function. However, when the available data is limited, the estimation is highly affected by outlying points, resulting in inaccurate estimates. It will produce a curve which best fits the data, as the technique purely relies on information contained in the data. This is also a challenge OP is facing at the moment. As long as the elasticities are not estimated accurately enough, the optimizer will not give reliable results. This is where there is room for improvement.

Instead of a frequentist approach, one can use a Bayesian approach. The advantage here is that the parameters are viewed as random variables. Information not contained in the data, can be included by means of a prior distribution. Rossi and Allenby [2003] give a detailed review about the usefulness of Bayesian methods when dealing with marketing problems. Furthermore, Bayesian modelling is also known to be able to handle relatively complex models. Bass et al. [2007] for example implemented a Bayesian dynamic linear model to capture the dynamic effects of different categories of advertising and applied Gibbs sampling (Gelfand and Smith [1990]) to estimate its parameters. More recently, Jin et al. [2017] employed a Bayesian framework to estimate the parameters of a media mix model which accounts for carryover and shape effects. In order to get more informative priors, they recommend to extend the framework to a hierarchical Bayes (HB) model, in a manner similar to Wang et al. [2017], where they pooled the data of similar brands within one category. By comparing this to a Bayesian model without hierarchy they showed that the HB model contains less uncertainty in the estimated response curve and produces more accurately estimated parameters. For this reason, we will follow this idea, but instead of pooling similar brands, we will pool across the hierarchy of the marketing channels and incorporate this structure into the priors.

2.4 Centralization of the data

When estimating a linear model, standardization of the independent variables does not change the overall quality of the model. However, when interaction terms are included, it is important to standardize the explanatory variables for several reasons. First of all, certain standardization techniques contribute to a more meaningful interpretation of the separate effects. In this paper we apply centralization, which means subtracting the corresponding mean of each explanatory variable (Shieh [2011]). This is explained in more detail in Section 4.1. Secondly, the standardization is necessary in combination with a hierarchical Bayes model. For this model we need the parameters of different channels to be on the same scale. Otherwise, it is difficult to assign them a hierarchical prior through which they share information. Thirdly, mean centering reduces the "micro" view of multicollinearity (Iacobucci et al. [2016]) meaning a substantial correlation between explanatory variables. Without standardization, the estimates will be imprecise.

Note that centralization is not the only standardization method available. Other commonly used methods are min-max or z-score standardization (Saranya and Manikandan [2013]). The weakness of these methods, however, is that they are dependent on the variation in the data (Jain et al. [2005]). When there is little variation in the data, the multicollinearity caused by the interaction terms can still be contained in the standardized data. In addition, these methods lead to a less straight-forward interpretation of the effects.

3 Data

The data used in this research originates from a telecommunication provider (hereafter called the client). We have information about three main online marketing channels, which are paid social, display and paid search. The structure of these channels is explained in Section 3.1. In addition, Section 3.2 gives more prior knowledge about each channel, which will be incorporated in the HB model. Furthermore, for each channel we have cost and attribution data aggregated on a weekly level. We specifically focus on attribution in terms of revenue, although attribution in terms of conversions can be modelled in an equivalent way. Section 3.3 goes into the summary statistics of the data. Note that a channel does not have to be active each week. If a channel is not active in a certain week, the attribution will be zero in that week. In this way, the data set can be viewed as an unbalanced panel, as these weeks are not incorporated in the data set.

3.1 Structure of the marketing channels

The marketing channels which we focus on can be classified according to a specific hierarchy, which we express in terms of levels. For example, Level 1 is display, Level 2 is branded or performance and Level 3 is awareness or consideration when Level 2 is branded, while Level 3 is prospecting or retargeting when Level 2 is performance. There is not just one way in which the data can be structured. Ideally, we want a tree for which we can assume the elasticities of the children are comparable. In that case, the model benefits from the information contained in the hierarchy. In this thesis we adopt the structure followed by the client, as they use this to make investment decisions. This means that the channel is at the top of the tree. It is not unreasonable to assume that different investments in the same channel have elasticities which are comparable, but not necessarily equal to each other. Figure 1 depicts the hierarchies for each channel. There are fewer data points on deeper levels, as not every kind of marketing is done every week. We will compare the performance of the models using the data containing the hierarchy of all three levels.

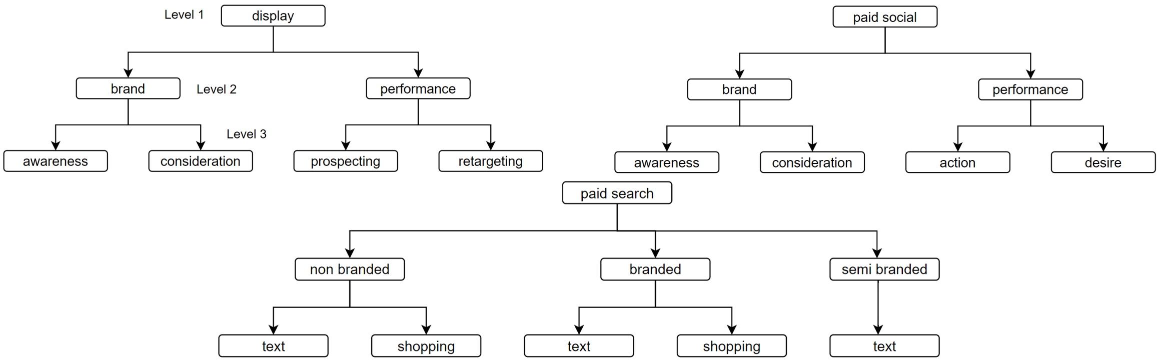

Figure 1: The hierarchy of display, paid social and paid search.

One should not forget that part of the customers directly navigates to the online store and makes a purchase without intervention of advertising. In the attribution model these so called "unpaid channels" are taken into account, as they are part of the customer journey. Accordingly, part of the generated sales is attributed to these channels. Nonetheless, when studying the effectiveness of the paid marketing channels, we should also investigate the potentially negative effect of advertising investments on this "direct" attribution. We call this the cannibalisation effect. For example, it could be that part of the sales attributed to the direct channel is transferred to display when a larger share of the media budget is allocated to this channel. Therefore, we perform a control regression with the weekly direct attribution as dependent variable and the weekly attribution per marketing channel as explanatory variables. Section B of the Appendix shows that there are no significant negative effects on direct attribution. Note, however, that this is just a small check and that the actual investigation of these side effects is out of the scope of this research. In case there would be significant effects, OP should incorporate this when giving the client advice about investments in this channel, or the estimated elasticities should be corrected directly. For a potential integration of these side effects into the model, see Section 6.2.

3.2 Prior knowledge of the elasticities

In general, we expect advertising elasticities to have a value between zero and one, as marketing investments are meant to have a positive effect on sales and we expect diminishing returns. For the HB model we can incorporate more specific information about the marketing channels into the prior distributions. First of all, paid search is known to be focused on customers which are already engaged with the brand or the product. In other words, they are in a further stage of their marketing journey and are more likely to convert. In addition, the return of investment on this channel is relatively high. On the other hand, in general paid social and display are more aimed at bringing in new customers. Especially display is a more indirect channel, meaning that it is intended to raise interest among potential customers and create brand awareness. From this perspective, the probability of converting is lower for this kind of channel. This information together with knowledge about estimated elasticities in 2020 of a company similar to our client, results in the expected average elasticities per channel which are shown in Table 1. These values will be used as the hyper prior values for the elasticities in the HB model explained in Section 4.2.2.

Table 1: Expected average elasticities per marketing channel.

channel	expected avg. elas.
paid search	0.7
paid social	0.5
display	0.3

3.3 Summary statistics

Next, we describe the characteristics of the data more specifically. The left plots in Figure 2 display revenue and campaign costs over time for each channel. The graphs show that there is a lot of variation in costs and revenue over the year. This can be explained by the fact that campaigns are often active for a certain period of time. This is also the reason why we do not treat or exclude potential outliers (see the box-plots in Figure 13 of the Appendix), as this is part of the "on and off" structure of marketing investments. Furthermore, on he right side of Figure 2 costs are plotted against revenue. These plots show that different deepest levels of the hierarchy cluster in the same area, in the sense that their cost-revenue combinations are close to each other. The data points of the display and paid social children nicely cluster in the same region. For example, for awareness and consideration from brand - display the hypothesis of comparable elasticities is quite plausible. However, for paid search the cost-revenue combinations are really diverse for different children. The performance of the HB model might suffer from this.

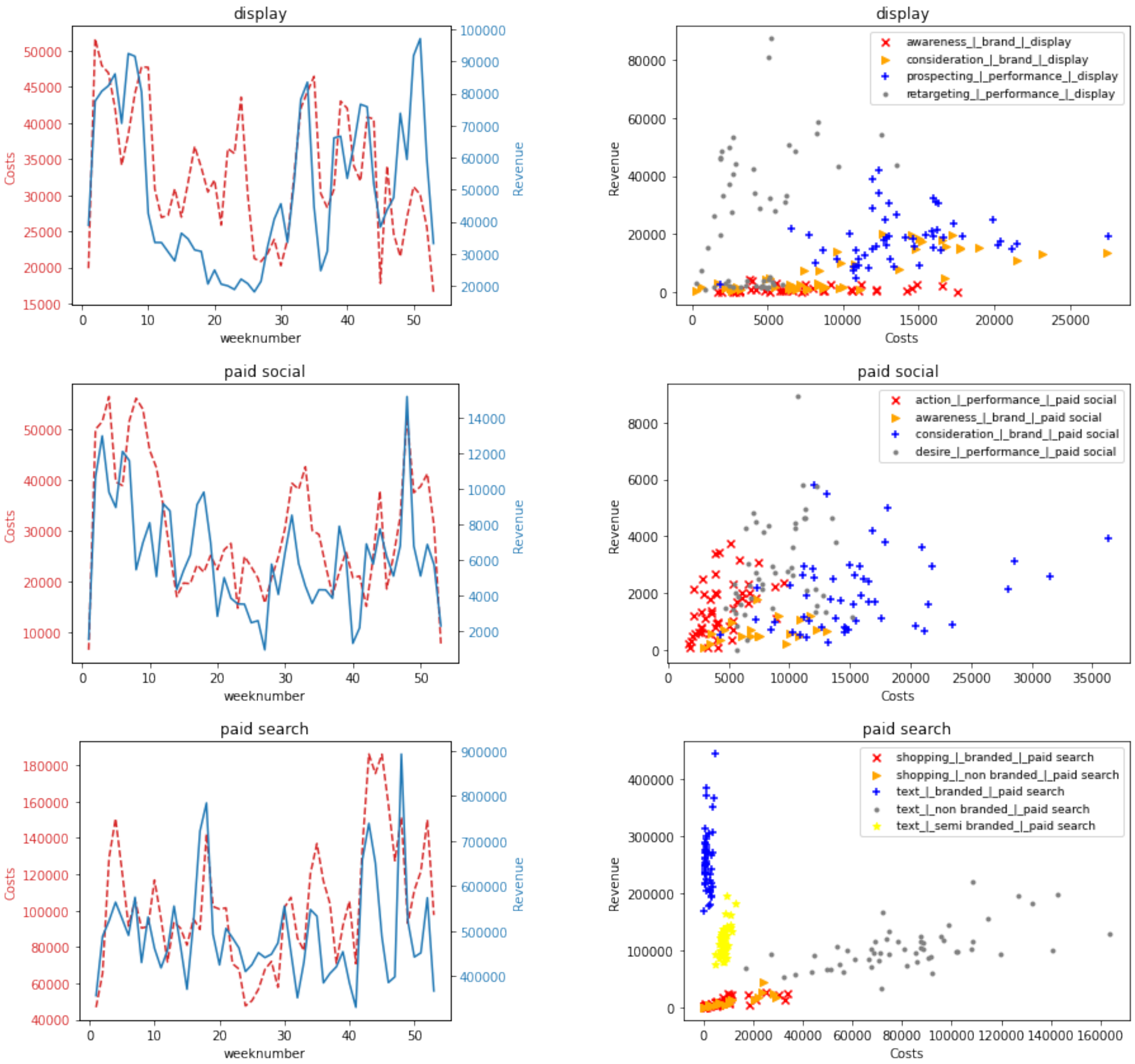

Figure 2: Graph over time of revenue and cost (dashed) data (left) and scatter plot of cost against revenue (right) for each marketing channel.

Furthermore, Tables 12 and 13 in the Appendix contain summary statistics of each deepest level for the cost and revenue data, respectively. In addition, histograms of the log-transformed data for each marketing channel are given in Figure 3. We examine normality of the log-transformed revenue, as this is the dependent variable in our linear model for which the disturbances need to be normally distributed. Table 2 shows that for most deepest levels the kurtosis ranges from 2 to 4 and the skewness between -1 and 1, with one exception for paid social. The kurtosis should be close to three and the skewness close to zero for the data to be normally distributed. We conclude that most deepest levels are close to but not equal to a normal distribution. Note that we do not perform a Jarque-Bera test as there are not enough data points for this asymptotic measure.

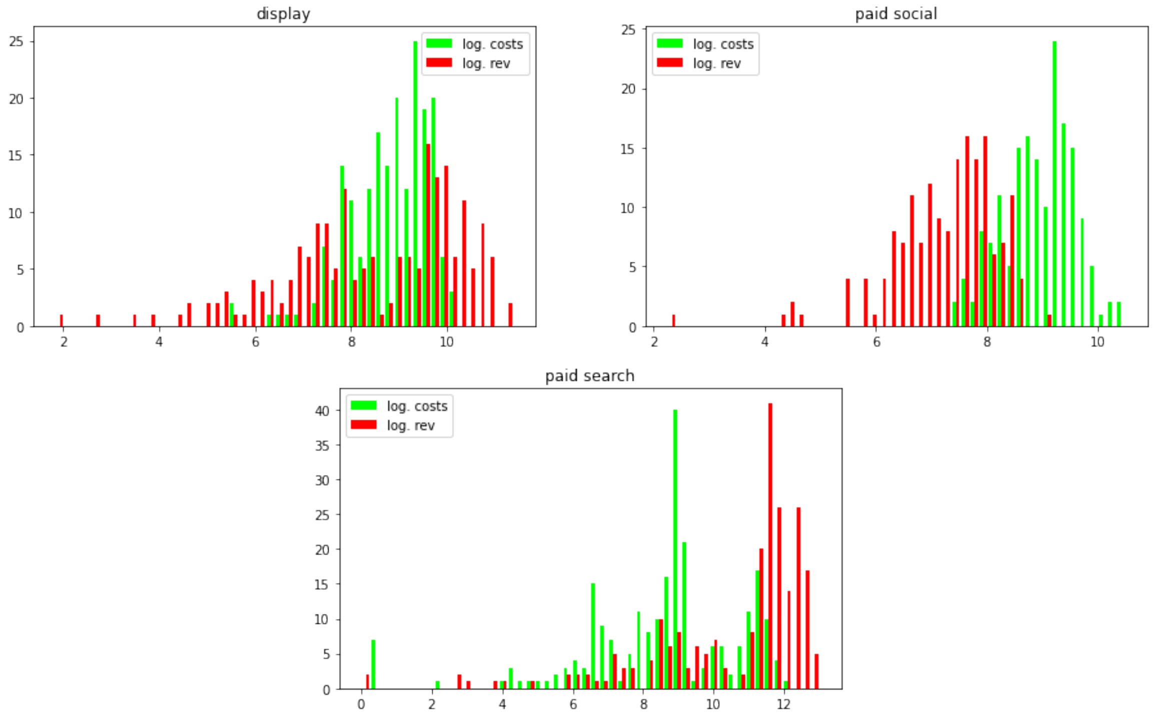

Figure 3: Histograms of log-transformed cost and revenue data for each marketing channel.

Table 2: Kurtosis and skewness values of log-transformed revenue data for each deepest level.

	kurtosis	skewness
shopping - branded - paid search	2.241	-0.496
text - branded - paid search	3.349	0.412
text - non branded - paid search	3.942	-0.182
text - semi branded - paid search	2.838	0.192
shopping - non branded - paid search	2.582	-0.957
awareness - brand - display	3.174	-0.609
consideration - brand - display	1.602	0.134
prospecting - performance - display	4.617	-0.894
retargeting - performance - display	1.338	-0.227
action - performance - paid social	4.217	-1.200
desire - performance - paid social	19.599	-3.494
awareness - brand - paid social	4.068	-0.856
consideration - brand - paid social	2.375	-0.303

4 Methods

The following section describes the econometric techniques which are part of this research. First of all, Section 4.1 specifies the elasticity model. Secondly, Section 4.2 describes the estimation both with the frequentist approach as well as the Bayesian approach. Followed by that there are a number of model choices and restrictions specifically for OP in Section 4.3. Finally, the models are evaluated by means of techniques described in Section 4.4.

4.1 Model specification

To estimate the advertising elasticity curve for each deepest level $l = 1 \dots L$ of each marketing channel $j = 1 \dots J$, we specify the dependent variable y_{jlt} as total revenue or conversions attributed to channel j at deepest level l at time t and the explanatory variable x_{jlt} as the budget spent for channel j at deepest level l at time t. Then, for each deepest level l of channel j we get the following specification:

$$\log(y_{jlt} + 1) = \alpha_{jl} + \beta_{jl} \log(x_{jlt} + 1) + \varepsilon_{jlt}, \tag{1}$$

where the addition of 1 ensures the log-transformed variables do not become negative. Even though we only consider this model for observations with positive x_{jlt}, we still need this addition of 1 as x_{jlt} can have a value smaller than 1. Furthermore, we want to incorporate interactions into the model as an aggregated form, as described in Section 2.2.1. Taking λ_{jl} as the interaction effect of deepest level l of channel j with all other deepest levels, (1) is extended in the following way:

$$\log(y_{jlt} + 1) = \alpha_{jl} + \beta_{jl} \log(x_{jlt} + 1) + \lambda_{jl} \sum_{i=1}^{J} \sum_{\substack{m=1 \\ (im) \neq (jl)}}^{L} \log(x_{jlt} + 1) \log(x_{imt} + 1) + \varepsilon_{jlt}. \tag{2}$$

Our main interest is the elasticity parameter β_{jl}, which can be interpreted as the percentage of change in y_{jlt} when x_{jlt} is changed by one percent. Note that this is an approximation because of the addition of 1 in the logarithmic transformation of x_{jlt}. When the explanatory variable is relatively large, the exact value of the elasticity will be approached. As we expect the attribution to increase when more is spent on the channel, we can assume $\beta_{jl} > 0$. To make sure β_{jl} does not become negative we apply the following parameter transformation:

$$\beta_{jl} = \frac{1}{c} \log(1 + e^{c\beta^*_{jl}}), \tag{3}$$

where c is a constant. We next treat β^*_{jl} as the parameter to be estimated. This makes sure that in case β^*_{jl} is larger than zero then β_{jl} is close to β^*_{jl}, while in case β^*_{jl} is negative then β_{jl} will be close to but just above zero. In this way, β_{jl} has a direct interpretation with respect to the degree to which β^*_{jl} is negative. Figure 4 displays the graph of (3) for different values of the constant c. This value should be chosen in such a way that the function is not too close to $\beta_{jl} = max(\beta^*_{jl}, 0)$, because this can cause issues in the Bayesian sampler. Besides, in case of the latter function, there is no difference between a β^*_{jl} which is just below zero or a β^*_{jl} which is very negative. We set c equal to 10.

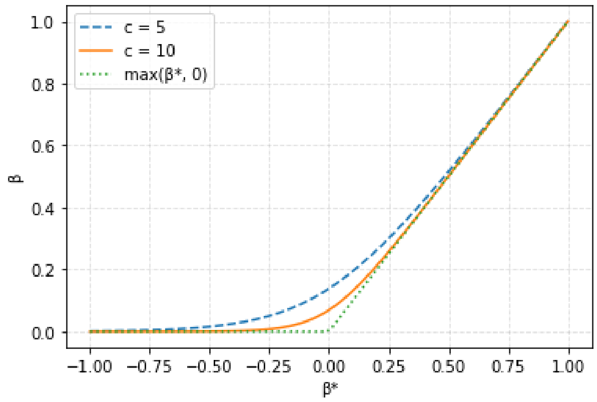

Figure 4: Transformation of the elasticity parameter as described in (3) for different values of c, leaving the subscripts out for simplicity.

As described in Section 2.4, we need to standardize the explanatory variables for several reasons. The log-transformed costs of each deepest level are standardized by subtracting its mean d_{jl}. Using the centralized values we determine the interactions. This changes (2) to

$$\log(y_{jlt}) = \tilde{\alpha}_{jl} + \tilde{\beta}_{jl}(\log(x_{jlt}) - d_{jl}) + \tilde{\lambda}_{jl} \sum_{i=1}^{J} \sum_{\substack{m=1 \\ (im)\neq(jl)}}^{L} (\log(x_{jlt}) - d_{jl})(\log(x_{imt}) - d_{im}) + \varepsilon_{jlt}, \quad (4)$$

where $\tilde{\alpha}_{jl}$, $\tilde{\beta}_{jl}$ and $\tilde{\lambda}_{jl}$ are the standardized parameters. Equation (1) is adapted in a similar way. With respect to interpretation, without centralization the impact of a moderate investment is represented by all parameters together, which is not convenient. With centralization, the parameters become more independent and the interpretation of the coefficients becomes more straight-forward. The variables now represent the difference from their mean; the interaction term is equal to zero when advertising investments of the other channels are at an average level. In this way, $\tilde{\beta}_{jl}$ captures the "main" effect of l relative to the average value of the other advertising investments with which it has an interaction. Simultaneously, $\tilde{\lambda}_j$ captures the effects when advertising investments are above or beneath their average. Note that subtracting a constant value other than the mean is also an option. However, as we want the difference to be zero in a common case, the mean is the most logical value.

4.1.1 Prior framework

In the context of estimating advertising elasticities, there is often less data available on deeper levels, since not every kind of marketing is invested in every week. As a result, the parameter estimates can contain a lot of uncertainty. Even when there is data available each week for all deepest levels, we still do not want to use data too far in the past as elasticities tend to fluctuate a lot over time. In order to get more accurate estimates we want to let similar marketing investments share information with each other. This can be achieved by incorporating the in Section 3.1 described tree structure into the model in the form of priors. As we want to tackle the problem of sparse data points on the deepest levels, we set the priors with a top-down structure. In this way, information of the data points belonging to a parent, for example paid social, are used as well to estimate the elasticities of the children, branded and performance.

Consider the regression model as stated in (1), where we assume the data points to be normally distributed. First of all, for the prior of the standard deviation σ_{jl} of the disturbance term we take a exponential distribution as is recommended by Simpson et al. [2017]. In a simulation study they show that this prior is rather insensitive to the choice of hyper-parameter λ, so we set this value equal to five, generating a weakly-informative prior. Secondly, we expect the advertising elasticities to be positive, therefore the first thought would be to assume a positive normal distribution for these variables. However, the transformation (3) guarantees the estimates do not become negative and as a result it suffices to take a plain normal distribution as a prior. Furthermore, in our model we want to put the information we have about a higher level into the prior of a deeper level. To ensure the draws of a deeper level are around the mean of the distribution of a level above we construct the following hierarchical prior framework:

$$\begin{aligned} \beta_j &\sim Normal(\mu_{\beta_j}, \nu_j^2), \\ \beta_{jk} \mid \beta_j &\sim Normal(\beta_j, \nu_k^2), \\ \beta_{jkl} \mid \beta_{jk} &\sim Normal(\beta_{jk}, \nu_l^2), \end{aligned} \tag{5}$$

where ν_j^2, ν_k^2 and ν_l^2 are the variances, set to a certain value. The hyper prior μ_{β_j} is set to the expected average of the elasticity of channel j as stated in Table 1. Here, for example when j is paid social and k is branded then l is awareness or consideration. Now it can be easily seen that there is another reason why a positive normal distribution is not convenient as a prior. With a truncation at zero, the elasticity of a deeper level will be overestimated as the parameter μ is not equal to the mean of the distribution. So on each level there would be need for a correction. Hence, we would not be able to give a meaningful interpretation to the parameters on higher levels anymore. In addition, sampling would become a lot more challenging.

The variance ν_l^2 expresses the degree of variation within the distribution of β_{jkl}. At the same time, it also represents how much we expect elasticities of channel j on this level to be different from each other since they have the same prior distribution. This hyper prior is not included in the hierarchy as a parameter to be estimated. This is because at each level there are only two or three different children so there is not enough information to estimate the variance between these children. The same goes for parameters ν_k^2 and ν_j^2. We set the values of these three hyper priors equal to 0.01, as we expect that the elasticities within the same channel are not very different from each other. To be precise, we expect the confidence interval of the distribution consisting of the differences between two elasticities on a certain level to range from -0.2 to 0.2.

Furthermore, for the intercept α_{jkl} we choose a flat prior, because we do not have a lot of prior information about this parameter. Unlike the elasticity we do not expect a certain value to be more common than other values. Besides, this parameter is not included in the hierarchy. The value of the intercept is dependent on the location of the data points. This does not have to be similar for different children of the same parent, for instance when the total investment in a channel is not equally divided over the children.

Lastly, for the interaction term λ_{jl} we assume a normal distribution with its mean equal to zero and variance equal to 0.04 as the interaction value can be positive as well as negative. Again, we do not estimate the hyper-parameters of this distribution, since we only have one interaction term on each deepest level.

4.2 Model estimation

In this section we describe how the model can be estimated using either the frequentist or Bayesian approach.

4.2.1 Frequentist approach

The frequentist estimation approach purely relies on the data. This means that it ignores the hierarchical prior framework as constructed in Section 4.1.1 and, instead, aims to find estimates for the parameters which best suit the data. One way of achieving this is by means of the Maximum Likelihood (ML) estimator, which is also used by OP. If we assume the data of deepest level l at channel j consists of T data points and is normally distributed with standard deviation σ_{jl}, this comes down to maximizing the log of the likelihood function

$$\log L(\theta) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(\sigma_{jl}^2) - \frac{1}{2\sigma_{jl}^2}\sum_{t=1}^{T}\Big(\log(y_{jlt}+1) - \log\widehat{(y_{jlt}+1)}\Big)^2, \qquad (6)$$

where vector θ contains the parameters to be estimated, $\log(y_{jlt}+1)$ is the log-transformed observed attribution and $\log\widehat{(y_{jlt}+1)}$ the predicted value as described by (1). To estimate the parameters, we apply a numerical search algorithm. In this study, we use one of the common used Quasi-Newton methods, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm.

4.2.2 Bayesian approach

In case of the Bayesian approach, parameters are viewed as random variables. It starts with the distribution of the data y conditional on its parameters θ, $p(y|\theta)$, also called the likelihood function. Together with a prior distribution, $p(\theta)$, which contains information about the uncertainty of θ, we can apply the Bayes theorem and estimate a posterior distribution for θ by

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \propto p(\theta)p(y|\theta), \qquad (7)$$

where the marginal distribution $p(y)$ is omitted as it is independent of θ and can be seen as a scalar factor. The resulting expression, $p(\theta)p(y|\theta)$, is called the kernel of the posterior distribution. One can say that the prior for θ has been updated with the information contained in the data. When the prior again is dependent on another hyper-parameter, we get a sequence of conditional distributions, often called a hierarchical model. For example, the first-layer prior can be specified as $p(\theta|\tau)$ and the second-layer prior as $p(\tau|h)$ where h is known.

To estimate the advertising elasticities we implement a similar hierarchical Bayes model, where the hierarchy of the marketing channels is incorporated into the priors as stated in Section 4.1.1. Note that we also include the non-hierarchical Bayesian (NHB) model in our study. In this way we examine whether the increased performance of the HB model in comparison to the ML model is not only due to the Bayesian estimation technique. In other words, we investigate the added value of the hierarchy incorporated in the Bayesian framework. The NHB model has the same prior distributions as the HB model, except for the elasticity prior. To ensure the difference in performance is not due to the hyper parameter values, we have to match the elasticity prior on the deepest level of the HB model in its unconditional form to the non-hierarchical elasticity prior of the NHB model. This is done by integrating over β_j and β_{jk}. For β_{jkl} we then get a $Normal(\mu_{\beta_j}, \nu_l^2 + \nu_k^2 + \nu_j^2)$ distribution with the same hyper parameter values as in the HB model.

The output of the Bayesian model is a posterior density for each parameter which was to be estimated. For further analysis, often one preferably works with a point estimator instead of a distribution. Therefore, the posterior mean can be used as an estimator. Calculating the posterior mean involves taking an integral of the posterior distribution, which often has no analytical solution. Therefore, we need to rely on sampling methods. In this study we implement the No U-turn Sampler (NUTS) which is a self-tuning variant of the Hamiltonian Monte Carlo sampling algorithm (Hoffman and Gelman [2014]). This method is available in programming languages which focus on Bayesian statistical modeling, for instance Stan, but also in a Python package called PyMC3 which we use in this research.

4.3 Model configurations

OP normally uses the last fifteen data points for which a channel was active to estimate its elasticity. The idea behind this decision is that a lot can change in fifteen weeks. Therefore, OP only uses data which is most recent. Even so, a lot of information is lost by excluding the weeks before this period. By varying the length of this period, we will investigate how much effect this restriction has on the estimates.

In addition, it might be the case that there are not enough data points available, as OP has set a restriction on the maximum weeks one can go back in time. In that case OP has chosen to let the elasticity of the parent be adopted instead. However, in the Bayesian approach there is no need for such a correction. In case there are few data points available for a channel, the Bayesian model automatically returns a posterior distribution close to its prior, which is the posterior distribution of the parent.

4.4 Model evaluation

This section describes how each model is evaluated. To start, Section 4.4.1 gives an outline of the simulation we implement to verify that the models produce estimates close to the true parameters. Furthermore, the forecast accuracy of each model is evaluated by means of the measures as defined in Section 4.4.2. In addition, for the Bayesian approach we also carry out a density forecast which is stated in Section 4.4.3.

4.4.1 Simulation study

When we use the available real data as input for the models, we cannot indicate whether the estimated elasticities are close to the "true" elasticities since these are unknown. Hence, we will start by conducting a simulation study to investigate whether the HB model will produce reliable estimates. In addition, this simulation is also of assistance when choosing sensible priors.

Under the hypothesis that marketing investments belonging to the same channel have comparable elasticities, we generate our own cost and attribution data containing the same hierarchical structure as the tree of paid social showed in Figure 1. This is done by choosing the intercepts and elasticity values of the deepest level, with the restriction that the chosen elasticities do not differ too much from each other. In each iteration of the simulation, we generate N points between a range comparable to the range of the paid social campaign costs. Plugging these points into the regression function (1) and adding a random error term, this creates N log-transformed "revenue" observations of each artificial child of the paid social channel. In this way, we call the artificial performance data "00" and "01", while we call the artificial brand data "10" and "11".

The elasticities are estimated using both frequentist and Bayesian approach and compared to the true parameter values. For the ML approach, the simulation is repeated a thousand times, to ensure the estimation error is not caused by the variation in the data. For the Bayesian approach the number of replications is only a hundred, as this method is less dependent on variation in the data and repeating the simulation more often does not change the results. Note that we include the parameter transformation (3) as described in Section 4, to get a more complete view of the performance of the model. After checking the performance of the models in this ideal situation, outliers are added to examine the performance of the models in a setting close to the environment OP is facing. To be more precise, for each child of the artificial paid social data, we add one outlying point. These data points are out of the expected range in either horizontal or vertical direction or in both directions. The outliers are encircled in Figure 6 of Section 5.1.

4.4.2 Point forecasts

After the simulation, we assess the forecast accuracy of the models using the real data as described in Section 3. To start, this is done by means of out-of-sample point forecasts using a moving window of R data points. Here, we set R equal to 15 as this is the number of weeks OP normally uses for predictions. In this way, for the prediction of week 20, the 15 most recent weeks are used for which the marketing channel was active. As performance measures we use the Root Mean Squared Prediction Error (RMSPE) and the Symmetrical Mean Absolute Percentage Error (SMAPE) which are defined as

$$RMSPE = \sqrt{\frac{1}{T-R}\sum_{t=R}^{T-1}(y_{t+1} - \hat{y}_{t+1|t})^2}, \tag{8}$$

$$SMAPE = \frac{100\%}{T-R}\sum_{t=R}^{T-1}\frac{|y_{t+1} - \hat{y}_{t+1|t}|}{(|y_{t+1}| + |\hat{y}_{t+1|t}|),/2} \tag{9}$$

where y_{t+1} is the observed revenue at time $t+1$ and $\hat{y}_{t+1|t}$ the predicted revenue at time $t+1$ given time t. We correct for the fact that the expectation of $e^{\varepsilon_{jlt}}$ equals $e^{\frac{1}{2}\hat{\sigma}^2_{jl}}$. The SMAPE is examined next to the RMSPE as it is more sensitive to under-forecasting. Besides, OP uses this measure for internal validation of models. We compare the frequentist approach to the Bayesian approach, as well as the different model specifications.

Note that we have chosen to forecast and assess the performance of the models based on the original data, meaning without log-transformation. The reason for this, is that our main interest lies in predicting the revenue, not log-transformed revenue. Therefore, we want to evaluate the prediction errors on this scale as well. Besides, peaks in the data are scaled down in case of log-transformed data, making them less important with regard to predictions. Using the original data, this results in a more realistic view of the prediction accuracy, in the sense that the penalty becomes larger when prediction errors are larger.

4.4.3 Density forecasts

The disadvantage of point forecasts is that it ignores the uncertainty of the estimation. Therefore, we also consider the prior and posterior predictive distribution for the Bayesian model. First of all, for the prior predictive distribution, the data is generated entirely by the prior distributions

$$p(y_{pred} \mid h) = \int_\theta p(y_{pred} \mid \theta) p(\theta \mid h) \mathrm{d}\theta, \tag{10}$$

where $p(y_{pred} \mid \theta)$ is the likelihood of an unseen data point y_{pred} and $p(\theta \mid h)$ is the prior distribution conditional on its hyper-parameter h which is known, as described in Section 4.2.2. This predictive distribution, which is the marginal distribution of y_{pred}, is helpful to check whether the priors generate a distribution with realistic y values. Secondly, after we have obtained the posterior distributions of the parameters based on the observed data points $y = y_1, \ldots, y_T$, we generate a future data point y_{T+1} from the model by means of the posterior predictive distribution

$$p(y_{T+1} \mid y, h) = \int_\theta p(y_{T+1} \mid \theta, y) p(\theta \mid y, h) \mathrm{d}\theta = \int_\theta p(y_{T+1} \mid \theta) p(\theta \mid y, h) \mathrm{d}\theta, \tag{11}$$

where $p(\theta \mid y, h)$ is the posterior distribution and $p(y_{T+1} \mid \theta)$ in y_{T+1}. Here, the last expression is achieved because we assume observed and future data points are conditionally independent given θ. In other words, this is the distribution of a future data point conditional on the observed data points.

5 Results

We start by discussing the results of the simulation study in Sections 5.1. After that, we estimate the parameters using a selected period of the real data. The estimations are discussed in Section 5.2. Furthermore, Section 5.3 provides further insights of the HB model regarding the prior and posterior distributions. In addition, the performance of the two approaches are compared in Section 5.4. For all Bayesian estimations, the number of chains is set to two, each with four thousand draws, half of which is used to tune. The remaining draws of both chains are used for analysis. The target accepting rate is set to 0.99, as a lower target sometimes results in divergences.

5.1 Simulation study

Figure 5 displays the simulated data of one of the replications and the corresponding true regression lines as described in Section 4.4.1. The 00 and 01 data, which can be seen as the performance data from the paid social channel, has an elasticity of 0.65, while the true elasticity of the 10 and 11 data is 0.55, which can be seen as the brand data from the paid social channel. For the Bayesian approach we assume the same prior composition as described in Section 4, where we assume a prior value of 0.6 for μ_{β_0}.

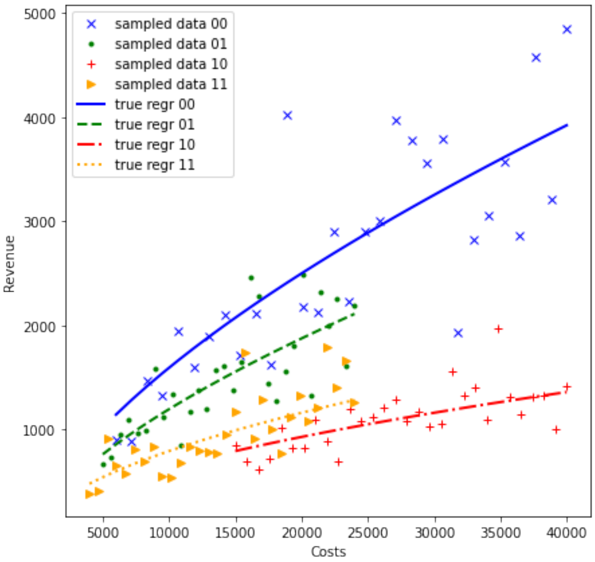

Figure 5: Simulated data and true regression curves.

Table 3 shows the true parameter values as well as the estimated parameters averaged over all replications, where a stands for the estimated intercept, and b for the estimated elasticity. In addition, Table 4 displays the mean absolute bias from the true values, as well as the two performance measures from Section 4.4.2. First of all, we set N equal to 15 as this is the number of observations OP is currently using to estimate the elasticities. We conclude that the HB model performs better than the ML model. Especially, the ML model appears to have a hard time estimating the intercept, as the corresponding average absolute bias is enormous in comparison to that of the HB model. As a result the RMSPE of the ML model is extremely large as well. Probably, this is caused by the small data size of the simulation. Therefore, we also conduct the simulation with a size of thirty data points instead, of which the results are shown in the right half of Tables 3 and 4. Now the ML estimates look much more stable, while the results of the Bayesian models hardly change. In this case the HB model still performs better than the ML model. We conclude that for the frequentist approach, fifteen data points is not enough to get reliable estimates.

Table 3: Estimated parameters of the simulated data sets, averaged over all replications.

		15 data points			30 data points		
	TRUE	ML	NH	HB	ML	NH	HB
avg. a							
00 data	4	68047.392	5.417	5.139	5.407	4.497	4.299
01 data	3	1198.679	4.191	3.917	4.199	3.381	3.223
10 data	4	242.333	2.910	3.240	12.341	5.525	4.733
11 data	5	211.639	8.142	7.220	7.078	5.335	5.140
avg. b							
00 data	0.65	0.641	0.643	0.646	0.663	0.646	0.648
01 data	0.65	0.625	0.635	0.639	0.650	0.652	0.653
10 data	0.55	0.545	0.614	0.599	0.550	0.560	0.563
11 data	0.55	0.516	0.542	0.555	0.537	0.562	0.563

Table 4: The relative average performance of the frequentist approach versus the Bayesian approach. The reported values are the average performance of the HB model as fraction of the average performance of the ML or NHB model.

	15 data points		30 data points	
	HB/ML	HB/NH	HB/ML	HB/NH
abs. bias a				
00 data	0.000	0.879	0.321	0.773
01 data	0.002	0.890	0.384	0.742
10 data	0.009	0.850	0.252	0.715
11 data	0.024	0.922	0.650	0.915
abs. bias b				
00 data	0.635	0.858	0.388	0.758
01 data	0.491	0.944	0.505	0.772
10 data	0.489	0.816	0.607	0.716
11 data	0.815	1.062	0.927	0.900
RMSPE				
00 data	0.000	1.002	0.810	1.000
01 data	0.014	1.005	1.057	1.002
10 data	0.250	0.997	0.917	1.003
11 data	0.065	1.004	1.022	0.998
SMAPE				
00 data	0.670	0.999	0.937	1.000
01 data	0.644	1.003	1.044	1.000
10 data	0.890	1.004	0.891	1.006
11 data	0.473	1.005	1.019	0.992

Furthermore, if we focus on the Bayesian approach and compare the NHB estimation to the HB estimation, we see that the HB model generates estimates with a bias smaller than the NHB model. However, with respect to the RMSPE and SMAPE we conclude that the NHB model performs slightly better in most cases or the Bayesian models perform about equal. Besides, to ensure the good performance of the Bayesian models is not only due to the choice of the prior distributions, we also run the simulation with a prior value of 0.4 or 0.8, with N equal to 15. It turns out that this hardly changes the results, which means there is enough info in the data to move the posterior distribution in the right direction.

Next to this ideal situation, we add some outlying points to the data, see Figure 6, and run the simulation again. Table shows that the HB model clearly performs better than the ML model in three out of four of the cases. Only for the 11 data ML performs slightly better.

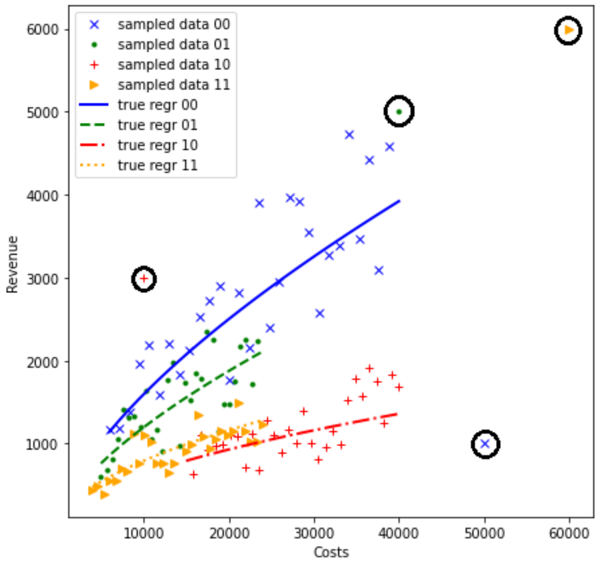

Figure 6: Simulated data containing outlying data points which are encircled and true regression curves.

Table 5: Average estimated parameters of the simulated data sets, each with thirty data points and outliers like Figure 6.

	avg. a			avg. b		
	TRUE	ML	HB	TRUE	ML	HB
00 data	4	409595.923	10.392	0.650	0.481	0.553
01 data	3	92.131	2.736	0.650	0.735	0.668
10 data	4	1201.568	16.939	0.550	0.070	0.423
11 data	5	1.284	2.438	0.550	0.712	0.635

Table 6: Average performance of the HB model as a fraction of the ML model. Each simulated data set has thirty data points and outlying points like Figure 6.

HB/ML	abs. bias a	abs. bias b	RMSPE	SMAPE
00 data	0.000	0.576	0.000	0.517
01 data	0.008	0.326	0.123	0.984
10 data	0.011	0.264	0.452	0.930
11 data	0.689	0.528	1.075	1.002

5.2 Estimated parameters

Now we concentrate on estimation of the parameters using real data. We select the last fifteen data points for which each channel was active counted from week 45. This period is chosen since most channels were active here. Table 7 gives a summary of the estimated parameters of both approaches. While for the frequentist approach we state the value of the estimated parameter itself, for the Bayesian approach this is the mean of the posterior distribution for each parameter. We see that the elasticities estimated by the ML model are clearly more "extreme" than the results of the HB model, in the sense that they vary a lot more. This is no surprise, as from the simulation we know that fifteen data points probably is not enough to generate reliable estimates in case of the frequentist approach, while in case of the Bayesian approach it is. For example, for the elasticities of brand - paid social we have a ML estimate of 0.236 for awareness and 0.845 for consideration. More realistic outcomes would be much closer together, like the HB estimates, which also rely on information from the priors.

It is important to note that there is no added value in testing whether the estimated elasticities are significantly different from zero. This is because we have assumed the elasticities are larger than zero and the reported values are estimated as in Equation (3). This means that when the elasticity has a value of zero, the "raw" value β^* has a value of 0.10. Instead, for the Bayesian elasticity parameters it might be of interest to investigate whether they are significantly different from each other. This will be studied in the next section.

Table 7: Estimated parameters of the ML model and the HB model, based on the last fifteen active weeks counted from week 45. Note that $\tilde{\alpha}$ is the standardized estimate and β is the estimated elasticity, meaning the transformed version as in Equation (3), not β^*, where we leave out the subscripts for simplicity. For HB the estimated model standard deviation is 0.845.

		ML		HB	
channel	variable	estimate	sd	mean	sd
shopping - branded - paid search	$\tilde{\alpha}$	7.845	0.236	8.011	0.259
	β	0.364	0.107	0.472	0.100
	σ	0.728	0.129	-	-
text - branded - paid search	$\tilde{\alpha}$	12.421	0.062	12.211	0.229
	β	0.087	0.088	0.511	0.137
	σ	0.205	0.037	-	-
text - non branded - paid search	$\tilde{\alpha}$	11.318	0.071	11.345	0.221
	β	0.822	0.252	0.683	0.128
	σ	0.344	0.063	-	-
text - semi branded - paid search	$\tilde{\alpha}$	11.487	0.033	11.538	0.220
	β	1.263	0.201	0.645	0.161
	σ	0.111	0.019	-	-
shopping - non branded - paid search	$\tilde{\alpha}$	6.969	0.329	6.968	0.221
	β	0.728	0.079	0.718	0.051
	σ	1.238	0.227	-	-
awareness - brand - display	$\tilde{\alpha}$	5.130	0.268	5.092	0.221
	β	1.294	0.409	0.623	0.143
	σ	1.057	0.187	-	-
consideration - brand - display	$\tilde{\alpha}$	8.246	0.120	8.344	0.215
	β	1.169	0.142	0.636	0.141
	σ	0.462	0.083	-	-
prospecting - performance - display	$\tilde{\alpha}$	9.770	0.095	9.793	0.219
	β	0.130	0.367	0.426	0.155
	σ	0.370	0.066	-	-
retargeting - performance - display	$\tilde{\alpha}$	9.966	0.221	9.970	0.218
	β	0.435	0.253	0.430	0.142
	σ	0.853	0.152	-	-
action - performance - paid social	$\tilde{\alpha}$	6.804	0.219	6.811	0.217
	β	0.640	0.747	0.544	0.169
	σ	0.846	0.152	-	-
desire - performance - paid social	$\tilde{\alpha}$	7.377	0.359	7.294	0.224
	β	1.809	1.248	0.563	0.165
	σ	1.378	0.252	-	-
awareness - brand - paid social	$\tilde{\alpha}$	6.335	0.099	6.314	0.221
	β	0.236	0.251	0.501	0.164
	σ	0.384	0.070	-	-
consideration - brand - paid social	$\tilde{\alpha}$	7.482	0.253	7.442	0.219
	β	0.845	0.947	0.512	0.168
	σ	0.904	0.162	-	-

5.3 Insights Bayesian model

Predictive distributions By sampling from the prior and posterior predictive distributions of the log-transformed dependent variable, we come to Figure 7. We generate five hundred samples, which are the blue curves. The predictive mean is then given by the dashed blue curve. The left plot shows the prior predictive distribution of the log-transformed dependent variable, as well as the corresponding observed values. The prior predictive distribution is very wide, which is due to the uninformative prior for the intercept. The right plot of Figure 7 shows that the posterior predictive distribution is neatly around the observed data points. This means that the prediction of an unseen data point will be in the same range as the observed data points.

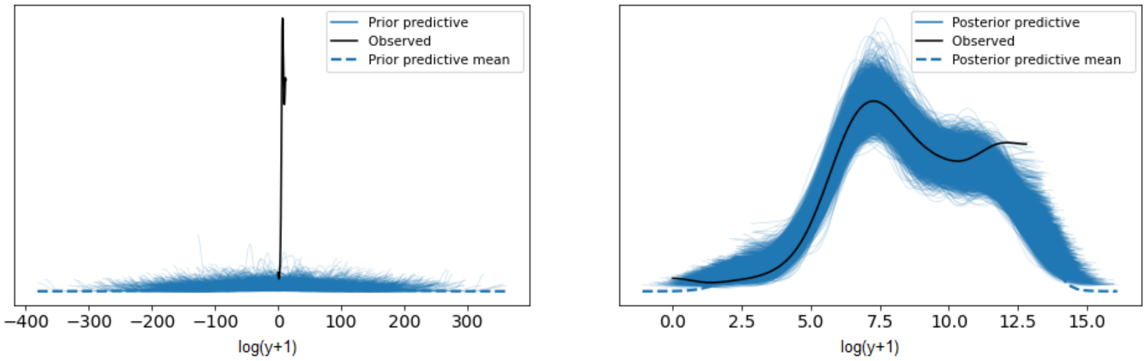

Figure 7: Five hundred samples from the prior predictive distribution (left) and posterior predictive distribution (right) of the dependent variable which is the log-transformed revenue.

Posterior distributions First of all, from the sample draws plotted in sequential order in Figure 14 of the Appendix we conclude that all posterior distributions have converged. There are two curves for each posterior, one for each chain, which show similar distributions. In addition, Figure 8 displays the posterior distributions of the model standard deviation and the intercepts. For the intercepts we only have posterior distributions on the deepest level as this variable is not integrated in the prior hierarchy. As a result, the intercept posteriors are not close together, with values ranging from four to fourteen. Besides, this is because of the flat prior given to this parameter.

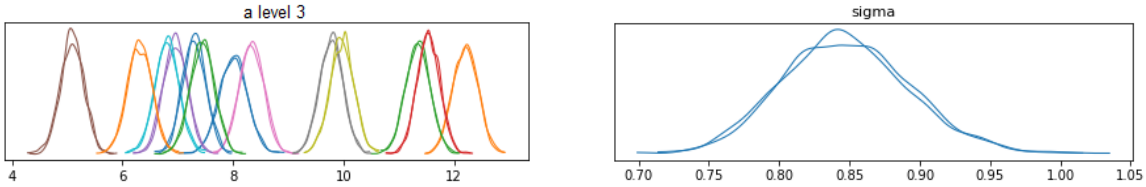

Figure 8: Posterior distributions of the intercepts on Level 3 (left) and the posterior distribution of the model standard deviation (right). There are two curves for each posterior, one for each chain.

Furthermore, Figure 9 shows the posterior distributions of the elasticities in the way the hierarchy is structured. These posteriors are much closer together than those of the intercepts, since the former share information in case they are from the same channel. As mentioned in the previous section, there is no value in testing whether the point estimates are significantly different from zero. However, for the Bayesian elasticity estimates we can investigate whether the posterior distributions on each level within a certain branch are significantly different from each other. Note that this chance is small, since children within the same channel have the same prior. However, when the data contains enough information, this should still lead to different posteriors. To investigate this, we take the posterior draws for example from the children of paid social, which are brand and performance. Then we determine the distribution consisting of all the differences and check whether zero is contained in the eighty percent highest posterior density interval. On Level 3 this is the case for all duos, while on Level 2 this is only the case for paid social. We conclude that on Level 3 the data does not contain enough information anymore, to generate an elasticity distribution that is different from their neighbors on the same level. This could be a reason to simplify the hierarchy to one with Level 2 as deepest level and thus assuming that children on Level 3 have the same elasticity. Another option would be to use daily aggregated data instead of weekly aggregated data, since then the data contains more information which could result in the elasticities to become different as well on Level 3.

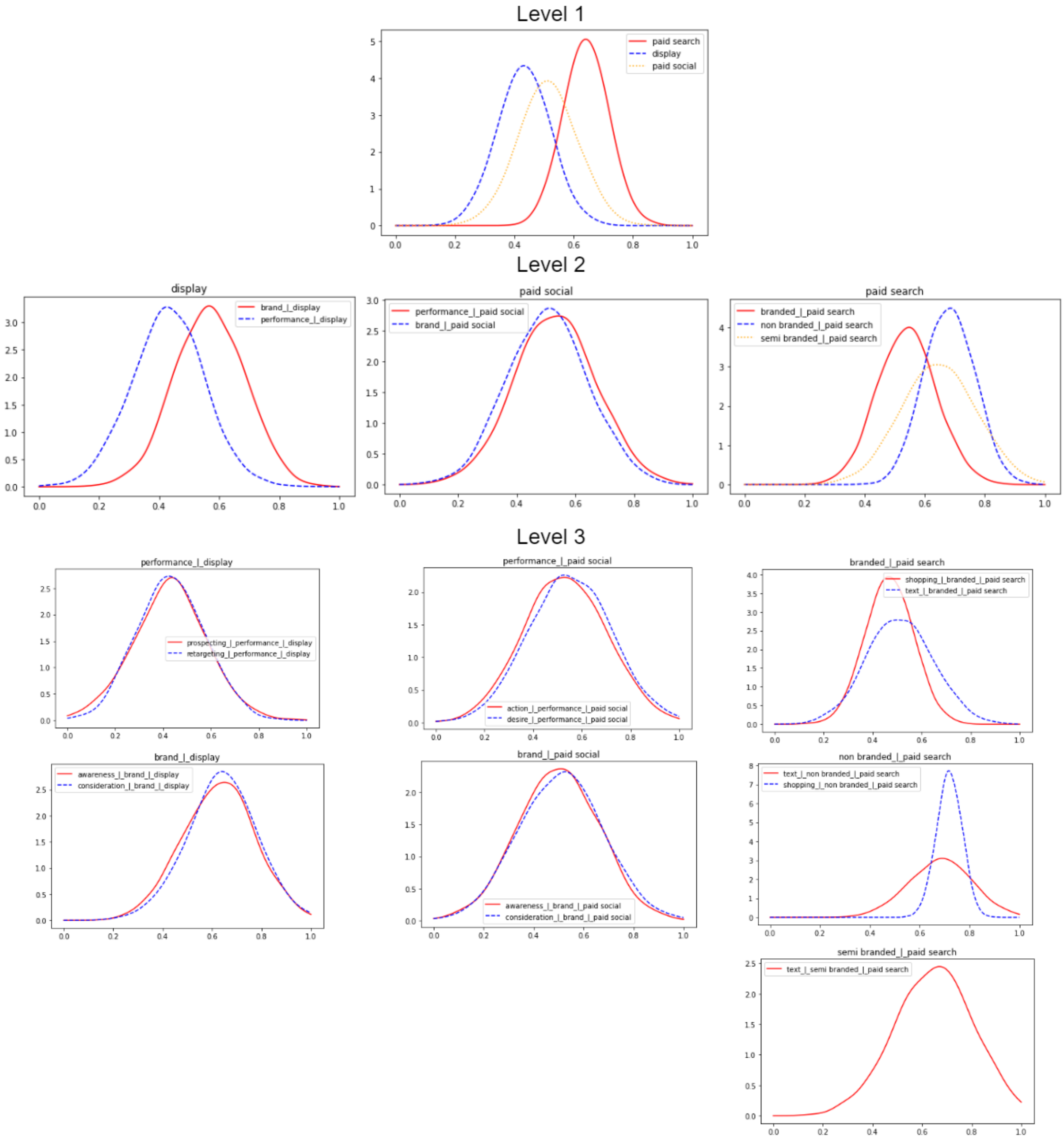

Figure 9: Posterior distributions of the elasticity parameters on each level in the way the hierarchy is structured.

5.4 Comparison of performance

5.4.1 Within sample performance

Table 8 shows the within sample performance of the frequentist approach and Bayesian approach, based on the estimated parameters as stated in Table 7. We have also included the NHB model to get a more complete view. If we compare the HB to the ML model we see that for display en paid social, we get mixed results. In half of the cases HB is preferred, for the other half ML is preferred or their performances are very close together. For paid search, we conclude that the HB model does not perform well in comparison to the ML model. This is not unexpected, as for the HB model we need the children within the same channel to have cost-revenue combinations which are not very different from each other, implying that they will have comparable elasticities. From Figure 2 in Section 3.1 we had already learnt that this channel does not meet this requirement. If we compare the NHB model to the HB model, we conclude that the latter is often preferred, but their performances are comparable.

Table 8: Within sample performance of the last fifteen weeks counted from week 45 based on the estimated parameters as stated in Table 7. For the RMSPE and SMAPE we use the revenue data which is not log-transformed.

	RMSPE		SMAPE	
	HB/ML	HB/NH	HB/ML	HB/NH
shopping - branded - paid search	1.332	0.971	1.126	0.990
text - branded - paid search	2.823	0.922	2.167	0.926
text - non branded - paid search	1.536	0.980	1.344	0.995
text - semi branded - paid search	3.800	0.961	4.035	0.965
shopping - non branded - paid search	0.347	0.934	0.806	0.985
awareness - brand - display	0.753	0.984	1.084	0.973
consideration - brand - display	0.725	0.954	1.075	0.948
prospecting - performance - display	1.447	1.028	1.603	1.024
retargeting - performance - display	0.997	1.013	1.000	0.976
action - performance - paid social	1.003	0.996	1.005	0.994
desire - performance - paid social	0.404	1.002	0.728	1.001
awareness - brand - paid social	1.409	1.001	1.424	1.001
consideration - brand - paid social	1.002	1.000	0.978	0.998

Furthermore, Figure 10 displays the estimated regression lines for the three different models for each deepest level based on the estimates in Table 7. The Figure shows that often, the differences between the two approaches is small. However, in four ML cases we have an estimated regression line of exponential form due to an estimated elasticity larger than one. This is not very plausible since media advertising investments are known to have diminishing returns. Besides, for the ML regression line of shopping - non branded - paid search there is an over-estimation caused by the estimated model standard deviation which has a value of 1.238.

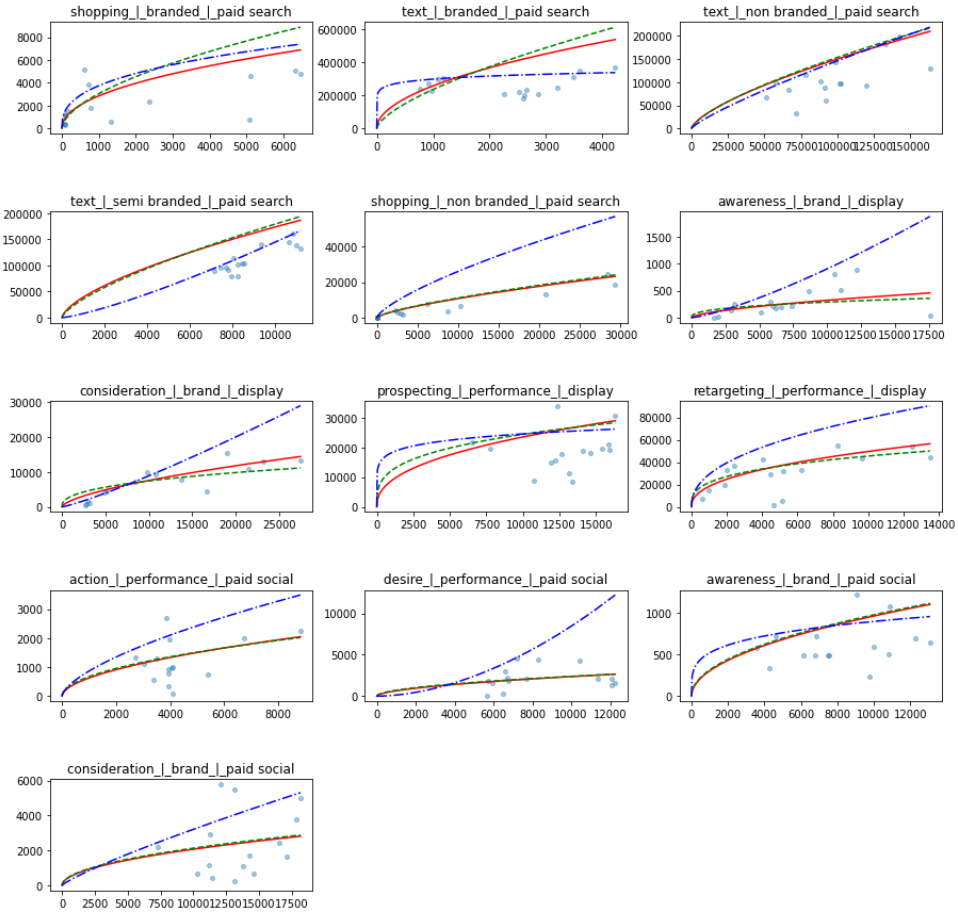

Figure 10: Estimated regression lines of the ML model (blue) versus the HB model (red) versus the NH model (green) for each deepest level, based on the last fifteen active weeks counted from week 45.

5.4.2 Out-of-sample performance

Next, we carry out out-of-sample point forecasts using a moving window. First of all, we set the length of the moving window to fifteen data points. Therefore, the earliest week to be predicted is week 16. The following table shows the average of the two performance measures over all forecasts. In addition, the last row contains the number of predictions. This number is not equal for all deepest levels, since we have an unbalanced panel. In other words, sometimes the test week is not part of the data set as not every deepest level has a data point every week.

First of all, we conclude that for paid social the HB model performs better than the ML model. In addition, for display we have mixed results, preferring the HB model in two of the four cases. However, for paid search we come to the same conclusion as in the previous section; the HB model does not perform well in comparison to the ML model. This means the children of paid search differ too much from each other to be incorporated in the hierarchy used in this paper. Especially text - branded and text - semi branded perform poorly. This is not unexpected looking back at the scatter plot on the right of Figure 2, which shows these two children have relatively small costs compared to the generated revenue. Next, when we compare the HB model to the NHB model, the former leads to better performance in most of the cases but the differences are small.

Table 9: Average performance over all out-of-sample predictions per deepest level using a moving window of 15 data points. The reported values are the average RMSPE (or SMAPE) of the HB predictions as fraction of the average RMSPE (or SMAPE) of ML or NHB.

	RMSPE		SMAPE		
	HB/ML	HB/NHB	HB/ML	HB/NHB	# pred
shopping - branded - paid search	1.182	1.011	0.976	0.994	38
text - branded - paid search	3.442	0.966	2.523	0.978	38
text - non branded - paid search	1.262	0.994	1.233	0.998	38
text - semi branded - paid search	2.359	0.983	1.966	0.986	38
shopping - non branded - paid search	0.477	0.989	0.648	1.003	25
awareness - brand - display	0.608	0.982	0.916	0.990	25
consideration - brand - display	1.521	0.992	1.255	0.993	38
prospecting - performance - display	1.544	0.993	1.413	0.994	38
retargeting - performance - display	0.003	0.990	0.971	0.994	37
action - performance - paid social	0.848	1.006	0.884	1.004	35
desire - performance - paid social	0.000	0.993	0.485	0.995	32
awareness - brand - paid social	1.106	0.995	1.059	0.997	9
consideration - brand - paid social	0.000	1.001	0.746	1.001	38

By way of illustration, Figure 11 displays the out-of-sample predictions of four of the deepest levels over time, compared to the actual revenue values. The upper two plots show two examples for which both frequentist and Bayesian approach follow the movement of the data quite well. The bottom right plot of Figure 11 shows that the NHB and HB models predict a peak around week 40 which is not there. This is probably caused by the before-mentioned observation that the hierarchy for this part of the data is not appropriate. On the contrary, for awareness - brand - display in the bottom left plot, the ML prediction is way too extreme around week 25.

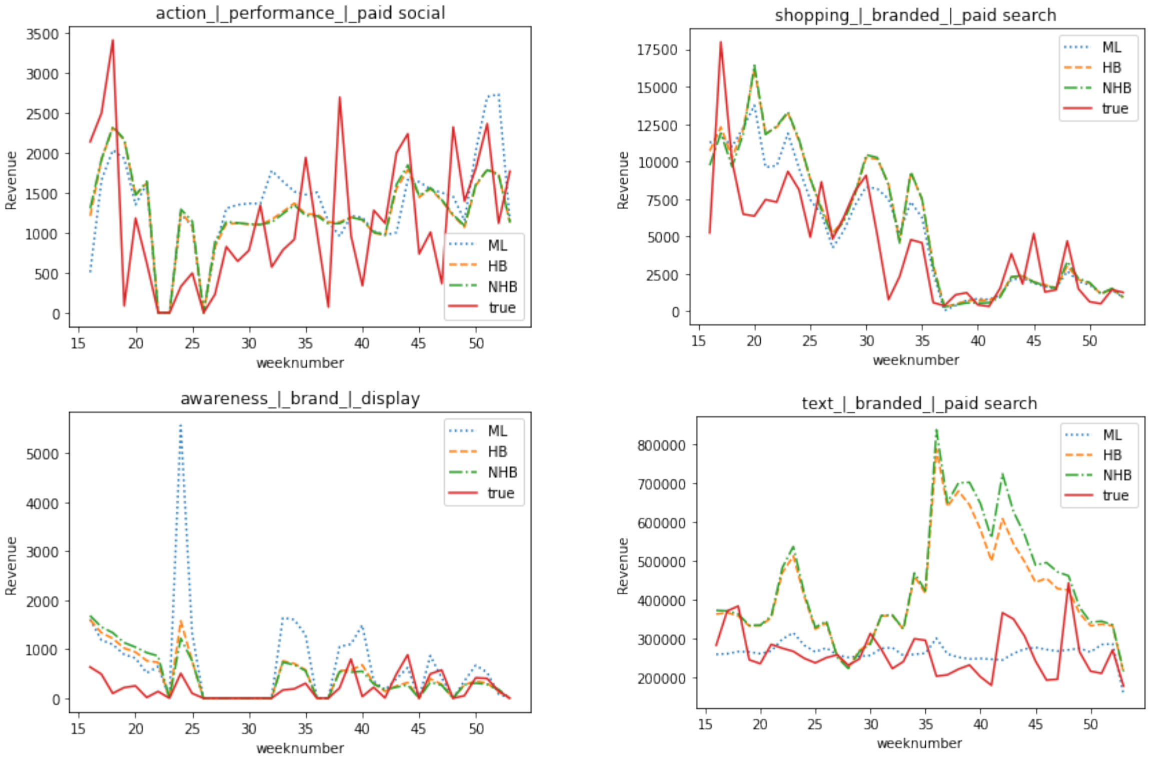

Figure 11: Out-of-sample predictions over time for four of the deepest levels, compared to the true revenue values.

Furthermore, we investigate some ML predictions with a very large RMSPE, which concerns two children from paid social, as well as one from display. It turns out that this is caused by a handful of the predictions with extremely large outcomes. Probably, this is because fifteen data points is not enough for the frequentist approach to get stable estimations, like we derived from the simulation. Therefore, we also consider the out-of-sample predictions using a moving window of thirty data points, instead of fifteen, of which the performances are compared in Table 10. As expected, the results of the ML model do not contain the aforementioned extreme outcomes anymore. Nevertheless, for the other deepest levels the moving window of fifteen data points is preferred. Similarly, for the HB and NHB model using the last fifteen active weeks for a prediction is preferred over using the last thirty weeks in most of the cases. This supports the fact that advertising elasticities are not constant over time and it stresses the importance to estimate the elasticities based only on the most recent data available.

Table 10: Relative out-of-sample performance of the predictions based on a moving window of the last fifteen active weeks versus those based on the last thirty active weeks.

	rel. RMSPE			rel. SMAPE		
	ML	HB	NHB	ML	HB	NHB
shopping - branded - paid search	0.006	1.223	1.254	0.714	0.780	0.805
text - branded - paid search	0.805	0.556	0.533	0.782	0.611	0.598
text - non branded - paid search	0.915	0.807	0.783	0.948	0.884	0.864
text - semi branded - paid search	0.752	0.657	0.654	0.780	0.668	0.670
shopping - non branded - paid search	0.775	0.826	0.830	1.244	1.297	1.304
awareness - brand - display	0.486	0.900	1.079	0.763	1.013	1.092
consideration - brand - display	0.408	1.675	1.468	1.074	1.634	1.428
prospecting - performance - display	0.877	0.905	0.886	0.740	0.814	0.807
retargeting - performance - display	1.721e2	0.483	0.487	0.949	0.718	0.715
action - performance - paid social	1.006	1.059	1.062	1.135	1.138	1.141
desire - performance - paid social	0.582	0.847	0.862	1.026	0.988	0.997
awareness - brand - paid social	0.941	0.907	0.912	0.961	0.955	0.958
consideration - brand - paid social	6.964e3	0.832	0.831	1.225	0.905	0.903

To clarify this, Figure 12 shows the predicted elasticities over time for two of the deepest levels based on a moving window of fifteen data points (ML15, HB15) versus those based on a moving window of thirty data points (ML30, HB30). We observe that the ML15 and HB15 elasticities fluctuate a lot over time, while the ML30 and HB30 are more flattened out. In addition, if we compare the ML and HB predictions it is noticeable that the former elasticities are located in a much wider range than the latter. For example, for text - non branded - paid search, the HB15 elasticities range from 0.6 to 0.8, while for the ML15 elasticities this is 0.2 to 0.9. This is caused by the information in the prior, which ensures that the HB elasticities stay within a range that is reasonable for the corresponding channel.

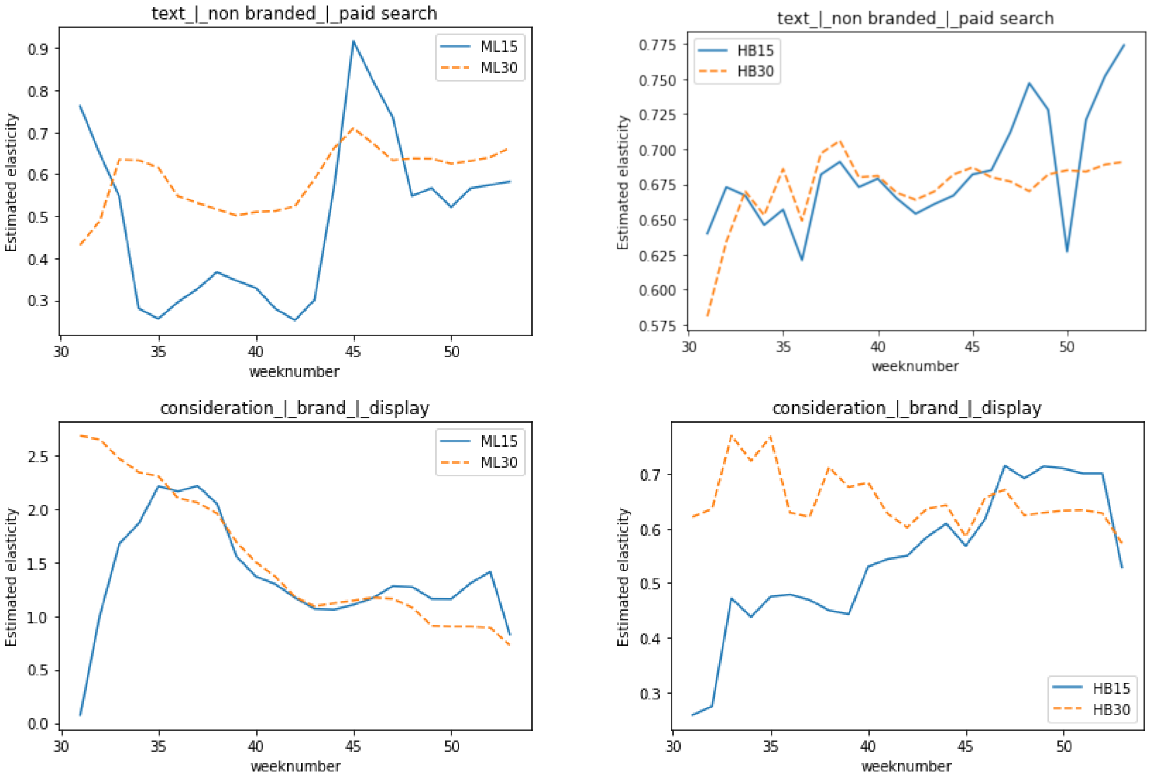

Figure 12: Out-of-sample predicted elasticities over time for two of the deepest levels, based on a moving window of fifteen data points (ML15, HB15) versus those based on a moving window of thirty data points (ML30, HB30).

5.4.3 Out-of-sample performance with interactions

As a side track of this research, we examine the change in performance if an aggregated interaction term is added to the model, as specified in Section 4.1. The estimated parameters of the ML model and the HB model based on the last fifteen active weeks counted from week 45 are stated in Table 15 of the Appendix. The interaction estimates range from -0.984 to 0.363 for the ML model, often with a standard deviation larger than the estimate itself, making them not significant. For the HB model the point estimates for the interactions are somewhat smaller, ranging from -0.286 to 0.107, due to the assigned prior. Also for this parameter, zero is contained in the eighty percent highest density interval for all cases, which is an indication the interaction term does not have a significant effect.

By comparing the out-of-sample performance of the models with and without the interaction term, we assess the effect of the aggregated interaction term. We use a moving window of the last fifteen active weeks to predict the next week, since we observed this is preferred as opposed to the last thirty active weeks. To get a more complete picture, the relative performance of the NHB model is also included. In Table 11 we see that for almost all cases, the performance of the model with interactions is worse than the model without interactions if ML estimation is used. The fact that this method completely relies on the observed data, together with the way in which the interaction term is constructed, cause the interaction term to have an effect which is sometimes out of proportion from what we would expect. However, for the HB and NHB estimation this is the other way around. Here, the model with interactions is nearly always preferred over the model without interactions, even though the difference in performance is small. This difference is caused by the influence of the prior on the interaction term. The prior we gave to this parameter ensures that the interaction effect is not too large. This indirectly ensures that in general it stays smaller than the effect of the elasticity itself, which is not an unrealistic assumption. We conclude that when one wants to incorporate interactions into the model specification, Bayesian estimation should be used.

Table 11: Relative out-of-sample performance of the predictions based on the model with interactions versus those based on the model without interactions. We use a moving window of the last fifteen active weeks.

	rel. RMSPE			rel. SMAPE			
	ML	HB	NHB	ML	HB	NHB	#pred
shopping - branded - paid search	1.134	0.845	0.850	1.025	0.937	0.937	38
text - branded - paid search	1.055	0.994	0.983	1.131	0.967	0.957	38
text - non branded - paid search	1.280e5	0.949	0.957	1.139	0.944	0.946	38
text - semi branded - paid search	1.015	0.904	0.899	0.996	0.927	0.923	38
shopping - non branded - paid search	1.127	1.938	1.936	1.121	1.336	1.340	25
awareness - brand - display	1.287	1.345	1.331	0.914	1.045	1.042	25
consideration - brand - display	1.035	0.861	0.857	1.106	0.957	0.955	38
prospecting - performance - display	1.108	0.929	0.937	1.038	0.949	0.956	38
retargeting - performance - display	1.035e10	0.988	0.990	1.027	0.985	0.981	37
action - performance - paid social	1.499e2	0.939	0.942	1.012	0.963	0.966	35
desire - performance - paid social	0.382	1.007	1.004	0.703	0.999	0.997	32
awareness - brand - paid social	1.155	0.999	0.998	1.119	0.990	0.992	9
consideration - brand - paid social	0.000	1.046	1.045	0.882	1.025	1.023	38

6 Discussion

In this section, we provide a compact summary of our findings and give advice regarding estimation of the elasticities. In addition, we point out several directions for further research, which emanate from the limitations of this paper.

6.1 Summary and conclusion

In this research, a hierarchical Bayesian approach is considered to estimate marketing advertising elasticities. By incorporating the existing tree structure of the marketing investments into the prior of the elasticities, we let investments belonging to the same channel share information with each other. In this way, we tackle the problem of sparse data points, as not every kind of marketing is done every week.

The Bayesian elasticity model is compared to the frequentist elasticity model, which purely relies on the data. Based on both the within and out-of-sample performances, we have found that the HB estimation is preferred over ML estimation when data points from the same channel have a comparable scale. In that case, information contained in data points belonging to investments higher up in the hierarchy, positively contribute to estimation of the elasticities deeper in the hierarchy. In addition, the result that the HB model nearly always performs as well as or better than the NHB model, shows that the increase in performance compared to the ML model is not only due to the Bayesian estimation technique. However, when the cost-revenue combinations for different children within the same channel are really diverse, this is not beneficial for the Bayesian performance. In that case, the advice would be to define the hierarchy differently. For example, split branded and non-branded paid search and consider them as two different channels. Another option would be to use the NHB model, for which the hyper parameter values of the elasticity prior are chosen based on information of the corresponding deepest level, instead of information of the corresponding channel as was done in this paper.

Next to this main finding, we have investigated what it means to use up to only fifteen data points when estimating the elasticities. From a simulation we have derived that in that case, we do not get reliable ML estimates. For the Bayesian estimation, however, we already get good performances with fifteen data points. This becomes even more evident when outlying points are added to the simulation data.

Even though there is less information contained in data points from the last fifteen weeks a channel was active, as opposed to data from the last thirty weeks, still the former setting is preferred given the results from the real data. These show that out-of-sample predictions based on a moving window of fifteen data points perform better than those based on a moving window of thirty data points. This is in line with the fact that advertising elasticities are not constant over time. Therefore, only the most recent data points should be used for predictions. Given these results, the advice is to further investigate the estimation of the elasticities based on daily aggregated data. In this way, information which was lost at first by aggregating the data at a weekly level, can decrease the estimation uncertainty, while still using data from the same period.

Moreover, when interactions are added to the model, the results have shown that the out-of-sample performances of the ML estimation decrease as opposed to those based on the model without interactions. On the contrary, for the Bayesian estimation the out-of-sample predictions show better performance when interactions are included in the model.

6.2 Limitations and further research

There are several topics which could be of interest regarding further research. This concerns the aggregation level of the data, the structure of the hierarchy, the side effects of the unpaid channels and the interaction terms.

Aggregation level of the data. As mentioned in the previous section, the elasticities are estimated using data which is aggregated on a weekly level. The disadvantage of this is that when a certain campaign is active for four days, a lot of information is lost by aggregating the data. Together with our finding that only the most recent data available should be used, it might be interesting to estimate the elasticities on a daily level instead. In that case the model should be extended, as we need to correct for the day of the week for instance. Note, however, that a potential disadvantage of using daily data is that it might be too granular for estimating elasticities.

Structure of the hierarchy. In this study we adopted the structure OP uses for the classification of the marketing channels, as this setting is most relevant to the client investment wise. Nevertheless, it could be that estimation wise, a HB model for which the data is structured with Level 2 on top (brand vs. performance for paid social and display), has more benefit from the information contained in this hierarchy. Furthermore, we have seen that the hierarchy implemented in this paper is not suitable for the paid search channel, as it does not meet the assumption that different kind of investments within this channel have comparable elasticities. Therefore, it might be of interest to OP to investigate whether they can develop a rule of thumb which assists in making a decision on whether investments from the same channel are qualified to have their information shared.

Side effects of the unpaid channels. The control regression in Section 3 showed that there are no significant negative effects of the three channels analysed in this paper on the direct attribution. However, there are significant positive effects for the paid search and display channel. The integration of these side effects into the elasticity model is an interesting topic for further research. Especially if the model is run for other marketing channels like email, there should be a correction. A potential solution could be to correct the dependent variable of the elasticity model beforehand. For example, assume a one percent increase in attribution of a paid marketing channel leads to 0.1% decrease in direct attribution. In that case, when the attribution of the paid marketing channel is corrected by this negative change beforehand, the estimated elasticity will be lower as well.

Interaction terms. A limitation of this research is that the interaction term incorporated in the model is only based on the data of the three channels we estimate the elasticities for. However, there are other marketing channels like television and radio, with which interactions can exist. If this is the case, they should be incorporated into the interaction term as well. Besides, to get more actionable insights, OP could consider a separate interaction term for each channel, instead of one aggregated interaction term.

References

F. M. Bass. A simultaneous equation regression study of advertising and sales of cigarettes. *Journal of Marketing Research*, 6(3):291–300, 1969.

F. M. Bass, N. Bruce, S. Majumdar, and B. Murthi. Wearout effects of different advertising themes: A dynamic bayesian model of the advertising-sales relationship. *Marketing Science*, 26(2):179–195, 2007.

J. D. Carroll, P. E. Green, and W. S. DeSarbo. Optimizing the allocation of a fixed resource: A simple model and its experimental test. *Journal of Marketing*, 43(1):51–57, 1979.

B. Dalessandro, C. Perlich, O. Stitelman, and F. Provost. Causally motivated attribution for online advertising. In *Proceedings of the sixth international workshop on data mining for online advertising and internet economy*, pages 1–9, 2012.

P. Doyle and J. Saunders. Multiproduct advertising budgeting. *Marketing Science*, 9(2):97–113, 1990.

A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

S. Henningsen, R. Heuke, and M. Clement. Determinants of advertising effectiveness: The development of an international advertising elasticity database and a meta-analysis. *Business Research*, 4(2):193–239, 2011.

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

D. Iacobucci, M. J. Schneider, D. L. Popovich, and G. A. Bakamitsos. Mean centering helps alleviate “micro” but not “macro” multicollinearity. *Behavior research methods*, 48(4):1308–1317, 2016.

H. S. Jagpal, E. F. Sudit, and H. D. Vinod. Measuring dynamic marketing mix interactions using translog functions. *Journal of Business*, pages 401–415, 1982.

A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285, 2005.

Y. Jin, Y. Wang, Y. Sun, D. Chan, and J. Koehler. Bayesian methods for media mix modeling with carryover and shape effects. 2017.

L. Krishnamurthi and S. P. Raj. The effect of advertising on consumer price sensitivity. *Journal of Marketing Research*, 22(2):119–129, 1985.

R. E. Quandt. Estimating the effectiveness of advertising: some pitfalls in econometric methods. *Journal of Marketing Research*, 1(2):51–60, 1964.

P. E. Rossi and G. M. Allenby. Bayesian statistics and marketing. *Marketing Science*, 22(3): 304–328, 2003.

C. Saranya and G. Manikandan. A study on normalization techniques for privacy preserving data mining. *International Journal of Engineering and Technology (IJET)*, 5(3):2701–2704, 2013.

G. Shieh. Clarifying the role of mean centring in multicollinearity of interaction effects. *British Journal of Mathematical and Statistical Psychology*, 64(3):462–477, 2011.

D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, pages 1–28, 2017.

D. Vakratsas, F. M. Feinberg, F. M. Bass, and G. Kalyanaram. The shape of advertising response functions revisited: A model of dynamic probabilistic thresholds. *Marketing Science*, 23(1): 109–119, 2004.

Y. Wang, Y. Jin, Y. Sun, D. Chan, and J. Koehler. A hierarchical bayesian approach to improve media mix models using category data. 2017.

A Descriptive statistics data

Table 12: Summary statistics of cost data per deepest level.

	count	min	50%	mean	max	std
paid search						
shopping - branded	53	62	5385	7154	33696	8370
text - branded	53	279	1073	1665	4445	1163
text - non branded	53	16837	80815	81515	163429	29858
text - semi branded	53	4464	8002	8109	13007	1470
shopping - non branded	26	0	5744	8800	29329	9761
display						
awareness - brand	39	1656	6170	7398	17580	4327
consideration - brand	53	284	8379	9544	27497	6217
prospecting - performance	53	1883	13033	13723	27584	4221
retargeting - performance	53	250	4188	4273	13493	2681
paid social						
action - performance	50	1672	3947	4392	9574	1941
desire - performance	47	4694	8579	8960	15181	2779
awareness - brand	20	2969	7426	7695	13086	3129
consideration - brand	52	4322	14589	15529	36334	6066

Table 13: Summary statistics of revenue data per deepest level.

	count	min	50%	mean	max	std
paid search						
shopping - branded	53	329	5239	7817	26711	7345
text - branded	53	167700	251615	259094	443738	54158
text - non branded	53	32100	98360	103949	219717	37794
text - semi branded	53	75917	118134	117434	194353	24748
shopping - non branded	26	0	4451	8534	43572	10608
display						
awareness - brand	39	5	487	945	4307	1121
consideration - brand	53	358	2709	6415	20078	6612
prospecting - performance	53	2719	17790	18316	41948	8392
retargeting - performance	53	1006	26203	24655	87435	23273
paid social						
action - performance	50	71	1313	1466	3769	955
desire - performance	47	8	2482	2890	8908	1697
awareness - brand	20	87	585	680	1795	407
consideration - brand	52	249	1752	2011	5788	1328

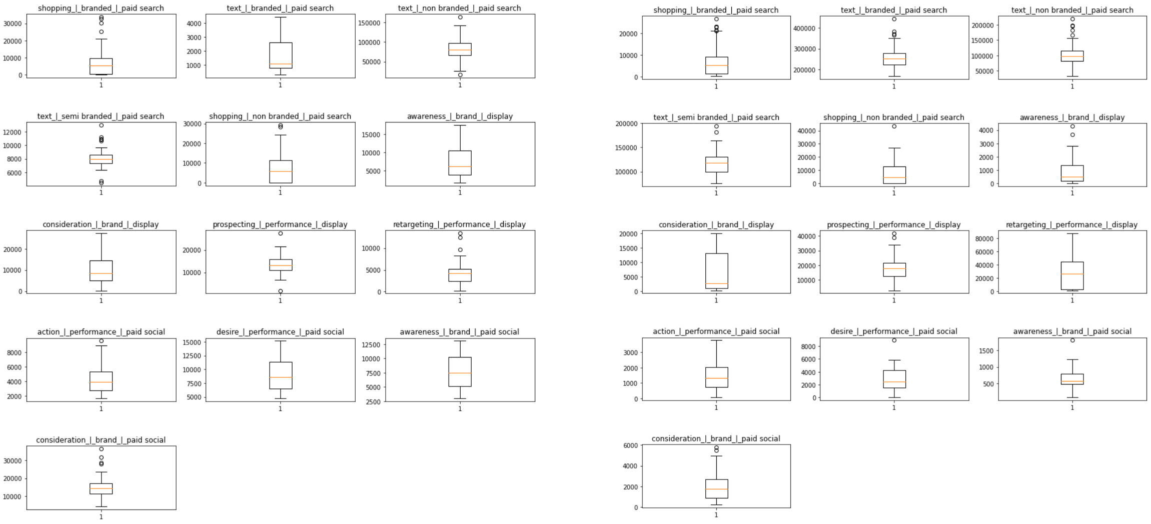

Figure 13: Box plots of the cost data (left) and revenue data (right) for each deepest level.

B Control regression

Table 14 shows the output of the control regression described in Section 3. We have also included the offline marketing channels, otherwise there could be an omitted variable bias. All variables are log-transformed and the explanatory variables are centralized. We conclude that for the paid search and display channel there is positive significant effect on direct attribution. For paid social there is no significant effect.

Table 14: Results of the control regression with direct attribution as dependent variable. Significance codes: "." is p-value less than 0.1, "*" is p-value less than 0.05, "**" is p-value less than 0.01, "***" is p-value less than 0.001.

	coef	SE
Intercept	13.674	0.009***
paid search	0.785	0.092***
display	0.085	0.035*
paid social	-0.030	0.027
affiliate	-0.071	0.043
email	-0.094	0.019***
referral	0.127	0.023***
tv	0.006	0.003
ooh	0.014	0.010
radio	-0.001	0.003
online radio	0.000	0.004

C Results

C.1 Hierarchical Bayesian parameter estimates

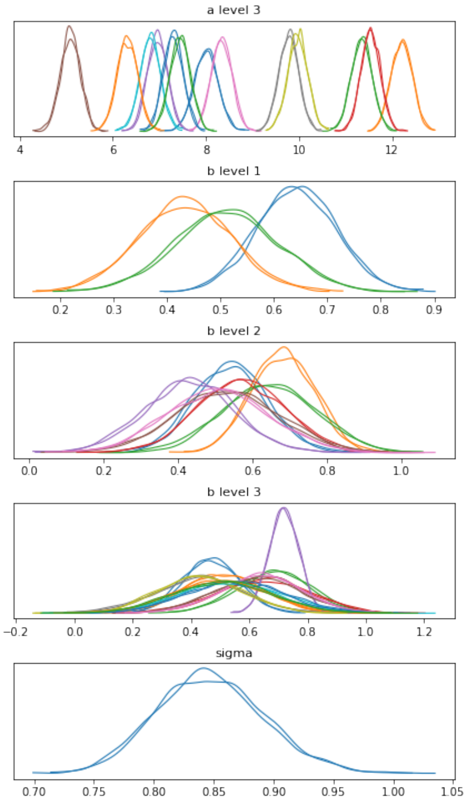

Figure 14: Posterior distributions (left) for each estimated parameter and the sample draws plotted in sequential order (right) of the HB estimation described in Section 5.2. Note that the colors are only to distinguish between different parameters and have no specific meaning.

C.2 Model with interactions

Table 15: Estimated parameters of the ML model and the HB model including interactions, based on the last fifteen active weeks counted from week 45. Note that for each parameter the standardized estimate is reported and $\tilde{\beta}$ is the estimated elasticity, meaning the transformed version as in Equation (3), not β^*, where we leave out the subscripts for simplicity.

		ML		HB	
channel	variable	estimate	sd	mean	sd
shopping - branded - paid search	$\tilde{\alpha}$	7.515	0.308	7.536	0.368
	$\tilde{\beta}$	0.496	0.130	0.565	0.114
	$\tilde{\lambda}$	-0.085	0.055	-0.098	0.057
text - branded - paid search	$\tilde{\alpha}$	12.412	0.049	12.178	0.225
	$\tilde{\beta}$	0.050	0.071	0.561	0.143
	$\tilde{\lambda}$	0.059	0.018	0.013	0.085
text - non branded - paid search	$\tilde{\alpha}$	11.312	0.079	11.308	0.226
	$\tilde{\beta}$	0.480	0.190	0.702	0.131
	$\tilde{\lambda}$	0.147	0.055	0.079	0.122
text - semi branded - paid search	$\tilde{\alpha}$	11.482	0.027	11.536	0.214
	$\tilde{\beta}$	1.182	0.073	0.674	0.159
	$\tilde{\lambda}$	0.059	0.043	0.025	0.169
shopping - non branded - paid search	$\tilde{\alpha}$	6.881	0.274	6.879	0.218
	$\tilde{\beta}$	0.749	0.067	0.739	0.048
	$\tilde{\lambda}$	-0.076	0.033	-0.074	0.024
awareness - brand - display	$\tilde{\alpha}$	5.051	0.271	5.039	0.224
	$\tilde{\beta}$	1.374	0.410	0.658	0.147
	$\tilde{\lambda}$	0.126	0.101	0.084	0.075
consideration - brand - display	$\tilde{\alpha}$	8.316	0.104	8.382	0.219
	$\tilde{\beta}$	1.204	0.117	0.667	0.141
	$\tilde{\lambda}$	0.113	0.043	0.077	0.083
prospecting - performance - display	$\tilde{\alpha}$	9.772	0.095	9.780	0.206
	$\tilde{\beta}$	0.277	0.486	0.432	0.157
	$\tilde{\lambda}$	0.060	0.110	0.047	0.137
retargeting - performance - display	$\tilde{\alpha}$	9.972	0.217	9.970	0.217
	$\tilde{\beta}$	0.427	0.231	0.434	0.141
	$\tilde{\lambda}$	-0.008	0.076	-0.008	0.067
action - performance - paid social	$\tilde{\alpha}$	6.646	0.230	6.733	0.228
	$\tilde{\beta}$	0.001	0.990	0.515	0.169
	$\tilde{\lambda}$	0.283	0.179	0.108	0.114
desire - performance - paid social	$\tilde{\alpha}$	7.377	0.323	7.322	0.211
	$\tilde{\beta}$	0.437	1.380	0.533	0.171
	$\tilde{\lambda}$	-0.984	0.580	-0.286	0.169
awareness - brand - paid social	$\tilde{\alpha}$	6.392	0.102	6.353	0.221
	$\tilde{\beta}$	0.498	0.289	0.507	0.166
	$\tilde{\lambda}$	-0.161	0.101	-0.089	0.136
consideration - brand - paid social	$\tilde{\alpha}$	7.419	0.226	7.436	0.210
	$\tilde{\beta}$	0.663	0.951	0.512	0.167
	$\tilde{\lambda}$	0.363	0.346	0.107	0.166

D Python code

The Python code written for this paper is attached in a separate file. Table of contents:

1. Packages
2. Configurations
3. Data read-in
4. Performance measures
5. Select data for within sample performance
6. ML approach with and without interactions
7. (Non)hierarchical Bayesian approach
8. Out-of-sample performance Bayesian approach
9. Simulation
10. Control regression