STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's thesis in economics
Academic year 2019–2020

Sales Modeling and Local Factor Decomposition for Optimal Investment Decisions in MMM: A Monte Carlo Simulation Study

Daniel Heimgartner (41644)

Abstract: Media Mix Models (MMM) are used to understand drivers behind key performance indicators and to measure the effectiveness of media channels. The key metric to report causal impacts of media investments on sales is return on investment (ROI). The shape of ROI-curves crucially impacts optimal allocation. Different modeling and decomposition approaches are scrutinized in an effort to estimate such response patterns. OLS, SVR, GAM and TVEM models in combination with WFD, ALE and SHAP decomposition are employed. TVEM allows the marketer to pool data, thereby leveraging larger samples despite potential structural changes. The different methodologies are tested and compared in a Monte Carlo study and in a virtual MMM environment: AMSS is a micro-founded demand model which is calibrated to real data. We find that an information criterion can be used as a proxy for the goodness of ROI-curve fit. Additive models should be combined with SHAP whereas ALE produces better estimates for multiplicative models. A bias-variance trade-off can be observed with additive models producing lower bias but higher variance. The multiplicative power model specification yields the most consistent estimates. TVEM is able to slightly reduce the variance of estimates and is not very sensitive to the degree of dynamic change. Strong funnel effects impose a challenge for all approaches. SVR is among the best performing methodologies when media channels are considered individually. GAM is overall the most balanced approach. We believe that our testing environment is a valid tool to be explored in further research.
Keywords: Media Mix Modeling, Shapley Value Regression, Generalized Additive Models, Time-Varying Effect Models, Model Agnostics, Monte Carlo Simulation
JEL: M310, M370

Supervisor: Rickard Sandberg
Date submitted: 18 May 2020
Date examined: 25 May 2020
Discussant: Piotr Józwik
Examiner: Andreea Enache

Contents

List of Figures

List of Tables

Preface

Economics is often criticized for its argumentation being built on simplifying assumptions. The art of explaining reality lies in the right choice between complexity and simplicity. That is to isolate the problem by disentangling reality.

On the other end of the scale, people perceive statistics as a precise science similar to engineering, or physics. But of course, statistics is not free of assumptions either. Additionally, we only observe a finite number of realizations of a given random variable and infer from these observations the underlying ground truth. Yet, certainty does never exist. In this sense, probability is just a concept to deal with imperfect knowledge. If we'd know the ground truth of all processes, probability would become obsolete.

More or less by chance, I ended up in the fields of economics after having intermediately studied Swedish, architecture, mathematics, English and sports. I want to thank my father for always supporting me and my life-choices. After all, the time-span of a mammal being dependent on the parents greatly correlates with the intelligence of a species. Let us for once take correlation for causation and for once don't question statistics!

I would also like to thank Nepa for supporting this thesis. In particular, thank you Goran Dizdarevic and Stefanie Möllberg for your help. Also, Nepa's Data Science team has my gratitude and is for me an example of great Swedish workplace culture.

Lastly, thank you Rickard Sandberg and Ulrich Matter for your supervision.

1 Introduction

Media Mix Models (MMM) are used to understand drivers behind key performance indicators (KPIs) such as sales and to measure effectiveness of different media channels. The aim is to infer the optimal media mix (media allocation) and hence to maximize returns on advertising (Jin et al., 2017). These models are usually based on weekly aggregated data, including sales, price, product distribution, media spend and external factors such as macroeconomic variables, weather data, holidays and others - commonly referred to as the 4Ps (Product, Price, Place, Promotion). To derive valuable insights from MMM, marketers are required to draw causal inference from their models which is usually achieved utilizing linear regression (Jin et al., 2017).

Of course, inferring causation from correlation is non-trivial and as presented in this thesis further complicated in the context of MMM. Reasons for this are manyfold: Data availability of weekly instances usually constrains the sample size which limits traditional causal inference techniques. Media expenditures provide a weak signal compared to other drivers in the 4Ps and are usually highly correlated due to advertisers aligning their media spending with the underlying seasonality of their promotion cycles. Advertisers further resist to greatly vary their spend from historic patterns which further complicates the task of disentangling each media's impact on KPIs (Wang et al., 2017). On the other hand, the randomized experimental benchmark for causal analysis is rarely applied in a marketing context because of high costs associated with such experiments among others. A further complication arises from yet another perspective: Data points are not only few (three years of weekly data results in 154 observations) but also not necessarily from the same data generating process. Markets are not static nor are consumer behavior patterns. From a technical perspective, this introduces a trade-off between data availability and recency and

gives rise to non-constant modeling parameters. Further and most importantly, response patterns are not linear. This is to say that the relation between media investments and sales (returns on investments, ROI) is expected to be concave or S-shaped. Such shape effects in ROI-curves are crucial for optimal media allocation and thus the key target of MMM and at the heart of this thesis. As marketing channels are expected to interact with each other and therefore benefit from potential synergies ROI-curves depend on other channels' spend. Mathematically speaking, ROI-curves become ROI-surfaces. Yet, marketers usually abstract from this perspective and decompose such multi-dimensional patterns to trace out two-dimensional response curves, subsequently used to optimize media allocation decisions. This dimension reduction refers to fairly redistributing synergistic effects to the contributing media channels and different decomposition methods are explored in this thesis.

This ultimately leads to the formulation of our research question: *How can shape effects in MMM be captured most accurately?* Now, the problem is, that both a historical or experimental analysis is not feasible by reasoning as outlined above. But there is a third option available: Simulation. Simulation allows us to generate ground truth in a controlled virtual environment which enables a methodological comparison. Recall, that the research question states "in MMM" which imposes that the simulated data follows patterns specific to a typical marketing environment with all its complexities. Our strategy of achieving this is to calibrate the simulation specification to mimic key metrics from a real-life example. Nevertheless, the above outlined complications need to be well-understood, simplified or abstracted (where legitimate) to grasp reality and reduce its complexity. One of these abstractions can be made with regard to the time-consistency of modeling parameters. The research question is hence applied once to a static and once to a dynamically evolving marketing environment. The results of this analysis should help marketers to decide which methodological approach to choose and how to handle samples over longer time horizons. Marketers are usually exposed to the latter question when firms would like to reevaluate their strategies and update the initial data sample (from the previous analysis) with more recent data.

In essence, a Monte Carlo simulation study is conducted, repeatedly estimating the media response patterns and comparing them to the ground truth. Arguing that the data generating process follows the real one allows us to compare the different methodological approaches to the same benchmark. Such argumentation will of course be strengthened with statistical facts.

The remainder of the text is structured as follows: In section 2 we introduce the reader to the current MMM methodology and highlight certain key simplifying assumptions of this framework. Subsequently, in 3 the current research frontier is outlined. This encompasses challenges and opportunities in MMM, a theoretical perspective on shape effects and literature tackling the problem of structural change in MMM. The literature section concludes by introducing the simulation framework: A micro-founded consumer demand model developed in a recent Google research project. Next, the underlying research strategy is presented in section 4 guiding the applied part of the paper. Section 5 outlines the methodology, namely different modeling and decomposition approaches and the evaluation strategy. Section 6 describes how the simulation specification is calibrated and presents key metrics comparing the simulated data to the real data. Sections 7 and 9 presents and discusses the results. As the simulation specification is not free from assumptions, robustness checks are conducted in 8 changing key parameters of the simulation process. Section 10 concludes.

This structure should alleviate an understanding of MMM in its complexity from the perspective of saturation effects. However, this perspective is not narrow but necessitates an understanding of other interrelated subjects such as interaction effects, multicollinearity, model agnostics, game theory, the Lucas critique, constrained optimization and many more. The aim of this thesis is thus

not only of technical nature but in that respect also an effort to understand and break down a multidimensional problem.[1]

2 A Primer on Media Mix Modeling

This section aims at a generic introduction to Media Mix Modeling as it is applied in the industry. It should give an intuitive understanding of the different steps performed and describes some crucial assumptions.

According to Hanssens et al. (2003, pp. 358), the determination of the optimal mix begins with the specification of a policy preference function. If the goal is to maximize the firm's profits then the optimization could be specified in a illustrative manner by

$$\max_{\mathbf{x}} \Pi = \underbrace{pQ(p, \mathbf{p}^*, \mathbf{x}, L, K, \mathbf{EV})}_{\text{revenue}} - \underbrace{pc_q - \mathbf{c_x x}}_{\text{cost}} \tag{1}$$

where $Q(\cdot)$ is the sales function, reflecting both the market equilibrium (potentially taking into account a competitive market environment) and the production function, p is the product price, $\mathbf{p}^*$ the competitors' price vector, $\mathbf{x} = \{x_m; m = 1, \dots, M\}$ the vector of M distinct marketing variables, L, K stand for the production factors and $\mathbf{EV}$ captures all other environmental variables possibly influencing sales. c_q is the cost of production whereas $\mathbf{c_x}$ represents the vector of marketing costs. Qualitatively, the firm's sales (captured by $Q(\cdot)$) depend on the own and competitors' pricing, but also on the production technology, available factors of production, environmental variables (for example weather or holidays) and marketing efforts.

Now, the here presented MMM approach abstracts from the potentially complicated problem in equation 1 by several assumptions. First, it separates the problem into a purely marketing-related one by writing

$$\max_{\mathbf{x}} \Pi_x = Q_x(\mathbf{x}) - \mathbf{c_x x} \tag{2}$$

where $Q_x(\cdot)$ now stands for the sales driven by marketing efforts. Clearly, $Q_x(\cdot)$ is a multidimensional function reflecting a ROI-surface in the n-dimensional marketing space (where n is the number of interacting marketing channels). For illustration, a ROI-surface is depicted in figure 1 where sales are an increasing function of two media channels. Moreover, the effect is larger if the marketer invests in both channels (synergies).

Secondly, the subsequently presented methodology makes the further abstraction that

$$Q_x(\mathbf{x}) = Q_1(x_1) + Q_2(x_2) + \cdots + Q_M(x_M) \tag{3}$$

This is to say, that the marketing-sales function is basically additively separable which implies that we can think of the ROI-surface as M independent two-dimensional ROI-curves. It is exactly this dimensionality reduction that urges the modeler to fairly decompose the synergistic media effects and attribute a fair share to each channel.

[1] The R scripts can be accessed via `https://github.com/dheimgartner/master-thesis-mmm`

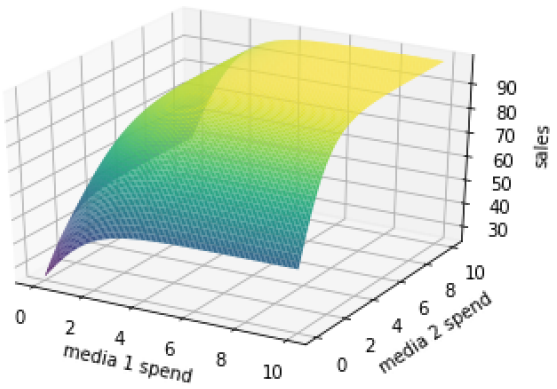

Figure 1: Illustration of a ROI-surface.

The standard first order condition in 2 taking into account 3 implies

$$\frac{\mathrm{d}Q}{\mathrm{d}x_i} = c_{x,i}, \qquad \forall i = 1, \ldots, M \tag{4}$$

The optimality condition shows that the shape of the response curves (ROI-curves) matters for optimal allocation. We should invest in each channel until the marginal benefit equals marginal marketing cost. Or in the spirit of Hanssens et al. (2003), allocation decisions do influence the sales response.

Estimating the ROI-curves in 3 is the crux of MMM and a multistage procedure: First, one should be aware that there is a contemporaneous and lagged effect dimension to the problem. The ROI-curves depict the 'instantaneous' impact of marketing expenditure on sales. Yet, marketing is known to exhibit so-called *carryover* effects which reflect that a single ad exposure makes a customer aware of a brand. This awareness is assumed to diminish over time and finally fade out.

Carryover effects are considered by transforming the time-series of media spend through the *adstock* function (Jin et al., 2017)

$$\text{adstock}(x_{t-L+1,m}, \ldots, x_{t,m}, L) = \frac{\sum_{l=0}^{L-1} w_m(l) x_{t-l,m}}{\sum_{l=0}^{L-1} w_m(l)} \tag{5}$$

where w_m is a non-negative weight function. The cumulative media effect is a weighted average of media spend in the current week and the previous $L-1$ weeks. L is the maximum duration of the carryover effect assumed for a medium m. Different functional forms can be chosen for the weight function w_m but the intuition of equation 5 should be straight forward and is visualised in figure 2 where a single investment is spread out over L time periods.

These two problems (estimating lagged and contemporaneous effects) are not interrelated and can thus be separated. Our study does not further consider carryover effects but assumes that the marketer is able to model them correctly. Hence if we refer to media variables, then these variables are already transformed as described in 5.

After having transformed the data accordingly, the sales process is modeled by some suitable model.

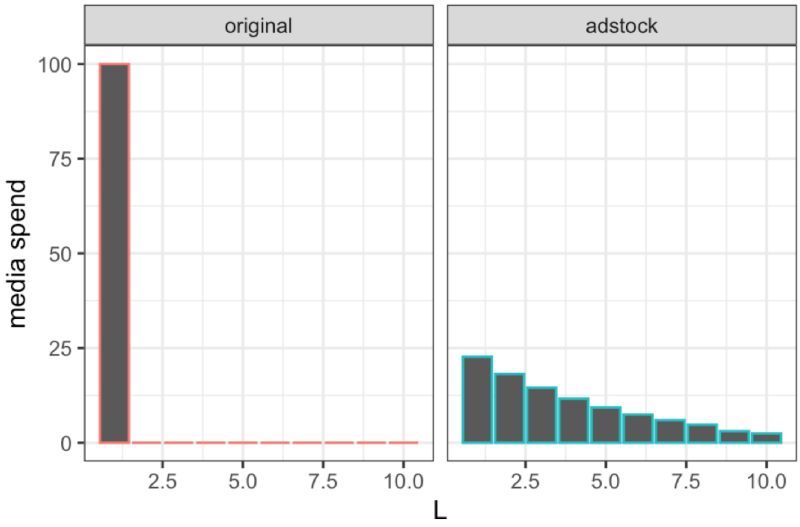

Figure 2: Example of spending in a media channel before and after the adstock transformation.

Importantly, drawing causal inference imposes stringent requirements as later discussed. Once the sales process is modeled, it has to be decomposed such that a sales' contribution for each observed media spend can be recovered. As already mentioned, if the model allows for interaction effects, synergies need to be fairly allocated to the individual channels. This decomposition exercise results in a scatter in the media spend - sales plane. A curve fulfilling some desirable properties fitted to the scatter is the final ROI-curve which is also referred to as response curve. The here presented analysis aims at scrutinizing suitable model and decomposition methodologies.

In summary, MMM is a multistep procedure. Media variables are transformed according to the *adstock* function which reflects the lagged effect, also known as carryover. The shape effect, on the other hand, reflects the curvature of the ROI-curve which is to be understood as a contemporaneous effect. ROI-curves, also referred to as response curves are traced out by decomposing a suitable model, after it has been fitted to the sales data. The MMM process requires simplifying assumptions of which the reader should be aware of.

3 Literature Review

The literature review will be structured in four parts: First, challenges faced in MMM are outlined with the aim to make the reader cautious. Second, relevant theory regarding shape effects in MMM is presented followed by a selection of literature directly proposing and testing solutions to capture media response curves. The third part elaborates on the problem of structural change in the modeling period. Again, we report on more general insights taking up the discussion on why structural change might emerge in a marketing context before diving into theoretical propositions how to actually deal with it. Generally, the relevant literature is extensive as dynamic changes are well understood in the time-series literature (for example Bai and Perron, 2003) and therefore focus lies on a small marketing related subset. Finally, the reader should hopefully understand the need to build a simulated environment in order to test and fairly compare different modeling strategies. Hence a google research project is presented, describing the theory behind the micro-founded model which is later leveraged in our sales and marketing simulation study. As it is to our best knowledge,

a cross-comparison of different methods stands short and this paper is one of the first to address the debated issues by deriving insights from the Aggregated Marketing System Simulator (AMSS).

3.1 Challenges and Opportunities in Media Mix Modeling

Chan and Perry (2017) note that MMM is concerned with one of the most demanding problems in applied statistics, namely causal inference. Recall, that MMM is all about determining the causal impact of ad spend on sales at any given level. To answer causal questions randomized experiments are the gold standard.[2] Because such experiments are very expensive in a marketing setting marketers turn to historical data in order to trace out causal relations.

The authors discuss this approach with the help of the Rubin causal model for causal inference which describes the difficulties of approximating the difference between potential and counterfactual outcomes (causal impact). This estimated causal impact is prone to a so-called 'selection bias' which refers to any biases in the treatment selection mechanism that are also correlated with the outcome (sales). In the context of MMM this means potentially everything that determines both ad spend and sales. Chan and Perry (2017) note that the matching estimator could alleviate the problem of selection bias but is most often infeasible in MMM due to data limitations. They therefore propose to turn to regression models. Regression models, on the other hand, are only trustworthy if there are enough data points, if there is useful and independent variability in the features and if all confounding variables are included (correct model specification).

Some of the above-mentioned prerequisites will now be immersed because they are later encountered in the text. If flexible functional response forms are required (to capture non-linear response curves) the regression specification exhibits many parameters. Together with limited data availability, the usual prerequisite of 7-10 data points per parameter falls short in MMM.

Further, when fitting a model to highly correlated input variables (multicollinearity) the coefficient estimates depict a high variance. As seen in figure 3 both estimated response surfaces fit the sales data well despite having very different slopes. The slopes correspond to the rate of change of the linear response curve in this case. The multicollinearity issue might therefore directly impact the ROI-curve estimates.

Yet another problem can be referred to as extrapolation bias which describes the difficulty to trace out the true response curve when only a limited range of ad expenditures is observed. The problem is illustrated in figure 4. The simulation approach allows, that the whole range of relevant ad spend is realized. This is to say that the simulated data stretches over the relevant range, similar to observing the whole scatter in figure 4. Nevertheless, in practice, this might not be the case and difficult to scrutinize.

To correctly estimate the response surface (see 3) the statistician not only needs to disentangle multicollinearity but also pay attention to so-called 'funnel effects'. An example of such an effect is a TV campaign driving more related queries, which in turn increases the volume of paid search ads. When an ad channel also impacts the level of another ad channel (funnel effect) but the model specification does not correctly account for such interactions, the resulting response curves are biased (Chan and Perry, 2017).

Another crucial point is related to model selection. Usually, model selection is based on some infor-

[2]For one of the rare studies concerned with the experimental approach in MMM the reader can refer to Lewis and Rao (2015).

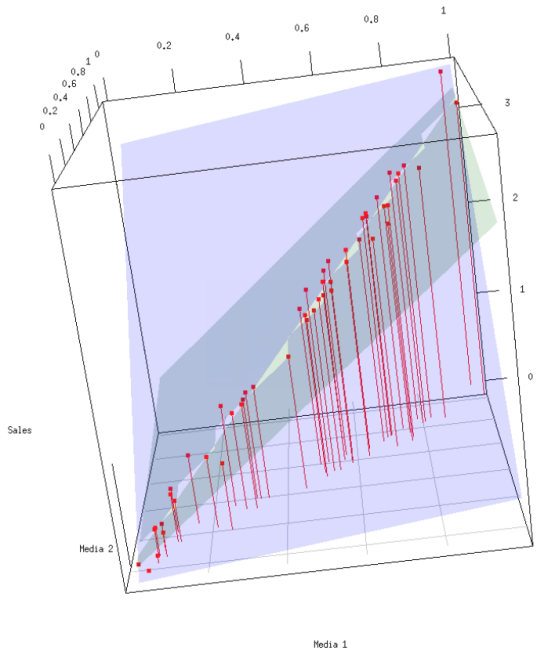

Figure 3: Multicollinearity yields two estimated response surfaces which lead to very different response curves, Source: Chan and Perry (2017).

mation criterion (such as R2). Chan and Perry (2017) question the validity of such an information criteria (IC) because it does not necessarily imply the best ROI-curve fit. This is because media variables contribute a small signal and hence don't contribute much to the IC. The quality of the R2 proxy can be gauged in the simulation analysis.

The authors propose three broad areas constituting opportunities in MMM: 1) Better data 2) Better models 3) Model evaluation through simulations. This text regards all of these three points. The first one by scrutinizing the potential of pooling data across time which increases the sample size. Additionally, the reader will encounter the here presented challenges explicitly or implicitly throughout the remainder of the text. It will also alleviate the discussion towards the end of the paper trying to derive valuable practical insights.

MMM is also criticised for only considering the short-run sales-driving capacity of marketing. Long-run brand-building properties are usually ignored. Such a more holistic point of view could be approached by for example leveraging Cointegration and Error Correction Models as proposed by Cain (2008). It is though yet another challenge to capture shape effects in such a setting. The here presented analysis focuses on the short-run perspective only and abstracts from potential brand-building capacities.

3.2 Response Curves and Shape Effects

This section introduces prominent functional forms in response patterns and presents the findings of relevant modeling literature. It is very important to understand that ROI-curves are implied response functions by the modeling approach and thus model contingent.

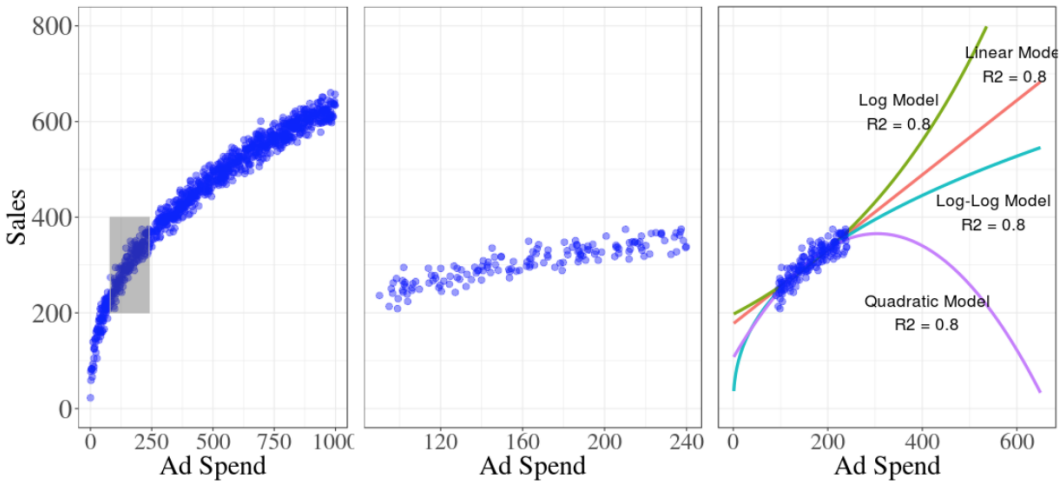

Figure 4: Potential extrapolation bias when models are fitted to a limited range of data, Source: Chan and Perry (2017).

Possible properties of a sales response function include what happens to sales when marketing expenditure is close to zero or very large. Additionally, the rate of change in sales as marketing activities increase can be analysed using the concept of increasing returns, decreasing returns to scale or some sort of threshold effect like a minimum advertising investment (Hanssens et al., 2003, pp. 94). This should make it obvious, that the discussion evolves around marketing elasticities which describe what happens percentage-wise if the marketing effort is marginally increased (by 1%). As seen in the previous section 3.1, there are potential funnel effects (synergies) between media channels, which makes the ROI-curves surfaces rather than two-dimensional curves (compare with figure 3). The following discussion refers to the media spend - sales plane and assumes thus holding other media investments constant.

Hanssens et al. (2003, pp. 95) argue that sales response functions are generally concave and only in a few instances S-shaped. Further, if the marketing driver has a relatively limited scale (compare figure 4), a linear approximation may be chosen. Importantly, linear approximations potentially lead to large extrapolation biases. This is because optimal allocation decisions attribute all the media spend to the channel with the highest ROI (the slope of the linear curve). The reader is encouraged to visualize this scenario with the help of figure 4. Concave response curves are linked to diminishing returns to scale: Each additional unit of marketing effort brings less in incremental sales. Hanssens et al. (2003, pp. 103) state that empirical evidence favors such a behavior. S-shaped response patterns (convex-concave functions) might arise if marketing efforts are characterized by threshold effects. Tellis (2006) argues that S-shaped ROI-curves are the most plausible because, at some very low level, advertising might not be effective at all but gets drowned out in the noise. Further, it implies that elasticities depict an inverted bell-shaped pattern in the level of advertising which is the most appealing form (linking 'no effect' at zero and some saturation point with a 'positive effect' in between). Different product categories might depict different sale response functions. The three potential response functional forms are presented in figure 5. One pattern that scholars and marketers agree upon is saturation (Hanssens et al., 2003, pp. 111). Saturation implies that no matter how much marketing effort is expended, sales won't react to that stimulus.

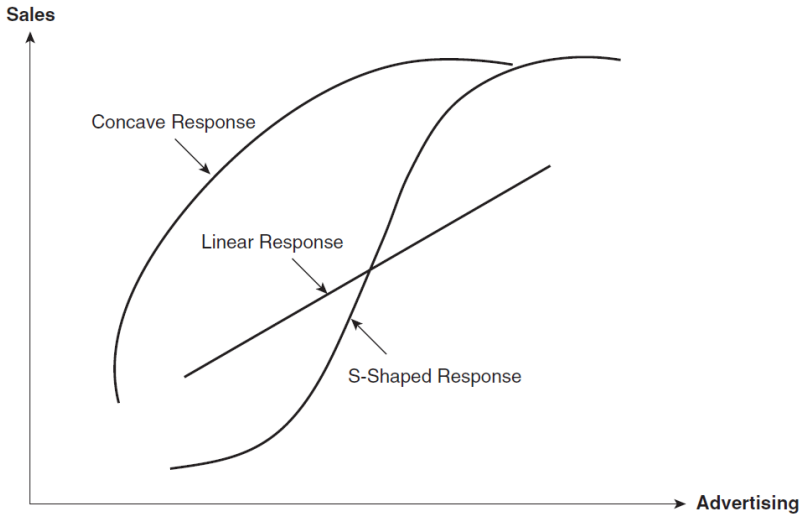

Figure 5: Potential media response patterns, Source: Tellis (2006).

Such is the case if buyers become insensitive or have binding budget constraints.

The reader is now introduced to the relevant literature concerned with capturing shape effects in MMM. Jin et al. (2017) propose a media mix regression model capable of capturing both carryover and shape effects of advertising at the same time. As the model is no longer linear in parameters estimation is nontrivial and achieved by means of a Bayesian approach. The Bayesian method additionally allows that prior knowledge can be incorporated into the model. More precisely, they use an adstock transformation (compare equation 5) to capture the carryover effects and model response patterns with the help of the (beta) Hill function ($\beta Hill$). The authors realize that the $\beta Hill$ function is very poorly identifiable which makes it challenging to estimate the parameters well with any statistical method. As a consequence, they propose investigating estimation with the help of regression splines (related to the estimation of Generalized Additive Models which will be presented in section 5.1.3). The model specification also abstracts from synergy effects by choosing the marketing channels to enter additively in the regression equation.

Jin et al. (2017) test their model both with real and simulated data. The simulation specification is of the exact same nature as the model specification which is a classical Monte Carlo approach. Additionally, the variance of sales explained by media is chosen to be higher than encountered in real-life scenarios which comes with the consequence of an unrealistic strong signal to noise ratios (our approach tries to correct for such unrealistic settings by calibrating the simulated data). Also, the authors propose to resample the simulated data and generate 500 datasets because of the randomness in the simulation process. They evaluate their findings both on a realistically small sample size and a large one.

The authors mention that estimating the coefficients of the $\beta Hill$ function yields low bias but high variability in the large sample whereas even the bias is high for the small sample. More precisely ROI-curves are downward shifted. Jin et al. (2017) argue that the bias was introduced by wrongly specified priors. This makes the Bayesian approach very sensitive to prior beliefs. The study by Jin et al. (2017) has many parallels with our approach from a general viewpoint as will become evident at later stages.

The same group of authors published a closely related paper (Sun et al., 2017) in which they tackle the problem of limited data availability. MMM data is usually aggregated at the national level but could be leveraged at a finer granularity by a Geo-level Bayesian hierarchical media mix model (GBHMMM). The authors follow the same approach as outlined above and compare the GBHMMM to the aggregated counterpart, which is basically the model presented in Jin et al. (2017). Indeed, there is a reduction in error due to having more observations and useful variability in media spend. This insight is more of practical value and is closely related to the extrapolation bias and multicollinearity issue introduced in section 3.1. An interesting note is that when bringing the GBHMMM to real data 'TV' exhibits concave returns to scale whereas 'Search' (online marketing) depicts an S-shaped response pattern. Both of the presented papers propose the usage of the Aggregate Marketing System Simulator (introduced in 3.4) as a simulation tool.

Wang et al. (2017) is yet another paper concerned with the data availability problem. Usual modeling windows consist of about 50 to 250 observations. Given the need to specify the model extensively in order to draw causal conclusions and given that each media variable has at least three parameters to be estimated, there are only a few observations per parameter. But even if a longer time horizon would be available to fit the model such is not desirable as the market dynamics could have shifted drastically. This is the classical trade-off between data availability and relevancy and is further explored in section 3.3. Where the previously mentioned paper addresses this issue by pooling data according to geographic entities Wang et al. (2017) suggest pooling datasets from multiple brands within a product category. Again, they adopt the Bayesian hierarchical framework presented in Jin et al. (2017). This approach also alleviates the multicollinearity issue because media expenditures are not expected to be highly correlated across independent brands. In a Monte Carlo approach, the authors compare their pooled strategy to the single brand model under different simulation scenarios and considering different priors (informative versus weak). Compared to Jin et al. (2017) the authors note, that the informative prior narrows the confidence interval for the respective response curves and the parameter accuracy considerably.

Liu et al. (2014) do a very similar exercise as Jin et al. (2017) but don't apply a Bayesian framework. Also, they specify the decay transformation (to estimate carryover effects) as a Gaussian convoluted exponential decay and leverage the Gamma (instead of Hill) function to capture saturation effects. This identification strategy leaves them with 6 parameters per media channel. The estimation feasibility is proven in a Monte Carlo study.

There are several commonalities shared by all the presented papers: First, they abstract from synergies and hence assume an additive model which second, makes the decomposition straight forward as response curves are directly estimated (for example the $\beta Hill$ function is also the ROI-curve). Third, they follow a Monte Carlo simulation setting and force the world to behave like assumed. From this observation, a common simulation framework might be very valuable to cross-validate different modeling strategies. Synergistic effects make it difficult to trace out response curves (as estimated surfaces have to be transformed into two dimensions). General decomposition methods, applicable to a wide range of models are needed.

3.3 Modeling Structural Change

There are several reasons why a marketer might assume that marketing channels vary in efficiency over time: Certainly, a marketing campaign for ice cream has a different impact on sales when lanced in winter compared to warm summer months. Of course, one could simply include a seasonality interaction in the model to account for this possibility. Still, it is likely, that the whole structure

of how the marketing channels behave is evolving caused by for example the customer's changing mindset. Also, the very nature of ad channels is fast-moving (for example, influencer marketing is depending on the person's popularity which could melt like ice cream in the sunshine). On the other hand, some underlying relations might be stable over time. From this perspective, allowing for structural changes in the modeling approach could yield more accurate ROI-curves.

In the MMM context, it is frequent that businesses would like to reevaluate their marketing efficiencies given the actual data availability. If the whole economic relation is assumed to have changed from the first MMM period to the now actual one, only the most recent data points should be considered. Contrary, if the relation was fundamentally stable, fitting the model to the pooled dataset is preferred, yielding better estimates because of the more data points. Hence, there is a potential trade-off between data availability and relevancy as already mentioned in the previous section. On the other hand, if some fundamental sales' drivers (such as seasonality, holidays, etc.) are time-constant whereas marketing campaigns are bound to vary over time, a time-varying parameter model would allow to leverage the pooled data and still account for potential dynamic shifts. It is the aim of this thesis to shed light on the above intuitive problem description from a more technical perspective.

As noted by Pauwels et al. (2004) and as introduced in section 3.1 there is yet the trade-off between the endogeneity bias ('more variables') and data points per parameter ('fewer variables'). With this in mind, data cardinality becomes crucial. The procedure suggested in the previous section was to pool data across geographical or business unit entities. In this respect, the potential to pool data across time is explored here.

Pauwels et al. (2004) mention that if the hypothesis of parameter constancy is rejected, one could alleviate the problem by formulating a time-varying parameter model. However, the authors state that the reduced form estimation makes the model suspect to the Lucas critique which will be explored at a later stage. Therefore marketers could circumvent this critique by focusing on impulse-response function analysis. On the other hand, this allows only for a marginal analysis and does hence not yield ROI-curves as an output. Generally and as becomes evident when reading Pauwels et al. (2004), the problem of changing dynamics is widely acknowledged by scholars but mostly approached from an inter-temporal perspective (and thus more concerned with carryover effects) whereas we are interested in evolving contemporaneous effects. Concerning this perspective, the authors note that even in the absence of structural breaks advertising effectiveness declines over the life cycle of a product. The authors conclude that important aspects of marketing effectiveness indeed are time or occasion dependent which opens a new set of research opportunities.

The Lucas critique, in a nutshell, states that economic agents are forward-looking and anticipate policy changes. As a consequence, the past policy interventions might alter the very economic relation by updating expectations. Van Heerde et al. (2005) claim that this problem threatens the validity of marketing models because they are backward- rather than forward-looking. For example, certain products are regularly sold at a discount. Once consumers anticipate this discount cycle, they stop purchasing the product at regular prices which seemingly boosts the marketing intervention. But truth is, that the discount campaign cannibalizes regular sales. The Lucas critique might only apply to certain marketing campaigns fulfilling some prerequisites. Wrongly ignoring the Lucas critique yields biased predictions of the effects of marketing policy changes. The key outcome of the Lucas critique is that response parameters change as a function of policy changes. Time-varying parameter models directly address the core of this problem.

Pauwels and Hanssens (2007) acknowledge that current market-response research does not offer

a framework to either identify performance regime changes or to isolate their causes. Instead, performance and marketing spending are either classified as evolving or stationary over the full data period. The authors investigate how marketing actions impact performance regime changes. Assessing regimes is achieved by time-varying parameter models among others. The time-varying model allows to directly assess marketing effects in a single stage without the need to identify switching regimes (join points). Their conclusion is twofold: First, even in mature markets, performance stability is not the only observed business scenario but markets behave as *punctuated equilibrium*. Second, marketing actions play an important role in influencing these performance regimes and inducing a switch from one to the other equilibrium.

Tucci (1995) give an overview of classical time-varying parameter models and group the approaches into *systematic changing* or *stochastically evolving* coefficients. The first idea assumes that parameters change discontinuously at certain points in time whereas the latter allows coefficients to evolve in a random way which can be either stationary or non-stationary. In the first case (also known as switching regimes) the modeler simply includes a dummy variable discriminating between the regimes. This approach can be referred to as 'recency split' in the marketing context. Yet another important aspect is to identify the joinpoint (breakpoint) correctly and the fashion of how the economic relation switches between the regimes. Statistical test procedures exist and in the later simulation study it is assumed that the joinpoint was correctly identified (or is known).

The other way of modeling time-varying parameters is by considering parameters as random variables with different realization in each time period. This random variable can either be stationary, non-stationary or non-stochastic. In either case, the modeling approach follows a multi equation specification where now a transition equation for the parameters (contingent on one of the three cases) is defined. Demidenko and Mandel (2005) applies a random coefficient model to trace out linear ROI-curves and finds that the ad efficiency is very different compared to the efficiency derived by regular OLS. Also, the random coefficient model demonstrates higher predictability.

Hastie and Tibshirani (1993) introduce yet another possibility to characterize varying-coefficient models which does not require a transition equation and not a fundamental statement about the very nature of change. In fact, the only assumption is that the time-varying coefficients evolve in smoothly over time. Given all the identifying difficulties in the previously mentioned approaches, we will follow this elegant solution and refer to the proposed methodology as Time-Varying Effect Model (TVEM).

Greene (2014) claims to be the first (and to our knowledge only) study to leverage TVEM to measure the effectiveness of media mix elements in a given industry. Most likely, the model is misspecified (as it assumes an additive relation with rather randomly chosen interaction terms and transformations) which leaves prices having a positive effect on sales. The author comments on the dynamic efficiency changes in marketing channels by considering the coefficient functions.

In summary, there are two critical trade-offs faced in the practical implementation of MMM: First, the endogeneity bias and the limited number of observations which results in few data points per parameter and second, the trade-off between recency and relevancy. The latter is caused by changes in market dynamics. Such changes might be either caused by simply evolving efficiencies of certain marketing channels or more economically, by forward-looking agents which is known as the Lucas critique. Empirical evidence hints that even stationary markets are subject to performance regime changes. One prominent suggestion to tackle the problem (no matter the cause) is to consider some sort of time-varying parameter model. These models come in different specifications and usually require the definition of a transition equation which characterizes the evolution of the

model parameters separately. Such models are applied in the marketing context but mainly from a marginal perspective and concerning the inter-temporal (not contemporaneous) evolution. This in turn implies, that the connection to shape effects in the time-varying modeling framework has not yet been made. The most suitable approach for the analysis is TVEM which does not require prior knowledge of parameter dynamics and does allow for non-linear response functions given suitable model specification.

3.4 An Introduction to the Aggregate Marketing System Simulator (AMSS)

Vaver and Zhang (2017) realized that new capabilities are needed to evaluate different measurement methodologies in the context of MMM. They further claim that simulation can be an essential tool for evaluating and comparing analysis options. Therefore they developed in a google research project the Aggregate Marketing System Simulator (AMSS) capable of generating aggregate-level time series data related to marketing measurement and ground truth for marketing performance metrics. The capabilities provided by AMSS create a foundation for evaluating and improving measurement methods such as MMM.

The need for simulation arises among others because randomized experiments are rare because, on the one hand, they are expensive and on the other hand require a complex experimental design to accurately measure small effects of marketing. The authors further elaborate on the limitations of drawing causal conclusions from historical data. Vaver and Zhang (2017) argue that such conclusions would require modeling assumptions concerning the nature of the marketing environment (e.g. how advertising changes user behavior, how ad channels interact, how pricing impacts sales, etc.).

Both experimental and observational methods require evaluation and validation which further enhances the importance of a ground truth against which the accuracy of estimates can be verified. The possibility of simulating data and running virtual experiments allows modelers to explore statistical issues, verify model performance and compare competing models (Vaver and Zhang, 2017). This section aims to introduce the data generation methodology of the AMSS. The underlying statistical assumptions should be made available to the reader such that he understands our parameter choice at a later stage.

The general structure of the AMSS is intuitive: AMSS splits the consumer population into different segments, each segment representing a unique consumer mindset towards the market and the brand. The mindset is influenced by natural forces (such as seasonality) and controllable forces (such as marketing interventions). These forces drive migration between the segments according to specified probability distributions. Different segments exhibit different purchase likelihoods. AMSS keeps track of all the migration processes, computes and aggregates sales and is able to generate ground truth by running counterfactual scenarios in the virtual environment.

The basic simulation structure is depicted in figure 6 in a simplified manner: The simulation is characterized by time-ordered events. The ordering of the sequential events matters because each event impacts all subsequent events through changes in the segmentation. Hence, this is the understanding of interactions and synergies. For example, if a TV ad increases favorability towards the brand of a segment and the segment is exposed to say a social media marketing campaign, then this market campaign is more effective because it profits from the more favorable inclined segment. Synergies are thus directional and can be controlled. Each event updates the population segmentation according to a specified probability distribution. Some events generate observable

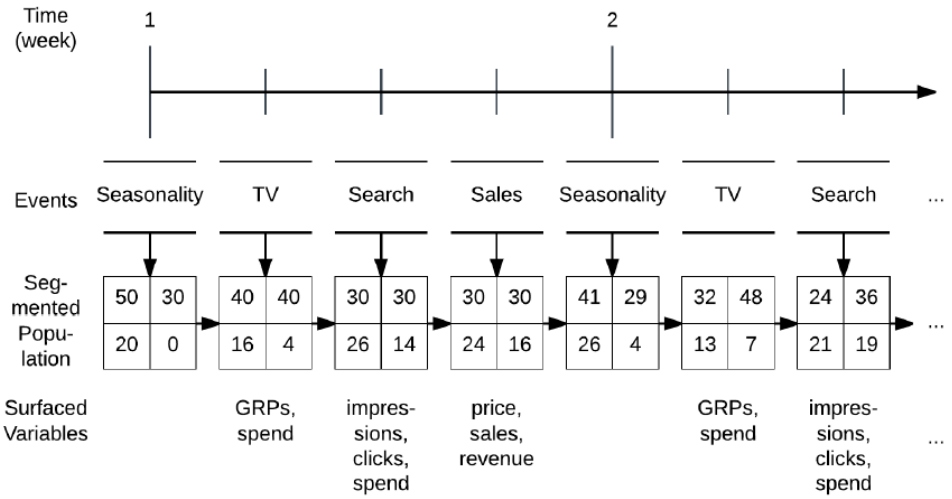

Figure 6: Overview of the AMSS structure, Source: Vaver and Zhang (2017).

State type l	**Potential values** $\mathcal{S}_l$
Category	
Market ($l = 1$)	in-market, out-of-market
Satiation ($l = 2$)	satiated, unsatiated
Activity ($l = 3$)	inactive, exploratory, purchase
Brand	
Favorability ($l = 4$)	unaware, unfavorable, neutral, somewhat favorable, favorable
Loyalty ($l = 5$)	switcher, loyal, competitor-loyal
Availability ($l = 6$)	low, average, high

Table 1: Dimensions of the population segmentation, Source: Vaver and Zhang (2017).

outcomes (surfaced variables). The last event of each time period is the sales event which derives advertiser's sales according to the final population segmentation and a defined demand schedule. This concludes the elaborations on why we need AMSS and its intuition. We will further explain how the population is segmented and in what manner specific events cause population migration.

Consumer mindset. AMSS conceptualizes the consumer mindset along six dimensions. There are three *category states* which characterize the market segment and three *brand states*. A given segment is denoted by the vector $\mathbf{s} = (s_1, s_2, s_3, s_4, s_5, s_6)^T$. Let $\mathcal{S}_l$ be the set of values the consumer mindset may take in the l-th dimension, so that $s_l \in \mathcal{S}_l \ \ \forall l \in \{1, \dots 6\}$. Table 1 provides an overview of the six dimensions and each potential qualitative value the respective dimension might take. A short description follows.

Market state. The market state constitutes the pool of potential customers for the category. Consumers who are 'out-of-market' don't make purchases. Changes are caused by natural migration (for example by seasonal fluctuations or more general trends such as the adaptation of a technological product). Market state can not be influenced by marketing efforts.

Satiation state. The satiation state refers to whether or not a person's demand has been satisfied by a past purchase. 'Satiated' individuals might become 'un-satiated' over time. A satiated individual

can not be driven to purchase by marketing activities.

Activity state. The activity state tracks the location along the path to purchase. Different activity states have different media consumption behaviors, different responses to marketing and different purchase behaviors. Consumers need to reach the 'purchase' state for the advertiser to make a sale.

Favorability state. The favorability state measures the opinion of the brand and thereby influences the purchase likelihood. For example, it is much more likely that a 'favorable' customer makes a purchase compared to an 'unfavorable' individual. Still, the unfavorable inclined individual might purchase the good because the brand is readily available.

Loyalty state. Consumers can be loyal to the advertiser, loyal to its competitor or have divided loyalties.

Availability state. The availability state describes how physically (or mentally) easy it is for a customer to make a brand purchase.

The set of all segments is represented by $\mathcal{S}$ and is a subset of the Cartesian product $\mathcal{S}_1 \times \cdots \times \mathcal{S}_6$. It is a subset because some state combinations are not possible (for example 'satiated' individuals can not have a 'purchase' intent). The restriction rules leave us with 198 different segments.

General migration. The general notation of how population migration is caused by some event $k = 1, \ldots, K$ is now introduced. The general notation alleviates an understanding of some peculiarities of marketing and sales events. Each event k is applied once within each time interval $t \in 1, \ldots, T$. Let the size of the population assigned to a segment $\mathbf{s} \in \mathcal{S}$ before the k-th event of time interval t be $n_{t,k,\mathbf{s}}$. Segments can be grouped: for $\mathcal{A} \subseteq \mathcal{S}$, $n_{t,k,\mathcal{A}} = \sum_{\mathbf{s}\in\mathcal{A}} n_{t,k,\mathbf{s}}$. The overall segmentation of the population at time t is denoted by the vector $\mathbf{n}_{t,k}$. The k-th event updates the population from $\mathbf{n}_{t,k}$ to $\mathbf{n}_{t,k+1}$ which reflects the change in the consumer mindset for each population segment.

The probabilistic consumer migration requires on the one hand a notation that pins down the migration process and on the other hand a way to control the migration probabilities. The modeler needs to define a sequence of transition matrices for each event k. These transition matrices characterize the segmentation update from $\mathbf{n}_{t,k}$ to $\mathbf{n}_{t,k+1}$. The k-th event of time t affects a subset of the population of size $\mathbf{a}_{t,k}$. Affected individuals migrate from population segment $\mathbf{s}$ to $\mathbf{s}'$ according to the transition matrix $Q^{(t,k)} = (q^{(t,k)}_{\mathbf{s},\mathbf{s}'})_{\mathcal{S}\times\mathcal{S}}$, where $q^{(t,k)}_{\mathbf{s},\mathbf{s}'}$ describes the probability of migrating from segment $\mathbf{s}$ to $\mathbf{s}'$. Recall that there are potentially 198 affected population segments which leave the transition matrix very high dimensional (198×198 in case of an event affecting all segments)! These are all the relevant ingredients to characterize the migration of individuals during the k-th event of the t-th time interval

$$\mathbf{m}_{t,k,\mathbf{s}} = (m_{t,k,\mathbf{s},\mathbf{s}'})_{\mathbf{s}'\in\mathcal{S}} \sim \text{Multinomial}\left(a_{t,k,\mathbf{s}}, (q^{(t,k)}_{\mathbf{s},\mathbf{s}'})_{\mathbf{s}'\in\mathcal{S}}\right) \tag{6}$$

where $m_{t,k,\mathbf{s},\mathbf{s}'}$ is the number of people migrating from segment $\mathbf{s}$ to $\mathbf{s}'$. Finally, the updated population segmentation is defined as

$$n_{t,k+1,\mathbf{s}'} = n_{t,k,\mathbf{s}'} - a_{t,k,\mathbf{s}'} + \sum_{s\in\mathcal{S}} m_{t,k,\mathbf{s},\mathbf{s}'}, \qquad \mathbf{s}' \in \mathcal{S} \tag{7}$$

So, there are a number of people in each population segment, some of which are affected by event k. The affected population migrates according to a multinomial distribution which is characterized by

the entries of the transition matrix. Equation 6 and 7 constitute really the core of AMSS. It is most intuitive when thinking of an event as a function that takes the current population segmentation as input and returns an updated population segmentation, along with related observable variables. The remainder of the section explains peculiarities of some events which are worth to understand to fully grasp the simulation specification in section 6.

Market size. The first event of each time interval determines the market size. Market size changes naturally over time as people move in and out of the market. This allows the modeler to specify seasonal changes or other drivers of fluctuations in market size (such as holidays). The modeler defines an 'in-market' target rate which pins down the number of individuals across all population segments with market state equal 'in-market'. Further, this target rate is multiplicatively composed according to

$$\rho_t = \rho_t^{(seas)} \rho_t^{(trend)} \tag{8}$$

where ρ_t is the in-market target rate, $\rho_t^{(seas)}$ is a seasonal trend and $\rho_t^{(trend)}$ an overriding, more general trend.

Natural migration. For each dimension potentially affected by marketing events, a natural transition matrix has to be specified which represents equilibrium values. As marketing events won't have a lagged effect in our case, the population segmentation will immediately jump to that equilibrium in the subsequent period and new marketing efforts impact again on top of that equilibrium population segmentation.[3]

Marketing interventions. Generally, marketing interventions drive population segments from less favorable to more favorable states. Marketing can drive changes in dimensions $l \in \{3, \dots 6\}$ which is to say it can neither influence market state ('has not the power to increase market size') nor satiation state ('can not force already satiated individuals to consume even more'). The media channel is controlled via four components: *audience*, *spend*, *volume* and *effect*. These components are now discussed sequentially.

The *audience* size $a^*_{t,k,s}$ is the population that interacts with the channel and is defined by a population segment's *reachability*. So, the modeler specifies the reachability likelihood $\pi_s^{(a)}$ of each population segment which further pins down the audience size via a Binomial distribution. Part two of this thesis leverages the freedom to choose $\pi_s^{(a)}$ in order to control the media efficiency and impose structural change.

The component media *spend* needs no further explanation than that it characterizes the spending pattern of a given marketing campaign (also known as flighting pattern). It is more important to understand that media volume $v_{t,k,s}$ is controlled via a constant cost function describing the cost per media exposure. The audience size impacts the media reach $a_{t,k,s}$ which in turn together with media volume pins down the average number of ad exposure referred to as frequency $f_{t,k,s}$.

The final component is *effect* which updates the population segmentation. Here all the notation comes together: The amount of migration in the population segmentation depends on both reach and frequency. The former determines the number of people with the potential to migrate and the latter influences the migration probabilities together with the respective transition matrix $Q^{(k)}$. The transition matrix specifies maximal migration probabilities $Q^{(k)}_{\mathbf{s},\mathbf{s}'}$. Recall that there are 198

[3] In case the modeler adds lagged effects these natural migration matrices also determine the speed of convergence towards the equilibrium.

possible population segments and consumers from each segment $\mathbf{s}$ may migrate to each of the other segments $\mathbf{s}'$. The AMSS has to keep track of all probabilities in the high dimensional matrix $Q^{(k)}$. Therefore migration in each dimension $l \in \{1, \dots, 6\}$ happens independently. One is left to pin down transition matrices for each dimension and hence specify the low dimensional matrices

$$Q^{(k,l)} = \left(q_{i,j}^{(k,l)}\right)_{\mathcal{S}_l \times \mathcal{S}_l} \tag{9}$$

For example, transitions in activity state are determined by

$$Q^{(k,3)} = (q_{i,j}^{(k,l)})_{3\times 3} \Rightarrow Q_{1,3}^{(k,3)} = Q_{1,\text{'inactive'}}^{(k,3)} = (0.6, 0.3, 0.1) \tag{10}$$

where the right-hand side after the arrow represents the first row of the transition matrix for the third dimension which reads a 30% chance to convert an 'inactive' individual to 'exploratory' and a 10% chance to migrate to a 'purchase' segment. As seen, these matrices are much more intuitive to define and fully determine all the possible migration probabilities. For an exact understanding of how this step is achieved, the appendix C in Vaver and Zhang (2017) may be consulted.

Recall that $Q_{\mathbf{s},\mathbf{s}'}^{(k)}$ are maximal probabilities. Maximal probabilities are scaled against the frequency $f_{t,k,\mathbf{s}}$ according to a sigmoid function (Hill function) which can be utilized to simulate concave or S-shaped response curves by parameterising this function accordingly.

Sales event. The sales event is the final event of each time period and computes the market-clearing sales quantity according to the final population segmentation and its likelihood to purchase. The sales event also drives population migration because the very act of purchase can change the consumer mindset towards the product. Sales per segment are controlled via linear demand schedules in the price - likelihood space

$$r_{t,\mathbf{s}} = (\alpha_{\mathbf{s}} - \beta_{\mathbf{s}} p_t) \tag{11}$$

where $r_{t,\mathbf{s}}$ is the purchase likelihood of a consumer in segment $\mathbf{s}$ at time t, α_s and β_s are the segment-specific demand intercept and slope. This concludes one full sales cycle in t and we move to the next interval $t+1$ to start again from the equilibrium relation. To summarise, AMSS keeps track of population migration (according to 6 and 7) caused by different events which are characterized by transition matrices or probability distributions along each dimension.

Ground truth. Ground truth is obtained empirically by generating multiple random instances of data. For example, to generate ground truth for media contribution, media spend is set counterfactually to zero in that particular time interval t. By comparing the counterfactual scenario to the actual one finitely often, the estimated contribution converges to the true contribution by the law of large numbers. We use the package **amss** for the implementation in R.

4 Research Strategy

In this section we'd like to provide insight into how we try to tackle the research questions exactly. The reader should subsequently have a guideline that leads through the remainder of the text. Also, we aim to emphasize that our approach is only one perspective on the questions posed in the preceding sections. As always, when asking a generic question, the answer will be provided

in a simplified setting by making necessary assumptions. The reader should be aware of these assumptions.

As introduced, simulating data generates ground truth. Ground truth is not equal to general truth in the sense that it does not necessarily reflect the real world but is, of course, depending on the simulation specification. The exact simulation specification can be found in section 6 but some assumptions are mentioned at this stage already: Shape effects are scrutinized by simulating two media channels which are labeled *media 1* and *media 2*. Media 1 obeys a concave response pattern whereas media 2 follows an S-shaped response curve. Additionally, media 2 profits from synergistic effects, whereas media 1 does not. This can be achieved by time ordering of the media channels, that is media 2 impacts the population segmentation after media 1.

We divide our empirical strategy into two parts: The first part reflects on shape effects in a static setting whereas the second part assumes a dynamic marketing environment with structural change. We begin by elaborating on the first part and subsequently introduce our way of thinking about dynamic change to pin down the strategy in the second part.

In the first part, data is generated, matching some key characteristics of a real dataset (see section 6). This observation window is referred to as *window 1*. Subsequently, all the different models (see section 5.1) are fitted to model the sales-generating process. After that, the models are locally decomposed (see section 5.2). Again, local decomposition refers to tracing out the factor contributions for each instance of a given factor. Concerning media channels, this yields a scatter plot in the media spend - sales plane where one scatter point for each observed media investment is retrieved. Fitting a curve with some desirable properties to the scatter yields the ROI-curve. The implied ROI-curve can finally be compared to the ground truth (see section 5.3). This concludes part one.

Introducing part two, we first discuss our way of thinking about dynamic change in a marketing environment and by taking into account the standard approach to MMM. As has been extensively discussed in the time-series literature (see section 3.3) there exist several strategies to account for structural changes contingent on the very nature of the structural change (break vs. transition, stochastic vs. deterministic). It is its own research interest to discover the changing efficiency of marketing interventions over time. It is important to emphasize that the here followed methodological approach to MMM requires some stability because one specific media response curve is traced out over a time window. If the model coefficients would change for each observation in that time window then each point lies on a different response curve. This insight is crucial because it very clearly highlights the limitations of this MMM approach to deal with dynamic change. This, in turn, implies that we have to narrow down the understanding of dynamic change which we do now.

Structural change occurs only in the sense of a shift in media efficiency. Graphically, this scales the ROI-curve along the sales dimension. Still, there are many reasons why such might be the case. In particular, with relatively new and fast-changing marketing channels (such as social media platforms) such a shift could be triggered by an evolving reach of that channel. The reach of a channel might depend on the popularity of the respective media platform in a particular time period. Hence, to implement the research strategy, a break in the audience size of each media channel is imposed.

This understanding pins down the strategy in part two. Again, AMSS simulates data according to the very same simulation specification as in window 1 except for the media reach. This observation window is referred to as *window 2*. By the reasons discussed above, there exists a structural break

between window 1 and window 2. A time-varying parameter model (see section 5.1.4) is fitted to the pooled data (window 1 and 2). The implied ROI-curve for window 1 can then be compared to the models of the first part. The following example illustrates this conception: Assume the marketer has previously made a MMM analysis for window 2. Subsequently, he would like to update the media allocation and in the meantime, he has the data of window 1 available. The marketer is aware that advertising efficiency has changed between the two windows. Should he pool the data and fit a time-varying effect model or should he make a recency split and only consider data collected during window 1? This is the question and this concludes part two.

5 Methodology

The following sections introduce the relevant theoretical foundation of all the modeling approaches. Further the three different factor decomposition approaches are explained. The concept of Shapley values is crucial for both Shapley Value Regression and Shapley Additive Explanations. The reader is therefore introduced to a general notion of the Shapley values which can be leveraged in both sections. All the applied estimation techniques are mentioned and intuitively but briefly explained where beneficial. The section should yield insights for why given methodologies were chosen, what their advantages and disadvantages are and how they differ from one another. This lies the foundation for a meaningful discussion of the results. The attentive reader might be able to anticipate some of the later discussed outcomes. Finally, as the ultimate interest lies in comparing the estimated response curves to the ground truth, a notation for similarity between curves is required. Throughout the remainder of the text, upper case letters denote random variables whereas their lower case counterparts stand for a particular realization.

5.1 Modeling Approaches

As should be clear intuition by now, the model specification has to allow for the assumed shape of the response curves. Following that proposition, several modeling approaches are discussed, highlighting respective strengths and weaknesses. The subsequent sections should hence provide different potential modeling strategies but also caution the reader why some strategies might fail to capture implied ROI-patterns in the simulation study.

5.1.1 Ordinary Least Squares (OLS)

We begin by introducing a general regression MMM specifying a parametrized sales function of the form

$$y_t = F(x_{t-L+1}, \dots, x_t, z_{t-L+1}, \dots, zt; \Phi) \qquad t = 1, \dots, T \tag{12}$$

where y_t is the sales at time t, $F(\cdot)$ is the regression function, $\mathbf{x}_t = \{x_{t,m}; m = 1, \dots, M\}$ is a vector of ad channel variables at time t and similarly $\mathbf{z}_t = \{z_{t,c}; c = 1, \dots, C\}$ is a vector of control variables. Φ is the vector of parameters to be estimated in the model.

The remainder of the section 5.1 centers around the discussion of the appropriate multidimensional functional form of $F(\cdot)$ which is inherently linked to the one-dimensional response curve and hence pins down changes in sales caused by a change in one particular ad channel (Chan and Perry, 2017).

As can be seen in the general form of 12, the model specification includes lagged variables which link to the ad stock discussion in section 2. As elaborated there, we follow the approach of separating the lagged effects by an appropriate ad stock transformation which can be considered as a different problem. Therefore, after the transformation, the feature space includes only contemporaneous effects. Given that problem separation, the modeler is yet left to argue for his specification of $F(\mathbf{x}_t, \mathbf{z}_t, \Phi)$. Chan and Perry (2017) argues that both the functional form and the choice of members $\mathbf{x}_t$ and $\mathbf{z}_t$ to include are ambiguous due to the complexity of the sales response process. To guide the thinking process, Hanssens et al. (2003, Chapter 3) links the discussion to returns to scale and threshold effect arguments. The following example provides intuition

$$y_t = \beta_0 + \beta_1 x_{t,1} + \cdots + \beta_M x_{t,M} + \beta_{M+1} z_{t,1} + \cdots + \beta_{M+C} z_{t,C} + \varepsilon_t \tag{13}$$

Clearly, the choice of this functional form implies no interaction (synergistic) effects between media variable nor between media and control variables. The implied response curve is linear for each ad channel as $F(\cdot)$ features constant returns to scale. This illustrative example should make it clear, that such a specification is not appropriate for the task of scrutinizing saturation effects.

To simplify notation and as we treat media and control variables symmetrically in our theoretical discussion, we represent all independent variables in the vector $\mathbf{x} = \{x_p; p = 1, \ldots, P\}$ and $P = M + C$. As we only consider contemporaneous effects, time indexing is omitted.

A first natural extension of the model in 12 is polynomial regression with interaction terms

$$y = c_0 + c_1\mathbf{x} + c_2\mathbf{x}^2 + \cdots + c_n\mathbf{x}^n + \beta \text{ Interactions} + \varepsilon \tag{14}$$

where all coefficients now are vectors, n is the polynomial degree. *Interactions* can either be specified between two or multiple media or control variables or between media and control variables. If an S-shaped response curve is expected to reflect the most complex ROI-pattern, three degrees are sufficient. As described in the subsequent section 6 the simulated data has three features that are available to the marketer to model the sales process. The features will be described later, but the model is described in terms of these variables here. The polynomial regression specification reads

$$\begin{aligned} sales = & c_0 + c_{1,1}(media.1.spend) + c_{2,1}(media.1.spend)^2 + c_{3,1}(media.1.spend)^3 + \\ & c_{1,2}(media.2.spend) + c_{2,2}(media.2.spend)^2 + c_{3,2}(media.2.spend)^3 + \\ & c_{1,3}(market.rate) + \beta_1(media.1.spend \times media.2.spend) + \\ & \beta_2(media.1.spend \times market.rate) + \\ & \beta_3(media.2.spend \times market.rate) + \varepsilon \end{aligned} \tag{15}$$

Yet another frequent choice in MMM is multiplicative modeling. Hanssens et al. (2003, pp. 102) and Tellis (2006) write that the multiplicative power model is the most popular one among marketers as it allows for the highest order of interaction between the variables and for flexible, non-constant behaviour of response curves. The model reads compactly

$$y = \beta_0 \left(\prod_{p=1}^{P} x_p^{\beta_p} \right) \varepsilon \tag{16}$$

This model is also known as *log-log* specification as it can be log-transformed and rewritten as

$$\ln y = \ln \beta_0 + \beta_1 \ln x_1 + \cdots + \beta_P \ln x_P + \ln \varepsilon \tag{17}$$

where $\ln \varepsilon$ is now assumed to be normally distributed. The slope of the response curve is characterized by the first partial derivative

$$\frac{\partial y}{\partial x_p} = \beta_0 \beta_p x_1^{\beta_1} \dots x_p^{(\beta_p - 1)} \dots x_{t,P}^{\beta_P} \tag{18}$$

Clearly, the slope of the ROI-curve has many degrees of freedom as it depends on the other feature instances and coefficients. Still, it is widely known that the power (log-log) model exhibits constant returns to scale as the coefficients can directly be interpreted as elasticities. Constant elasticities further imply either increasing ($\beta_p > 1$) or decreasing ($\beta_p < 1$) returns to scale and thus either concave or convex response patterns.

In terms of the three simulated variables the model reads

$$\ln \textit{sales} = \beta_0 + \beta_1 \ln \textit{media.1.spend} + \beta_2 \ln \textit{media.2.spend} + \beta_3 \ln \textit{market.rate} + \varepsilon \tag{19}$$

5.1.2 Shapley Value Regression (SVR)

Shapley values originate from coalition game theory being concerned with the fair allocation of a payout to each contributing member of a group. Applying the concept of Shapley values in a regression setting was first done by Lipovetsky and Conklin (2001). The Shapley Value Regression's (SVR) desirability stems from OLS not being able to handle strong multicollinearity (MC) in the feature space thereby destabilizing the regression coefficients (Mishra, 2016). As mentioned by Chan and Perry (2017) MC arises as a natural problem in MMM as marketers tend to align multiple marketing campaigns which leaves media spend on different channels correlated.

SVR proposes a unique strategy to assess the contribution of regressor variables to the regressand variable. To formalize the concept we render the problem as follows: The value of R2 is known after fitting a regression model $F(\mathbf{x}; \Phi)$ and considered as the value of a cooperative game played by $\mathbf{x}$ (whose members $\{\mathbf{x}_p; p = 1, \dots, P\}$ work in a coalition) against y (explaining it). The analyst only knows the total joint contribution but would like to infer the individual contributions of each player x_p. The Shapley value decomposition imputes the most likely contribution of each individual player to R2 (Mishra, 2016).

Having formalized the game which the dependent variables play in a regression setting, a more general definition of Shapley values follows. The definition can be leveraged later in section 5.2.3 where the concept will be encountered again, with the aim of decomposing the model in its local factor contributions. The definition allows an exact connection to the game setting introduced in the previous paragraph. The Shapley value for feature j is (Molnar, 2019, Chapter 5.9.3.1)

$$\phi_j(val) = \sum_{S \subseteq \{x_1, \dots, x_p\} \setminus \{x_j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{x_j\}\right) - val(S)\right) \tag{20}$$

where S is a subset of the features used in the model, $\mathbf{x}$ is the vector of features to be explained and p the number of features. $val(\cdot)$ is some value function and depends on the game being played and in particular the payoff to be distributed. In the particular case of SVR the value function is the computation of the R2.

As is evident from 20 the computation of the 'feature importance' of j requires retraining the model for all possible subsets S. For each possible coalition S the difference between the R2 including feature j and the R2 of the model withholding feature j has to be computed. These differences are then weighted accordingly (Lundberg and Lee, 2017).

Shapley values possess some desirable properties which will be discuss in 5.2.3 where the fairness discussion (a fair payout distribution) of different attribution methods lies at the core.

As mentioned above, the actual computation of the exact Shapley values is intensive because of the many possible coalitions and the necessity of retraining the model twice for each coalition vector. The estimation follows by means of a Monte Carlo approximation where random coalitions are sampled and the feature of interest j is not replaced for the first value function evaluation whereas for the second, it is replaced from the sampled vector.

The crux of SVR is now to reweight the coefficients of the OLS estimation with the help of the Shapley value vector V. We denote these standardized regression coefficients by the vector α and compute them according to the quadratic programming problem (Mishra, 2016)

$$\min_{\{\alpha_p;1,\dots,P\}} f(\alpha) = \sum_{p=1}^{P} (\alpha_p(2T - U\alpha)_p - V_p)^2 \tag{21}$$

where U is the pair-wise correlation vector among regressors and T the pair-wise correlation vector between regressand and regressors. The quadratic programming problem is not intuitive but can be derived via an adjusted net effect formulation as fully described in Lipovetsky and Conklin (2001).

To summarize, SVR yields a more fair feature attribution by reweighting the regression coefficients with help of Shapley values. Such an approach is expected to be superior to OLS in cases of high multicollinearity between the features. In MMM such correlations between variables arise naturally because of cyclical marketing campaigns, alignment of marketing ad expenditures and synergistic effects between media channels. All of these causes are present in our simulated data by matching metrics from real data. We use the package **relaimpo** for the computation of relative feature importance (Shapley values) and write an optimization in line with 21 to implement SVR in R. Making SVR comparable to OLS, the same model specification as in 15 and 19 is used in our simulation analysis.

5.1.3 Generalized Additive Models (GAM)

Generalized Additive Models (GAM) have been developed to capture non-linear relations between dependent and independent variables. One might argue, that OLS already achieves this by transforming the features accordingly. As introduced in section 5.1.1, polynomial regression does so by taking independent variables to the power of some degree. The choice of the appropriate degree is not trivial and optimally requires some sort of prior knowledge or repeated fitting and cross-validation. In the context of MMM the model might be overfit when imposing a polynomial of third degree to a media channel which in fact exhibits a concave response pattern. GAM allevi-

ates this problem by modeling the dependent variable as a sum of smooth functions (Hastie and Tibshirani, 1986). The model reads

$$g(E_Y(y|\mathbf{x})) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p) \tag{22}$$

where $g(\cdot)$ is the link function and $f_i(\cdot), i = 1, \ldots, P$ is some smooth function.

The *Generalized* in GAM refers to the fact that the outcome variable might not be normally distributed (but allows for any distribution of the exponential family) (Molnar, 2019, Chapter 4.3.3). From now on we assume the outcome to be Gaussian distributed which leaves the link function being the identity function and hence can be omitted in the notation. Clearly, the simple OLS is a nested model when we replace $f_i(x_i)$ by $\beta_i x_i$.

The question is how to learn such non-linear functions and what degree of smoothness should they have? There exist many different possibilities to achieve the estimation of such smooth functions. For an elaborate discussion, the reader can refer to Hastie and Tibshirani (1986). We follow in the remainder of the section Larsen (2015) and give a brief intuitive introduction to *Smoothing Splines* and the approach to determine the degree of smoothness in order to avoid over-fitting. We elaborate briefly on the procedure of fitting one single spline and mention towards the end the routines that are applied to simultaneously fit multiple splines and hence estimate the GAM.

A spline curve $s(x)$ is a piece-wise polynomial curve. The smooth function $f_i(x_i)$ is now approximated with help of several spline curves. For one single spline the penalized sum of squares is minimized

$$\sum_{i=1}^{P} (y_i - f(x_i))^2 + \lambda \int \left(s''(x)\right)^2 dx \tag{23}$$

which is a combination of the residual sum of squares and a penalty term weighted by λ which controls the trade-off between model fit and smoothness. A straight line has the same slope for each value of x and hence the integral evaluates to 0.

As mentioned, when fitting GAM, multiple spline functions have to be estimated simultaneously which is achieved by maximizing the *Penalized Likelihood Function* with some sort of *Backfitting Algorithm*. The final smooth function is then a composition of the fitted splines. The optimal smoother λ is determined by the *Generalized Cross Validation Criteria* which is based on a 'leave one out' cross-validation approach. This involves repeatedly fitting the model to all but one data point and then estimating the prediction error for that particular point.

To summarize, GAM is a flexible additive modeling approach which is more suitable than the strategy to transform variables (such as polynomial regression) because the appropriate degree of smoothness is determined internally. This is to say that GAM requires no prior knowledge of the response curve's functional forms nor extensive and explicit cross-validation of different model fits. On the other hand, GAM is not immune to the problem of multicollinearity which might make it inferior to SVR. We use the package **gam** for the implementation in R.

Applied to our simulated data, we specify the model as

$$\begin{aligned} sales = &\beta_0 + f_1(media.1.spend) + f_2(media.2.spend) + f_3(market.rate) + \\ &\beta_1(media.1.spend \times media.2.spend) + \\ &\beta_2(media.1.spend \times market.rate) + \\ &\beta_3(media.2.spend \times market.rate) + \varepsilon \end{aligned} \tag{24}$$

which is a combination of GAM and regular regression. The interaction terms do not handicap GAM over the polynomial regression approach. This alleviates the comparison between GAM and OLS.

5.1.4 Time-Varying Effect Models (TVEM)

"No man ever steps in the same river twice, for it's not the same river and he is not the same man" (Heraclitus). The possibility of changing underlying economic relations over time imposes an additional difficulty when it comes to modeling. The time-series literature suggests multiple modeling approaches to account for such evolving dynamics. One subset of proposed techniques is collected in the term *Varying-Parameter Models*. As discussed in section 3 these models mainly differ in their underlying assumptions with respect to, for example, how parameters evolve over time. A general framework not requiring any prior knowledge was introduced by Hastie and Tibshirani (1993). Interestingly, he introduces Varying-Coefficient Models by mentioning the link to GAM. However, the generalizations refer to the proposition that models are linear in the regressors but parameters are allowed to change smoothly with the value of other variables which are called *effect modifiers*. In this respect Tan et al. (2012) simply define the effect modifier to be time and call the result Time-Varying Effect Models (TVEM).

Compared to other traditional analytical approaches, TVEM does not require strong parametric assumptions about the nature of change between time-varying covariates and the outcome variable thus allowing to model the change in a flexible manner (Tan et al., 2012). Similar to the functional forms in GAM, the change $\beta(t)$ is smooth. We therefore write

$$y_t = X_t\beta(t) + \varepsilon_t \tag{25}$$

where $X_t = \{\mathbf{x}_{t,p}; p = 1, \ldots, P\}$ is the matrix of time-ordered features and $\beta(t) = \{\beta_p(t); p = 1 \ldots, P\}$ is the vector of now time-varying parameters. The error terms are assumed to be normally and independently distributed.

In principle and as mentioned, the effect modifier could be any other variable and has not to be time. Additionally, the time-ordered observations don't have to be evenly spaced which implies for the context of MMM that different observation periods can be linked (with missing observations in between). Such might be the case if businesses would like to have a reevaluation of their media allocation after some time. For model comparison, we choose the polynomial model specification introduced in equation 14. Tan et al. (2012) mention that the modeler is free to incorporate prior knowledge on some coefficient functions.

The estimation procedure follows similarly to GAM either a spline-based or kernel-based approach. The p-spline method is very nicely described in Tan et al. (2012). The interested reader is also referred to Hastie and Tibshirani (1993). In any case and as intuitively described in section 5.1.3

there is a trade-off between smoothness and over-fitting which is again resolved by the *Generalized Cross Validation Criteria* (Tan et al., 2012).

To summarise, TVEM does not require any prior knowledge and assumptions on coefficient functions other than that they have to be of continuous and smooth form. The coefficient functions allow for additional insights in how the efficiency of marketing channels changed over time. Time-ordered observations don't have to be evenly spaced which makes it suitable for applications in MMM where reevaluation of marketing strategies happens irregularly. We use the package **tvReg** for the implementation in R. To compare TVEM to the previous approaches, we apply the same model specification as in 15 in our simulation analysis.

5.2 Local Decomposition Approaches

This section presents three local factor decomposition approaches. 'Factor decomposition' refers to the contribution of each variable towards the predicted value (sales in this case). 'Local' implies a decomposition for each observation and hence for each realized factor value. If the independent variable is media spend for a particular channel then these factor contributions are the media's sales driver. Some sort of fitted curve can be interpreted as ROI-curve.

The most intuitive approach (weighted factor decomposition) is first introduced and should be perceived as the current benchmark. We further proceed by explaining more elaborated techniques which correct for particular shortcomings. Most of the decomposition approaches stem from recent efforts to explain 'black box' machine learning models (such as neural networks). This field of *Model Agnostics* is concerned with separating the explanation from the Machine Learning model and has the advantage of not being model-specific (Molnar, 2019, Chapter 5). The derived methodologies can thus be applied to simpler parametric models as well. Under certain conditions, several decomposition methods yield the same or similar results as will become evident.

To simplify the discussion we outline the problem with an example where, again, upper case letters denote random variables

$$Y_t = \beta_1 X_{1,t} + \beta_2 X_{2,t} + \beta_2 X_{1,t} X_{2,t} \tag{26}$$

How much does $X_{1,t} = x_{1,t}$ contribute to $Y_t = y_t$? Clearly, it depends on $X_{2,t}$ as the two factors interact. The question is, how can the collaborate contribution be fairly allocated to the two correlated variables. In brief, the *Weighted Factor Decomposition* approach relies on one particular realization of $X_{2,t}$ whereas *Accumulated Local Effects* consider its conditional density. On the other hand *Shapley Additive Explanations* follow a game-theoretic approach from coalition theory as will be explained in detail. All approaches share the idea of more or less elaborate weighting schemes. They differ in the manner these weights are determined.

5.2.1 Weighted Factor Decomposition (WFD)

The most trivial of such a weighting scheme is applied in WFD. This makes it ideal as a first intuitive benchmark approach to which the two subsequent procedures can be compared to. The basic intuition follows Suarab et al. (2014) but we extend the discussion of the methodology and choose a notation in line with Molnar (2019) to simplify the comparison to the methods introduced at a later stage and to generalize the approach to non-parametric models.

The WFD approach follows a three-step procedure described in the algorithm below. The reader might think of $x_s \in X_S$ as being media spend of one particular channel s at a particular point in time (time subscripts are omitted).

for *factor* X_S *in* X **do**
  for *each* $x_s \in X_S$ **do**
    1. substitute each instance x_s of feature X_S by mean(X_S);
    2. compute unscaled contribution C_{x_s} according to 28;
    3. reallocate difference between actual value y and sum of unscaled contributions according to weights calculated from the absolute values of unscaled contributions according to 27;
  end
end

Algorithm 1: WFD

The original proposition by Suarab et al. (2014) is to set the respective coefficient to 0 instead of substituting feature values. Of course such is only possible in a parametric approach and thus not feasible in a GAM model. Instead of setting the coefficients to 0 one might set the feature to 0. After all, this follows the intuition of setting the feature effect to 0. But, if the model specification is of multiplicative form, it follows that the unscaled contribution is equal to y for each factor (because the model evaluates to 0). To make the WFD procedure applicable to all the proposed methods, each instance of the particular feature of interest can be substituted by its mean value (as is common in other decomposition methods).

The discussion above centers around the understanding of *'leaving a factor out'*. In mathematical notation $x_{\setminus} s$ denotes the vector of one particular observation 'without' factor s. The two necessary computations according to the algorithm above are

$$\hat{f}_{x_S,WFD}(X_S) = C_{x_s} + \underbrace{\frac{|C_{x_s}|}{\sum_{i=1}^{n} |C_{x_i}|}}_{\text{weighting scheme}} \underbrace{\left(y - \sum_{i=1}^{n} C_{x_i}\right)}_{\text{reallocate}} \tag{27}$$

$$\text{where } C_{x_s} = y - \hat{f}(x_{\setminus s}) \tag{28}$$

where x_s is the feature instance for which we compute the contribution, C_{x_i} stands for the unscaled contribution of feature instance x_i, $\hat{f}$ is the estimated model to be decomposed, and n is the number of dependent variables. The notation $\hat{f}_{x_s,WFD}(x_s)$ refers to the estimated contribution of x_s following the WFD approach and being a function of the random variable X_S.

It should be clear, that in equation 28 the unscaled contribution is solely based on one single realization of each feature in the feature space. To take up the introductory example, one realization of X_2. Literally, this makes the individual contributions calculated by the WFD random. The resulting scatter in the *media spend - sales contribution plane* can therefore be expected to have a large variation along the sales' axis caused by variation of the variables X_i in a small neighborhood around x_s.

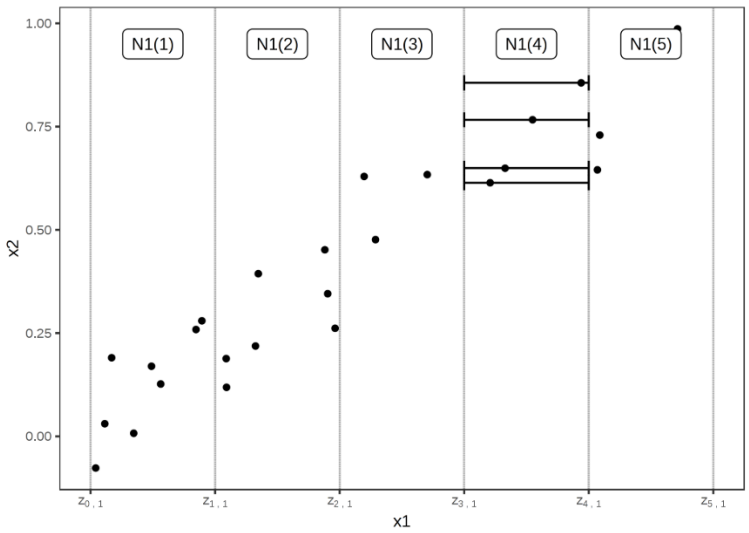

Figure 7: Calculation of ALE for feature x1, which is correlated with x2. Source: Molnar (2019, Chapter 5.3.1).

5.2.2 Accumulated Local Effects (ALE)

To correct for this randomness of the first approach ALE describes how a feature affects the prediction on average. In particular, ALE respects the correlation relation between features by considering the conditional density of all X_i when computing the contributions.[4] Intuitively, at a grid value of x_s predictions of instances with similar x_s are averaged. Still, we are left to split up the correlation (synergistic) effect by computing the difference in the predictions instead of taking the averages (Molnar, 2019, Chapter 5.3).

The intuition from figure 7 is best understood by going through the following algorithm.

for *factor* X_S *in* X **do**
1. divide the features into neighborhoods (vertical lines) $NS(1) - NS(k)$;
2. for the data instances in each neighborhood calculate the difference in the prediction when we replace the feature x_s with the upper and lower limit of the interval (horizontal lines);
3. accumulate and center these differences to get the ALE

end

Algorithm 2: ALE

The description of the procedure makes it very intuitive why it is called Accumulated Local Effects. *Local* refers to small neighborhoods and the instances in it (conditional density) whereas the *Accumulated* refers to the aggregation of these neighborhoods to provide an overall picture. By centering the ALE the feature effect stands in relation to the average prediction which makes it nice for interpretation. At this point, it should be mentioned, that ROI-curves should not be

[4]Partial Dependency Plots (PDP) are similar to ALE but consider the marginal distribution, thereby omitting the correlation relation between features making it an inferior approach in MMM.

centered. In order to invert the effect of centering, the theoretically sound assumption is imposed that response curves have to pass through the origin.

Leveraging the intuition from the algorithm together with figure 7 the below theoretical foundation of ALE follows naturally

$$\hat{f}_{x_S,ALE}(x_S) = \int_{z_{0,1}}^{x_S} E_{X_C|X_S}\left[\hat{f}^S(X_s, X_c)|X_S = z_S\right] dz_S - \text{constant} \tag{29}$$

$$= \int_{z_{0,1}}^{x_S} \int_{x_C} \hat{f}^S(z_s, x_c)\mathbb{P}(x_C|z_S) dx_C dz_S - \text{constant} \tag{30}$$

$$\text{where } \hat{f}^S(x_s, x_c) = \frac{\delta \hat{f}(x_S, x_C)}{\delta x_S} \text{ is the gradient} \tag{31}$$

where the notation closely follows the previously introduced definition in 27 and 28 and X_C stands for a correlated feature. z refers to the boundary points of the respective intervals where $z_{0,1}$ is the lowest such. In the actual estimation procedure, the gradient is replaced by the difference. The gradient represents an infinitely small interval around x_s.

The previous verbal discussion of the concept aligns even more with the actual estimation procedure characterized by

$$\hat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} \left[f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)})\right] \tag{32}$$

$$\hat{f}_{j,ALE}(x) = \hat{\tilde{f}}_{j,ALE}(x) - \frac{1}{n}\sum_{i=1}^{n} \hat{\tilde{f}}_{j,ALE}(x_j^{(i)}) \tag{33}$$

where equation 32 describes the uncentered ALE and equation 33 adds the step of centering by subtracting the average local effect as a constant. $N_j(k)$ stands for the k^th neighborhood for feature j, $n_j(k)$ is the number of instances in that neighborhood and k indexes the respective bounds of the neighborhood.

To fully grasp the concept, the formula 32 can explained in its individual parts beginning at the far right. The term in the squared brackets captures the differences in prediction, whereby the feature of interest is replaced with grid values z. This yields the effect the feature has for an individual instance in a certain interval. The preceding sum adds up the effects of all instances within a neighborhood which is divided by the number of instances in this neighborhood $n_j(k)$ to obtain the local effect. Lastly, these average effects are accumulated over all neighborhoods (Molnar, 2019, Chapter 5.3.3).

The choice of intervals (neighborhoods) has a certain influence on ALE curves. It should be clear that the ALE curve is a piece-wise linear approximation of the theoretical ALE. This is because the estimated ALE is linear on a given interval. The smaller these intervals the better the fit. But on the other hand, ALE is also depending on the number of instances in each neighborhood, crucially so if the local effect of the feature is depending on the other features (interactions). Hence, the modeler faces a trade-off between the size of the intervals (or similar the number of intervals) and the number of data points within each interval (Altmann et al., 2020, Chapter 7.1).

The intermediary conclusion is that ALE corrects for the randomness of WFD by utilizing conditional expectations. The conditional density accounts for correlation between features. We use the package **iml** for the implementation in R.

5.2.3 Shapley Additive Explanations (SHAP)

The concept of SHAP is closely linked to the computation of Shapley values introduced in section 5.1.2. SHAP is the only approach that respects some desirable properties (linked to a fair payout distribution) but falling short in accounting for correlation in the feature space as will be elaborated.

Like the other local decomposition approaches, SHAP aims at explaining the prediction of an instance x by computing the contribution of each feature to the prediction. SHAP is a permutation-based approach where feature values act as players in a coalition and the games' payoff is the predicted value (Molnar, 2019, Chapter 5.10.1).

Recall section 5.1.2 where the definition of Shapley values was introduced in equation 20 which is here repeated in 34

$$\phi_j(val, x) = \sum_{S \subseteq \{x_1,\dots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{x_j\}\right) - val(S)\right) \tag{34}$$

$$val_x(S) = \int \hat{f}(x_1, \dots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X)) \tag{35}$$

where (as previously) S is a subset of the features used in the model, x is the vector of features to be explained and p the number of features. $val(\cdot)$ is some value function and depends on the game being played and in particular the payoff to be distributed. Here, the value function is explicitly written out as the model prediction function.

The formula 35 suggests that multiple integrations need to be computed for each feature not contained in S. For example in the case of four features and the coalition S consisting of x_1 and x_3

$$val_x(S) = val_x(\{x_1, x_3\}) = \int_{\mathbb{R}} \int_{\mathbb{R}} \hat{f}(x_1, X_2, x_3, X_4) d\mathbb{P}_{X_2 X_4} - E_X(\hat{f}(X)) \tag{36}$$

The estimation follows by means of a Monte Carlo approximation

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{f}(x^m_{+j}) - \hat{f}(x^m_{-j}) \right) \tag{37}$$

where M is the number of iterations. Formula 37 is explained in more depth by going through the algorithm 3 which is based on Molnar (2019, Chapter 5.9.3.3). It describes the procedure to derive the Shapley value of the j^{th} feature when the game of prediction is played.

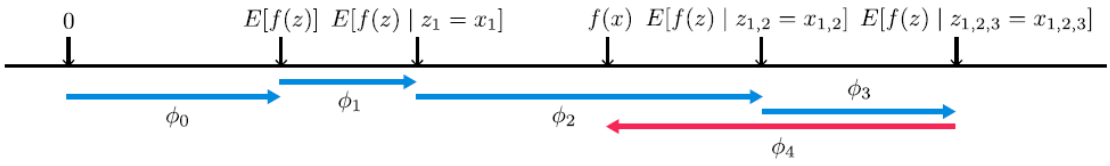

Figure 8: SHAP values attribute to each feature value the chain in prediction compared to some base value. Source: Lundberg and Lee (2017).

for *all* $m = 1, \ldots, M$ **do**
1. draw random instance z from the data matrix X;
2. choose a random permutation o of the feature values (order matters);
3. order instance x; $x_o = (x_{(1)}, \ldots, x_{(j)}, \ldots, x_{(p)})$
4. order instance z; $z_o = (z_{(1)}, \ldots, z_{(j)}, \ldots, z_{(p)})$
5. construct two new instances;
   (i) with feature j: $x_{+j} = (x_{(1)}, \ldots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \ldots, z_{(p)})$;
   (ii) without feature j: $x_{-j} = (x_{(1)}, \ldots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \ldots, z_{(p)})$;
6. compute marginal contribution (value function): $\phi_j^m = \hat{f}(x_{+j}) - \hat{f}(x_{-j})$;
7. compute Shapley value as the average: $\phi_j(x) = \frac{1}{M} \sum_{m=1}^{M} \phi_j^m$

end

Algorithm 3: SHAP

So far, the only thing that changed compared to the discussion in the previous section 5.1.2 is the definition of the value function. The extension of SHAP is the link of Shapley values to the local interpretability of other model agnostic methods (such as LIME) where a model $f(x)$ is *locally* approximated with an explainable model $g(x)$ for each instance of each factor X. This connection between the two concepts yields that the desirable properties of the Shapley values (mentioned below) hold at the local level (for each feature instance).

To formalize the notion we write (Lundberg and Lee, 2017)

$$f(x) \approx g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j x' \tag{38}$$

$$= E_X(\hat{f}(X)) + \sum_{j=1}^{M} \phi_j \tag{39}$$

where g is the explanation model, x' is the coalition vector, M is the maximum coalition size and $\phi_j \in \mathbb{R}$ is the feature attribution for a feature j, the Shapley values. The coalition vector consists of the feature value if the feature is present in the coalition and 0 otherwise. Formula 38 implies that we can additively explain each predicted value by feature contributions ϕ_j compared to some base value ϕ_0 (compare figure 8). The base value can be defined as the average predicted value which leads to 39 (Lundberg and Lee, 2017). In addition, the following properties hold:

- **Local Accuracy:** The feature contributions must add up to the difference of prediction for

x and the average.

- **Missingness:** If a feature instance is 0 then its contribution should be 0.
- **Dummy:** A feature value that does not change the prediction, regardless in which coalition, should get a contribution of 0.
- **Symmetry:** If two feature values are the same, then their contribution should be the same.
- **Linearity:** If a game has combined payoffs, then the contributions in each subgame can be added to arrive at the combined contributions.

SHAP is the only attribution method that fulfills these properties. For example, ALE violates the missingness property because it averages over an interval around the observation. Additionally, ALE does not fulfill the symmetry property either. The linearity property ensures that SHAP is even suitable for random forest algorithms and others, which train several models to arrive at one prediction. On the other hand, ALE respects the correlation between features by considering the conditional densities. By construction, SHAP does not consider conditional densities but samples from the whole distribution (of the training set). This permutation-based approach has the drawback that highly unlikely feature value combinations might result when features are correlated. We use the package **shapper**[5] for the implementation in R.

5.3 Similarity Between Curves

From a statistical viewpoint, similarity measures are often inversely related to some distance measure. Consider for example the Euclidean distance. A marginal right shift of two initially overlapping curves might lead to a large dissimilarity measure and a linear approximation might be considered more similar based on this approach. Recall from section 2, that the recovered ROI-curves are later to be used as part of the objective function in an optimization problem. As mentioned, the shape of a response curve might matter more than the actual location in space. Optimal media allocation is characterized by equating marginal return to marginal cost for a particular media investment as seen in equation 4. Hence again, a small shift of the curve does not translate to dramatic changes in optimal allocation.

In order to avoid the trade-off implied by choosing an appropriate distance measure[6], the problem can be tackled from another perspective. The factor contributions (retrieved from the local factor decomposition) represent a scatter in the media spend - sales plane. The actual ROI-curves are retrieved by fitting an appropriate curve. The curve has to fulfill certain properties: it should be non-decreasing, allowing for concave and S-shaped patterns and depict a saturation point. The logistic functional form fulfills these desired properties.

The logistic function evaluates to

$$f(x) = \frac{\text{Asym}}{1 + exp((\text{xmid} - x)/\text{scal})} \tag{40}$$

[5] **shapper** is a python wrapper library

[6] An appropriate measure of similarity could be derived from the Dynamic Time Warping (DTW) algorithm usually applied to compare multiple time series. But its application is not restricted to sequences with time dimensions only. In our case, we could consider the media spend to be the time axis.

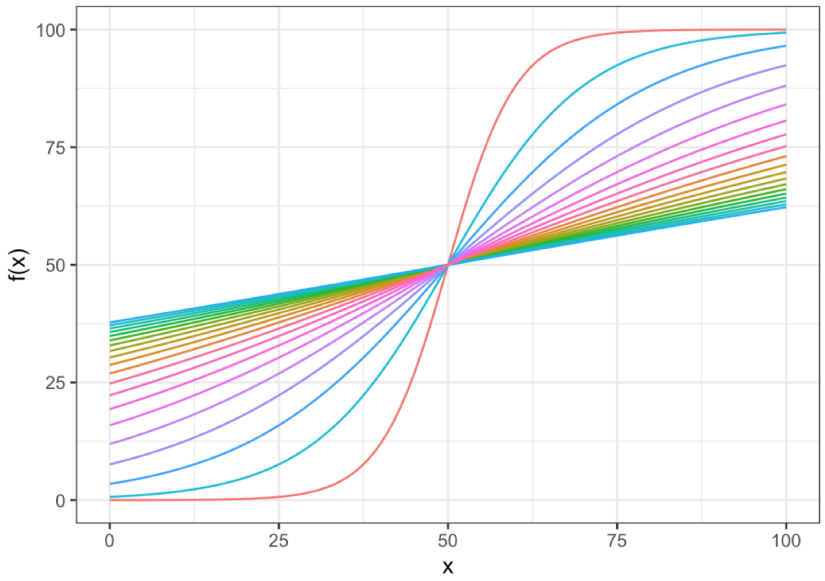

Figure 9: Logistic curves with different "scal" parameter values.

where $Asym$ represents the asymptote, $xmid$ is the x value at the inflection point of the curve. The value of $f(xmid)$ will be $Asym/2$ at xmid. Finally, $scal$ is the scale parameter characterizing the steepness of the curve. In figure 9, $Asym$ is set to 100, $xmid$ to 50 and the third parameter is defined according to $scal \in \{5, 10, \ldots, 100\}$ to give an illustration. Depicting all traced-out parameter values from the Monte Carlo simulation study in boxplots, alleviates an understanding of why the estimated response curves deviate from the ground truth by analysing the parameter fit individually (for Asym, xmid and scal).

Additionally, the Mean Absolute Percentage Error (MAPE) is computed across all 500 iterations for each parameter and each model, decomposition tuple according to the formula

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \tag{41}$$

where A_t refers to the actual value and F_t is the forecast value. The MAPE loss function was chosen as media channels differ in their sales' contribution and parameter values are defined over different scales. As a consequence the function allows to scrutinize whether a modeling, decomposition approach outperforms others either media-specific or for both media types. Similarly, a trade-off between logistic parameters when choosing a given modeling, decomposition approach would be indicative (for example depicting low MAPE values for Asym and xmid but a high value for scal).

6 Data

In this section, we leverage the theoretical description of the AMSS from 3.4 and define all the necessary parameters. As mentioned, the simulated data should follow some key characteristics of a real-world example which in this case is a big Swedish electronic retailer. The calibration to the real data is not an exact science and the reader should be aware that there are many degrees of

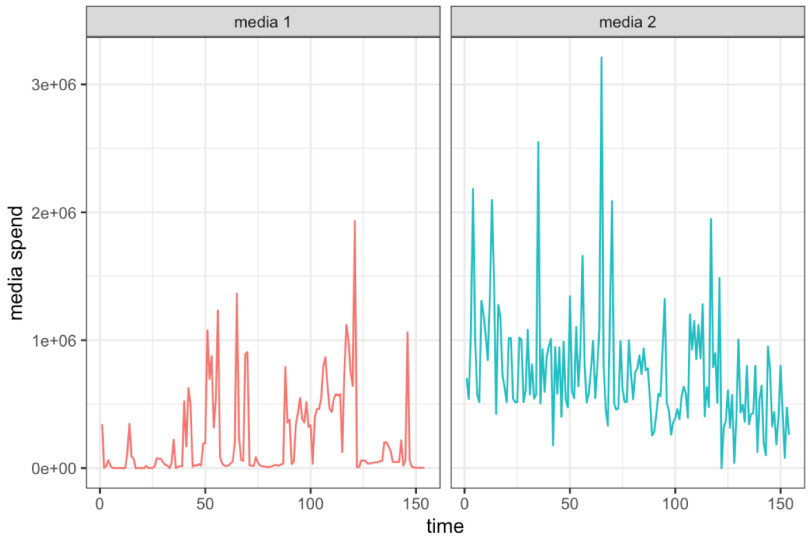

Figure 10: Flighting patterns of the two selected media channels.

freedom. The reason for this is, of course, the missing empirical counterparts for various parameters (such as transition matrices). Still, the choice of appropriate values should be defended. One can argue that all the models tested are exposed to the same potential misspecification bias. Further, by matching the characteristics, central properties of the simulated marketing environment behave realistically. The section is structured as follows: First, the real dataset is briefly described to subsequently define the key properties to be matched. Next, we describe our calibration routine and pin down all parameter values of the AMSS. Intuition for the choice is given where necessary. Then we describe how *window 2* is discriminated from *window 1* by altering media reach which imposes a structural break between the two windows. The section concludes by comparing the simulated data to the real data and thereby presents evidence for a successful calibration. Throughout the section, the terminology introduced in 3.4 applies.

As mentioned, the real dataset serves only as a source of key properties. These properties are:

- marketing expenditure (flighting) patterns
- signal strength of the media variables (sales to investment ratio)
- multicollinearity in the feature space
- signal to noise ratio

By achieving this, the testing environment is valid. The real dataset has dimensions (154×15) and consequently consists of 154 weekly sales' observations. Additionally, there are 9 marketing variables and 3 additional sales-driving variables (such as holidays) and a time index. Out of these 9 marketing variables we choose two, in order to align the simulated marketing interventions. The flighting pattern is exactly matched and depicted in figure 10. This concludes the calibration of the first key metric. If the simulated sales process follows the real sales sequence, a realistic sales to investment ratio follows which concludes the calibration of the second key metric.

The sales data is decomposed by leveraging the functionality of the **prophet** package. This decomposition pins down the market size referred to as ρ_t in equation 8 which is the in-market target rate.

This fluctuating variable is made known to the modeler and will be referred to as 'market rate'. A 10% noise ratio is added to the observed market rate (not observed by the modeler) which guides the signal to noise ratio. Marketers are usually capable of recovering around 95% of the variation in the sales process (R2). The specification is thus rather on the conservative side. As all the simulated variables (market rate and media spend) follow real-world patterns, the multicollinearity in the feature space should be realistic. This paragraph concludes the calibration of the last two key metrics.

All the necessary parameters of the AMSS are subsequently defined in the same order as presented in the theoretical disquisition of section 3.4.

Market size. The model economy is populated by 9 million people which roughly corresponds to the Swedish potential customer population. Market size (the percentage of the population being 'in-market') is determined by employing time series decomposition as elaborated above. The migration necessary to match the in-market target rate is the first event $k = 1$.

Natural migration. Recall that each dimension potentially affected by marketing events is governed by a transition matrix. Recall also, that these matrices characterize equilibrium relations. As described below, the marketing campaigns intervene with all possible ($l = \{3, \ldots, 6\}$) dimensions. Natural migration matrices are row-identical to abstract from lagged effects. As mentioned in section 2 the problem of estimating carryover effects can be treated separately and is not intended here. We therefore only report the first row of the quadratic matrices. The natural transition reads

$$Q^{2,\text{'activity'}} = (0.45, 0.30, 0.25) \tag{42}$$

$$Q^{2,\text{'favorability'}} = (0.03, 0.07, 0.50, 0.30, 0.10) \tag{43}$$

$$Q^{2,\text{'loyalty'}} = (0.50, 0.30, 0.20) \tag{44}$$

$$Q^{2,\text{'availability'}} = (0.30, 0.40, 0.30) \tag{45}$$

where k was set equal to 2 because natural migration is our second event. For example, equation 42 reads: The equilibrium population is segmented in 45% 'inactive', 30% 'exploratory' and 25% 'purchase' oriented individuals. The potential values of each dimension are reported in table 1.

Marketing interventions. As already mentioned, two media channels, *media 1* and *media 2*, are defined. Media 1 is time-ordered before media 2 which implies that the latter profits from synergistic effects. The synergies stem from both ad channels impacting on the favorability dimension. Apart from that common dimension, media 1 causes migration along activity and loyalty, whereas media 2 causes migration along availability. The transition matrices for media 1 read

$$Q^{3,\text{'activity'}} = \begin{pmatrix} 0.50 & 0.30 & 0.20 \\ 0.00 & 0.70 & 0.30 \\ 0.00 & 0.00 & 1.00 \end{pmatrix} \tag{46}$$

$$Q^{3,\text{'favorability'}} = \begin{pmatrix} 0.40 & 0.00 & 0.40 & 0.20 & 0.00 \\ 0.00 & 0.90 & 0.10 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.50 & 0.40 & 0.10 \\ 0.00 & 0.00 & 0.00 & 0.80 & 0.20 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix} \tag{47}$$

$$Q^{3,\text{'loyalty'}} = \begin{pmatrix} 0.50 & 0.25 & 0.25 \\ 0.00 & 1.00 & 0.00 \\ 0.30 & 0.00 & 0.70 \end{pmatrix} \tag{48}$$

where k was set equal to 3, indicating the chronological order of the events. The transition matrices for media 2 read

$$Q^{4,\text{'favorability'}} = \begin{pmatrix} 0.20 & 0.00 & 0.50 & 0.30 & 0.00 \\ 0.00 & 0.80 & 0.20 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.50 & 0.50 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.70 & 0.30 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix} \tag{49}$$

$$Q^{4,\text{'availability'}} = \begin{pmatrix} 0.25 & 0.50 & 0.25 \\ 0.00 & 0.50 & 0.50 \\ 0.00 & 0.00 & 1.00 \end{pmatrix} \tag{50}$$

where k was set equal to 4, indicating the chronological order of the events. These matrices are not in the appendix, exactly because the reader should go through them and strengthen his understanding of why the chosen values are appropriate. We give one example by guiding through equation 46: The activity state can take the three values 'inactive', 'exploratory' and 'purchase' (see table 1). The rows and columns correspond to these states. An inactive individual (first row) remains inactive by a 50% chance after exposure to media 1 (first row, first column). It migrates to the exploratory state by 30% chance (first row, second column) and forms a purchase intent with 20% likelihood (first row, third column). The second row refers to an exploratory individual. Most matrices are upper triangular because marketing interventions drive population segments from less favorable to more favorable states.

The *audience* is determined by a reachability likelihood. All individuals are treated symmetrically by setting the likelihood to 0.5. This reflects more traditional media channels (such as TV) where marketing interventions can not be targeted. As noted by Vaver and Zhang (2017) and Chen et al. (2018) paid search and traditional marketing channels should not necessarily be modeled equally.

Each media channel has its cost function which in turn characterizes volume and hence the average frequency of exposure. Importantly, the cost function is the necessary degree of freedom, such that the full ground truth ROI-curve is realized and not only one limited range which would result in identification problems as depicted in 4. Recall that the true ROI-curve is, among others, shaped by the Hill transformation which scales the maximal probabilities (transition matrices) against the

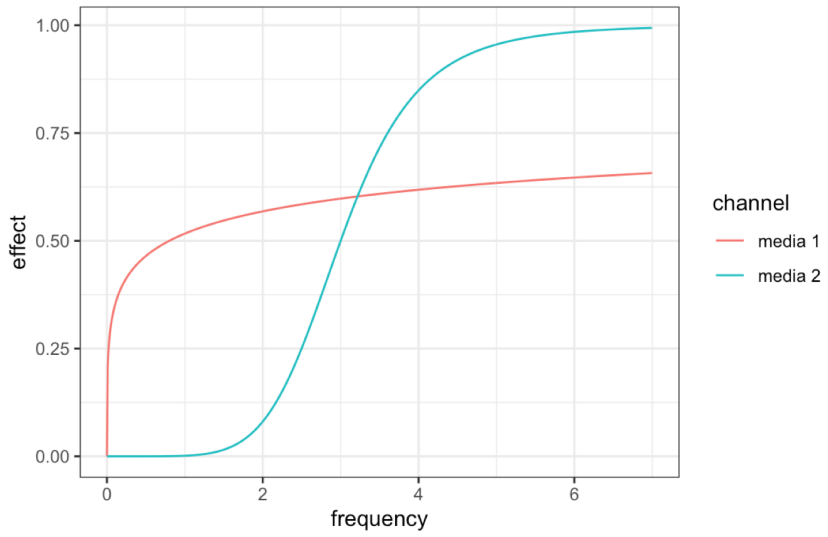

Figure 11: The Hill transformation scales marketing transition matrices against the frequency and determines the shape of the ground truth ROI-curves.

frequency. To control the shape of the ground truth ROI-curves we need to control the frequency which is achieved by adjusting the cost function. The imposed Hill transformation is depicted in figure 11 which leads to media 1 being of concave and media 2 of S-shaped nature.[7]

Sales event. Lastly, the market-clearing conditions are specified according to equation 11. As the company sets only one price (constant supply curve) the segment-specific intercept of the demand schedule determines the purchase likelihood of that segment. These likelihoods read

$$\alpha_{\text{'favorability'}} = (0.01, 0.00, 0.20, 0.30, 0.90) \tag{51}$$

$$\alpha_{\text{'loyalty'}} = (0.50, 1.00, 0.00) \tag{52}$$

$$\alpha_{\text{'availability'}} = (0.10, 0.50, 1.00) \tag{53}$$

Window 2 is simulated according to the exact same simulation specification with the exemption of media reach which is reduced to 0.2.[8]

Figure 12 shows that the simulated sales pattern follows closely the real sales sequence. Additionally, figure 13 highlights the similar correlation structure between the variables by reporting the correlation matrices for the real and simulated data. Together with the spending pattern, visible in figure 10, evidence is provided that the key metrics are matched.

The following assumptions summarize the data section:

- Excluding lagged effects which in practice are modeled separately.

[7] It is important to define the Hill transformation over the whole realized frequency range. Media 1 has a higher frequency range which explains the slower convergence to 1. Most segments have a frequency exposure below 2 which induced us to impose the strongest curvature below that value.

[8] This imposes a structural break as was affirmed by a simple CUSUM test based on the polynomial regression specification as in 15.

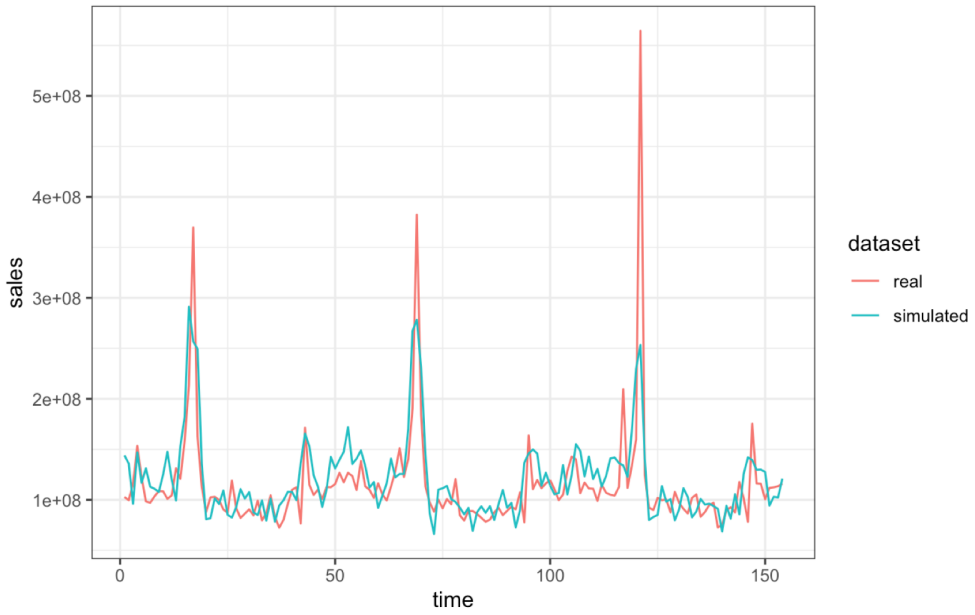

Figure 12: Simulated sales follow closely the sales time-series of the real dataset.

- The market rate is made known to the modeler which is equivalent to no omitted variable bias neither through actually not considering an important variable nor improperly decomposing the time-series.[9]
- Two media channels are simulated. The first depicts decreasing returns to scale and the second S-shaped response patterns. Additionally, there is a synergistic effect from media 1 to media 2.

7 Results

This section presents the results and guides the reader through the relevant output of the statistical analysis. For each figure, the results are first described from a normative perspective. The interpretation and explanation follows thereafter. The section concludes by summarising the key findings such that the reader is able to put them into a wider context in the discussion section 9 when the whole narrative of the paper comes together.

We start with a notation remark: The abbreviations of the previous chapters apply but we add an abbreviation for the model specification after the underline where relevant. For example *svr_poly* refers to the Shapley value regression modeling approach with the polynomial regression specification as in equation 15. We should also strongly point out that the curve fitting (where a logistic function is fitted to the contribution scatter in the media spend - sales plane as described in section 5.3) was not always successful. This implies that the curve fitting did not converge in each iteration for all the approaches which implies leads to unequal sample sizes. This, of course, might manifest itself in the distribution of the presented key measures. The failure rates are presented in figure 19 in the appendix A. The curve fitting procedure did never converge for the multiplicative models

[9]Still, a misspecified model can lead to similar consequences as omitted variables.

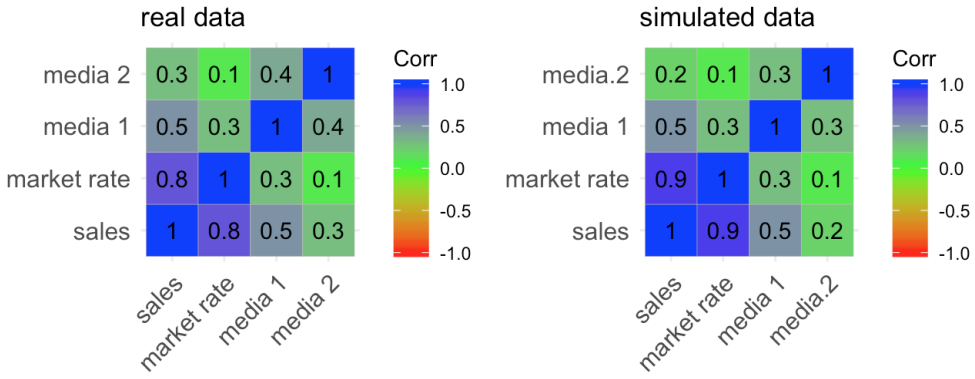

Figure 13: Similar correlation structures of the real and simulated data.

in combination with the weighted factor decomposition. This is because the multiplicative specification reflects the highest order of interaction between the features but each factor contribution computed by WFD is only based on one particular feature combination. The contributions are thus very much depending on the instance and the respective realized factor levels which introduces a large dispersion in the scatter plot leading to failing convergence. As a consequence, the reader will not find results for ols_multi and svr_multi in combination with WFD.

Marketers usually select the appropriate model with the help of some information criterion. But a good model fit does not necessarily translate into a good ROI-curve fit. For example, a linear model might well approximate the sales process but, of course, completely fails to capture shape effects. However, an alternative information criterion which allows the modeler to select the model according to the goodness of fit concerning ROI-curves can not exist because the real curves are counterfactual. Table 3 reports the R2 measures computed over the relevant data sample *window 1*. The multiplicative models (ols_multi and svr_multi) fit the sales data worse than the alternative specifications. The modeler would select the gam model or probably the tvem model if he intends to examine structural changes over the time window (recall, that tvem allows deriving further information by looking at coefficient functions). The reader is asked to remember this model selection based on the information criteria when considering subsequent results.

Figure 14 reports the mean percentage error (MAPE) measures for each parameter of the fitted logistic curves. The MAPE is calculated based on all simulation repetitions and computed for each tuple modeling, decomposition approach and each media channel separately. Modeling approaches are color-coded and decomposition approaches are denoted with the respective symbol. The first two graphs show the full picture: The very large MAPE values indicate that some modeling approaches are far off.[10] In particular, svr_poly is the worst performing model if combined with WFD for both media channels. Generally, the lines slope downwards which indicates that the parameter Asym tends to be captured worse than scal which in turn shows higher MAPE values than xmid. Recall

[10]MAPE can be greater than 1 (compare with formula 41).

Model	*ols_poly*	*ols_multi*	*svr_poly*	*svr_multi*	*gam*	*tvem*
R2	0.92	0.86	0.85	0.87	0.93	0.91

Table 2: Model validation, R2 averaged over all 500 iterations.

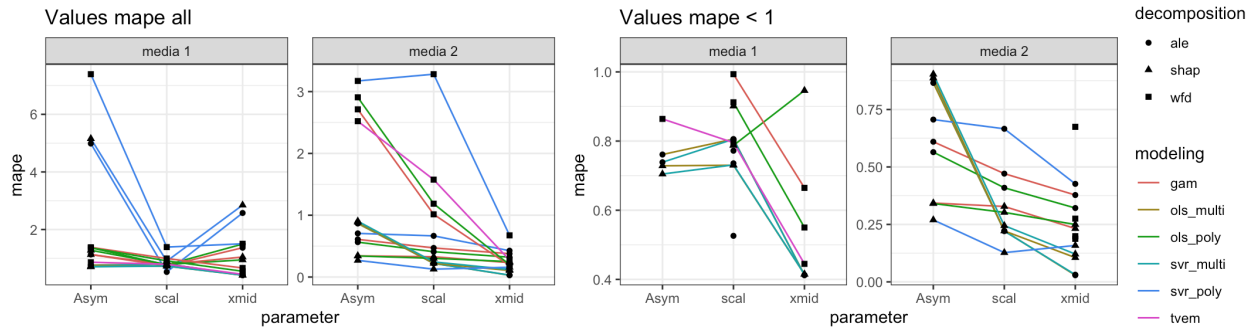

Figure 14: Parameter fit reported by MAPE values. The first row provides all MAPE values, whereas the second row considers MAPE values less than 1. The columns discriminate media 1 and media 2.

from section 5.3 that xmid represents the inflection point. The fact that xmid is the most precisely estimated parameter is somewhat encouraging because of its interpretation in MMM: The inflection point pins down the media investment, where increasing returns are replaced by decreasing returns to scale. This point on the domain is not necessarily the most relevant one for optimal media allocation (as marketers need to factor in marginal costs as seen in section 2) but still is worth knowing. In particular it represents the steepest part of the ROI-curve which corresponds to the most efficient media investment.

The residual two graphs zoom in on the best performing modeling and decomposition approaches with MAPE values being less than 1. First, the downward sloping tendency still holds. Second, it is not clear, which modeling, decomposition tuple outperforms the others. For media 1, svr_multi together with the SHAP decomposition is certainly a potential candidate. On the other hand, this methodology does not perform too well for channel 2 (which has an S-shaped response pattern). Surprisingly, svr_poly together with the SHAP decomposition is now among the best performing approaches. Recalling that the same modeling approach together with WFD was the worst choice, this hints that the decomposition has a big influence on tracing out the correct shape of the ROI-curve. The last takeaway is, that SHAP and ALE dominate WFD (compare the count of circles and triangles to the count of squares in the lower graphs). This concludes the discussion of figure 14.

A somewhat similar picture is provided by figure 15 with a focus on the distribution of the respective parameter values. The columns refer to the different modeling approaches whereas the rows are discriminated by the logistic parameters. Further, the boxplots for media 1 and media 2 are reported separately. The red dotted lines are the true parameter values.

The initial comment is about a commonality shared across media channels. All applied methodologies are biased. Also, there seems to be a bias-variance trade-off depending on the parameters and media channels: The lower the bias the higher the variance. Such is the case for channel 1 and xmid whereas for channel 2 it holds for the Asym parameter. On the other hand, this trade-off seems less pronounced for all the other parameters. The multiplicative models produce consistent estimates whereas gam in combination with SHAP and ALE is overall the least biased estimator but not very consistent. Generally, marginally lower bias comes at the cost of considerably increasing the variance. The two models gam and ols_poly perform very similarly.

When comparing media 1 to media 2 it is somewhat astonishing that the multiplicative models

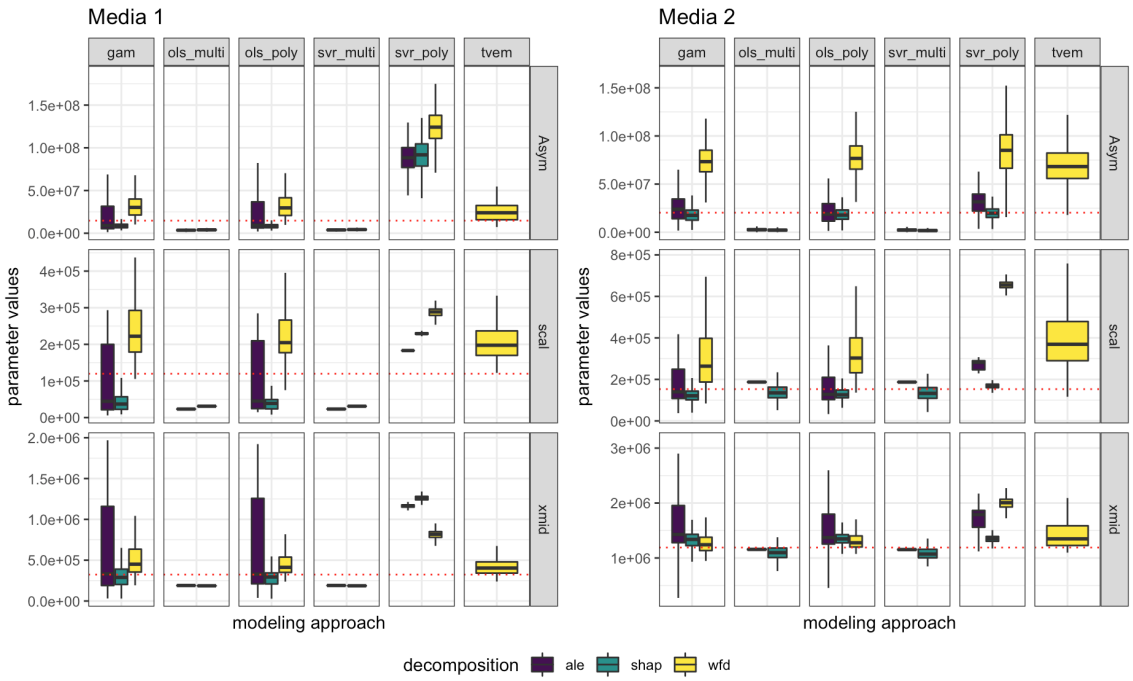

Figure 15: Parameter fit reported by boxplots. Each column reflects one modeling approach. The top three rows report on media 1.

perform better for media 2. We explain this observation by yet another trade-off: By construction, media 2 has an S-shaped response pattern (which handicaps the multiplicative models) but features synergistic effects (which benefits the multiplicative models). Recall from section 5.1 that the multiplicative specification considers the highest order of interactions among all chosen approaches.

As a concluding remark, the performance of tvem (which has been fitted to the pooled data sample, incorporating a structural change) is evaluated. The tvem model has the same specification as the polynomial regression and can hence be compared to the performance of ols_poly. The two models perform very similarly on average with a slightly lower variance for tvem at least for media 1. This has two implications: First, if the modeler is not certain whether or not a structural change occurred, the tvem methodology is among the top performers. Second and not surprisingly, the variance can most certainly be reduced by increasing the sample size. Yet, the modeler should be aware that increasing the sample size by pooling data across time comes with the trade-off between recency and relevancy.

Figure 16 provides the most intuitive insight. The figures are read left to right (columns first) and top to bottom (rows second). In each row, the results for two modeling approaches are reported. The true ROI-curves are depicted in red. The black curves are the estimated ROI-curves for each iteration. The degree to which these curves overlap is reflected by the black saturation (referred to as density). The green curve is the mean of all black curves and the dashed lines represent 90% confidence levels.

The previous discussion on parameters' MAPE measures cautioned not to draw conclusions about modeling performance from pure eyeballing: The vertical distance between estimated ROI-curve and ground truth is not the only relevant goodness of fit measure. For example, the estimated ROI-curve can perfectly match the parameters xmid and scal but completely miss the asymptote.

Yet, the asymptote might not be too important for media allocation because optimality decisions are a relative problem. This becomes evident when looking at the ROI-curves for svr_multi, media 1. Only considering the graphical illustration (figure 16) would lead to the conclusion that this modeling approach is useless. Complementing the discussion with the boxplots and MAPE measures yields the valuable insight, that the red and green curves are actually more similar than thought.

Again, the figures lead to the same conclusions as previously outlined, but provide a more graphical perspective. For example, the already mentioned bias-variance trade-off can nicely be visualized: Large dispersion of the black lines corresponds to a large spread in the boxplots. Higher density is indicative of lower variance.

By looking at the ROI-curves for media 1 the additional insight emerges, that WFD is the most heterogeneous decomposition approach. There is almost no area of high density visible in the graphs. This is not surprising as already discussed in the theoretical section 5 and in a preceding paragraph: WFD is depending on one single realization of the covariates which makes it random by construction.

Lastly, the svr_poly approach introduces a significant trade-off between the goodness of fit in media 1 versus media 2. In fact, the methodology in combination with SHAP matches the response patterns impressively for media 2 but disappoints for media 1.

We now summarize our findings. Contingent on the decomposition approach, the factor contributions can have a large spread in the media spend - sales plane. This implies, that the scatter to which a logistic curve is fitted can be very dispersed which makes estimation unstable. In particular, multiplicative models featuring a high degree of interactions in combination with WFD are problematic in that respect. Moreover, the decomposition approach matters significantly for the goodness of fit. SHAP and ALE outperform WFD across all models. SHAP might be preferable for more additive models, whereas ALE produces slightly better estimates for multiplicative models.

An information criterion (such as the R2) seems to be a good proxy for ROI-curve fit: Models with higher R2 values perform better.

Among the three parameters characterizing the ROI-curve, Asym is the worst estimated such, followed by scal and xmid. The inflection point xmid stands for the media spend where increasing returns are replaced by decreasing returns to scale and therefore reflects the steepest slope of the ROI-curve and the most efficient media investment. Still, all the proposed approaches produce biased estimates and there is a bias-variance trade-off which is not equally pronounced for all ROI-curve parameters and not shared across media channels. Most importantly, models not featuring a high degree of interactions have lower bias but high variance. On the other hand, models being able to capture interactions well, are prone to the synergy-shape trade-off (which is very specific to the chosen simulation constellation). In the presented case, the synergy out-weights shape: If media channels are likely to depict synergies, then a model reflecting a high degree of interactions might approximate ROI-curves better even though the model implied shape is not fully correct.

The last trade-off is the one between media channels: A methodology might perform very well for one channel but might estimate the shape of the other channel very poorly. Without having a clear preference for the importance of the parameter ordering Asym, scal and xmid, it is not evident which modeling, decomposition tuple dominates. For media 1 we identify svr_multi in combination with SHAP to be among the best performers whereas for media 2 we identify svr_poly with SHAP to be a strong candidate. Shapley value regression is hence an interesting methodology but might introduce additional bias under certain conditions (potentially caused by small sample size) as it

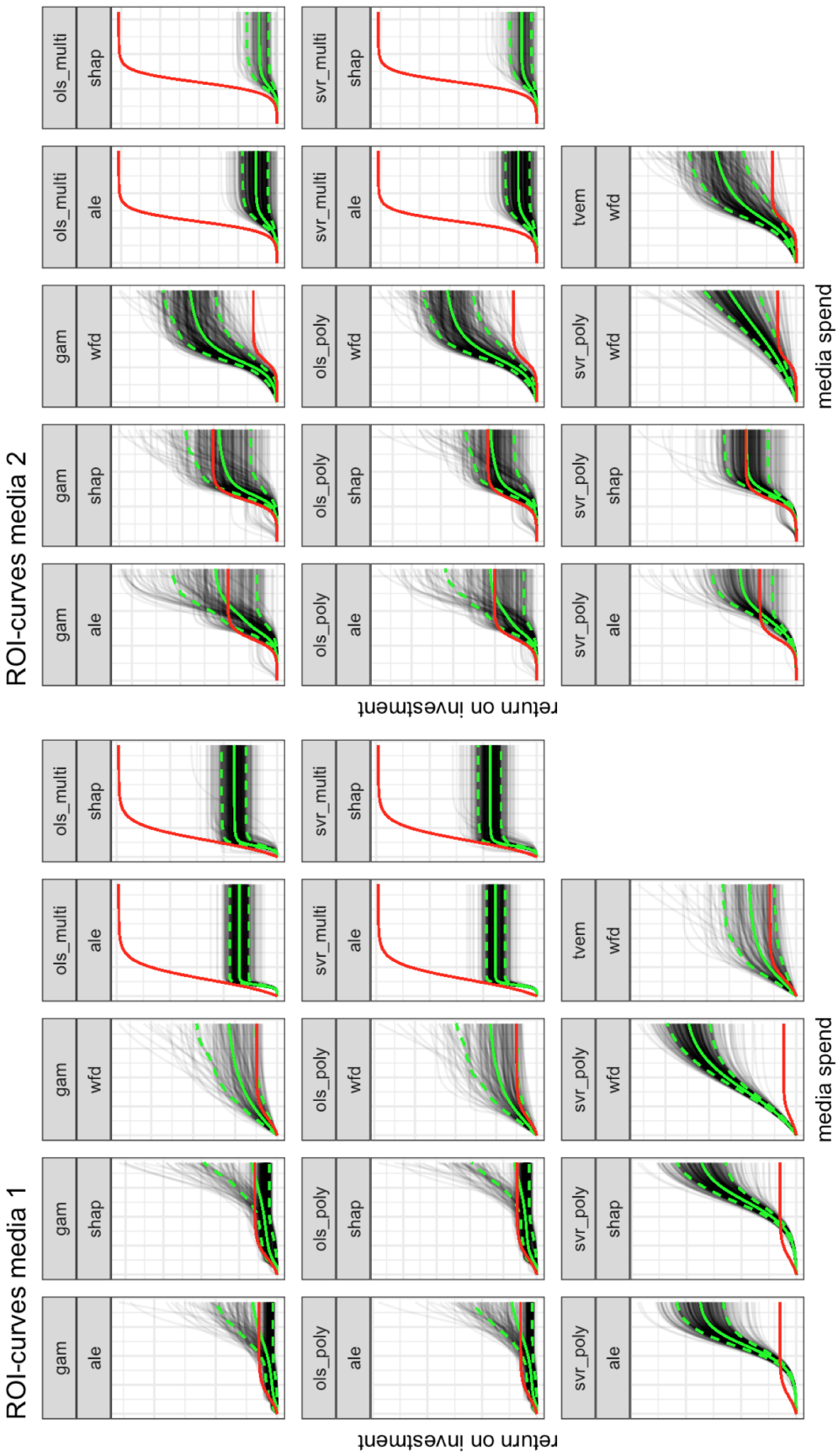

Figure 16: ROI-curves for media 1 and media 2. The red line represents the ground truth. The green line is the mean of all estimated curves and the dashed lines represent the 90% confidence interval.

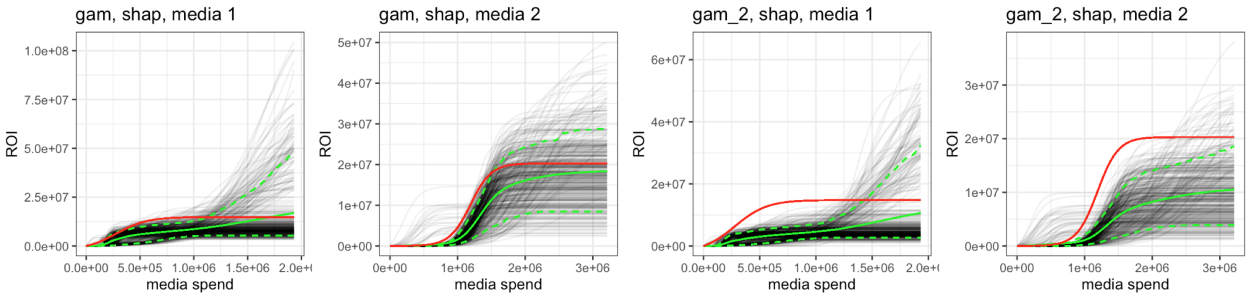

Figure 17: Recency split versus pooled data.

only aligns with the ground truth of one channel at a time.

The analysis incorporating structural change yields that the tvem methodology is a legitimate approach being among the best performing models. The tvem leverages the pooled and thus larger sample size and as a consequence reduces the variance slightly without being more biased.

8 Robustness Checks

As with real data, drawing conclusion from one analysis is never advisable. Our study is comparable to one empirical analysis conducted with one real data set. Section 6 already discussed that the simulation specification has many degrees of freedom which makes the simulated data to some extent arbitrary. But it is exactly this freedom of choice which allows us to understand MMM in a variety of scenarios. It is strongly recommended to more systematically change certain parameters in isolation to infer model performance. For example and as in Jin et al. (2017), the sample size could be increased to understand the role of limited data availability. This would be very important since the model should at least perform well in a counterfactual setting with large sample size. To increase sample size is unfortunately not trivial in the calibrated simulation setting and computation time increases considerably. This important robustness check is left to be explored in the future. Still, some obvious robustness checks will be conducted to strengthen the validity of the previous results.

These come in two flavors: First, the simulation specification is not adjusted and the gam model (one of the best overall performing models) is fitted to both observation windows (*window 1* and *window 2*). This is to say that the marketer does not consider the structural break but pools the data to increase the sample size. This allows us to further scrutinize the potential of the tvem approach on the one hand and to gauge the trade-off between dynamic change and data availability (in the specific setting) on the other hand. This is referred to as the *first* robustness check.

Results first check. Fitting the gam to the pooled data results in a reduced fit. More precisely, the R2 drops by 3 percentage points (compared to table 3). This already hints that the ROI-curve fit will diminish too. As is evident from figure 17 the response pattern does not change dramatically. Still, a deterioration is clearly visible (gam_2 is the model fitted to the pooled data). Once again, the degree of deterioration is very much depending on the magnitude of the simulated break. The simulated break was considerable with a 30 percentage point increase in media reach. Therefore, this robustness check hints strongly that the small sample size might severely constrain our ROI-curve fit.

Second, the robustness checks are conducted by altering the simulation specification. In particular, by changing the definition of the transition matrices, the demand schedules and the hill transformations which pin down the shape of the response curves. The migration behaviour and several dimensions of the medias' impact are altered as a consequence (whereas other dimensions such as media spend or media budget remain unchanged). Further, the nature of the structural break is changed more considerably. This is referred to as the *second* robustness check. Our reasoning for the new simulation specification goes as follows:

Window 1. From the original specification a synergy-shape trade-off was identified. We hence make the true response of the synergistic media channel concave and increase the synergies which should benefit the multiplicative models for media 2 considerably. To achieve this, the synergies via the common favorability transition matrix are increased and symmetry between the two channels is imposed, meaning that both channels impact the favorability dimension equally. These increased synergies and the symmetry might make it difficult for more additive models to attribute purchase behaviour to one or the other channel. The previous results are somewhat disillusioning with regard to tracing out ROI-curves accurately. One obvious limitation could be the available sample size. Arguing in the same direction, the simulation specification potentially underestimates the ability of media channels to nudge consumer behaviour. An extreme scenario, where advertising has a stronger ability to influence consumer mindsets is therefore defined. Further, the dimensions influenced by media shall provide a stronger signal which is achieved by clear discrimination of purchase likelihoods within each dimension. This implies that media can strongly nudge customers to more favorable states and given, they reach these states, after media exposure, more likely purchase the good. Generally, this should increase the ROI-curve fit for all modeling approaches. The concave shape pattern and the higher synergies benefits multiplicative models. The alternative simulation specification can be found in the appendix B.

Window 2. The results of the first robustness check, the trade-off between data availability and parameter stability is in favour of availability which yielded that the constant parameter model did not perform significantly worse compared to the tvem (see figure 17. We hence force the break not only to be in the 'intercept' but specify completely different marketing characteristics (transition matrices) and specify the hill transformation to be different between window 1 and 2. The break in the media reach is left unchanged to the original setting. Generally, this alternative specification is a robustness check concerning the tvem. To implement these propositions, the original simulation specification is employed for *window 2* and the alternative specification (as outlined above) for *window 1*.

Results second check. Most of the results derived under the original simulation specification are robust. The bias-variance and the media channel trade-offs translate to the alternative specification. But now, lower bias, higher variance only holds for the Asym parameter. As before, marginally lower bias comes at the cost of considerably increasing the variance. The multiplicative model specification still produces lower variance but again underestimates the Asym parameter. Figure 18 depicts all the methodologies (yielding MAPE values smaller than 1) under the original and alternative simulation specification. Importantly, SHAP and ALE still outperform WFD for almost all modeling approaches and parameters. The model svr_multi is again providing parameter fits with relatively small MAPE values for media 1. Surprisingly, the increased synergistic effects for media 2 under the alternative specification, does not benefit the multiplicative models. Interestingly, although the media channels are expected to provide a clearer signal under the alternative specification, the models tend to perform worse. We explain this fact by the increased synergies under the alternative simulation specification which impose a challenge for both more additive and

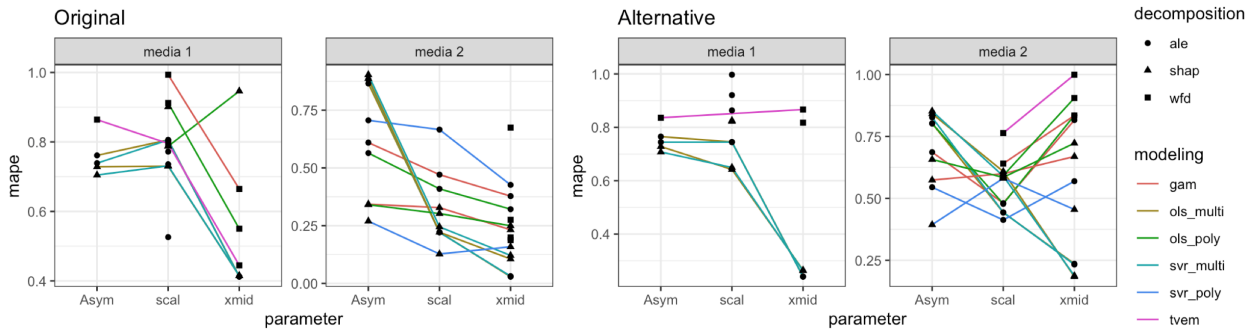

Figure 18: Parameter fit reported by MAPE values under the original and alternative simulation specification.

multiplicative models.

Further, by including window 2, we note that the gam model fitted to the pooled data now performs significantly worse. The R2 measure is only 42%. On the other hand, tvem is relatively robust with respect to the magnitude of the break and produces similar estimates to the original specification.

In summary, most of the results are robust under the alternative simulation specification. Even if media channels have a strong impact on consumer mindsets and clearly discriminate the purchase intent, MMM remains a difficult task. Surprisingly, synergistic effects impose a difficulty for both additive and multiplicative models. These interactions are hence hard to disentangle. From this perspective, a purely additive model as for example proposed by Jin et al. (2017) might proof valid. Tvem again provides relatively precise estimates even if the magnitude and nature of the break are altered. The complete output of the analysis can be found in appendix C.

9 Discussion

In this section, the results are put into context. Both insights and implications for the current MMM methodology are derived and a reference to the relevant literature is established. By doing so, several areas interesting for further research are identified. Additionally, we will start a broader discussion, where we put several encountered methodologies into a wider perspective. This should yield the insight, that the outcome of this thesis can potentially be leveraged outside of MMM.

This paper is the first one to leverage the AMSS in order to generate a realistic MMM testing framework and study shape effects. The underlying micro-founded model is able simulate synergistic effects on the micro-level which further makes it possible to incorporate media synergies in the virtual environment. To our best knowledge, all previous literature abstracts from such synergies and we therefore add this layer of complexity.

The simulation process being free from the model specification yields a fair cross-comparison between the different methodologies. This decoupling between simulation and model specification is an additional contribution and can be leveraged in further studies. Most importantly, the structure of the AMSS allows the researcher to study a problem of interest in isolation. Concerning the here presented difficulties in MMM this suggests the following agenda: Scrutinizing the sample size, the degree of multicollinearity and the magnitude of structural breaks where recency splits are superior to pooling.

This study follows the approach to test different methodologies in an as realistic as possible setting in an overriding effort to gauge MMM's capabilities to capture shape effects. Yet, methodologies should also be benchmarked in an optimal environment, free from utterly complicating factors. One could subsequently introduce complexities in a more structured manner and layer by layer in order to isolate crucial limitations. Still, the presented analysis could disentangle several complexities and hint to important implications in current MMM.

It is strongly recommend replicating the paper by Jin et al. (2017) following their Bayesian approach and testing the implications of different priors. Of course, there are several other interesting modeling approaches to be considered. Non-parametric modeling approaches such as neural networks or XGboost could prove more accurate. The here presented model agnostic methods can directly decompose such models.

An important remark should be made about ROI-curves. Once again, it is worth mentioning that the here proposed decomposition methods are concerned with additive separation. To reduce a complex problem with interaction effects to several two-dimensional spaces is clearly a misleading (but simplifying) abstraction and introduces an inconsistency in the current MMM methodology: On the one hand, the marketers pay great attention to include feature interactions in their models just to disregard them in the ROI-derivation. Our findings hint, even in the presence of synergies, additive models might not impose too severe a limitation compared to their multiplicative counterparts. On the other hand, some degree of synergies should potentially be included in the modeling specification as it might improve the model fit considerably and reduce the variance of ROI-curve estimates. This might especially hold in a more complex scenario with multiple marketing channels. The results hint that model fit is a good proxy for ROI-curve fit.

As soon as there is a multiplicative component to the model, decomposition is no longer straight forward and different reflections should enter the discussion of a fair attribution method. SHAP and ALE outperform WFD as expected. SHAP is a very general framework which fulfills some desirable properties characterizing a fair distribution. SHAP has many further interesting applications: For example, the method can be leveraged to derive synergy effects or variable importance measures. Still, if the feature space has a high degree of multicollinearity, we recommend using the ALE as it considers the conditional densities. A theory perspective for further research could hence be to derive a SHAP approach that forms coalitions and samples from the distribution taking conditional densities into account. One could think of either constraining the possible coalition space or weighting each coalition by some likelihood measure (based on the multivariate distribution).

Another approach could be to model synergies separately and include the derived interaction measures in the optimization routine (for example by once again the concept of SHAP or Friedman's H-statistic, Molnar, 2019, Chapter 5.4.2). The final output of MMM is usually a recommendation for optimal media allocation. The modeling approaches should therefore also be scrutinized in that sense after the optimization routine has been applied to the ROI-curve estimates. This potentially yields a more precise picture of the preference ordering between the different ROI parameters (Asym, scal, xmid).

Nevertheless, we prefer the picture of a single ROI-surface instead of multiple ROI-curves. Intuitively, it should be possible to formulate a game-theoretical approach to allow for coalition solutions. Let's picture a (causal) contribution-surface in three-dimensional space. Constrained optimization would be trivial. Constraining the solution space to the plane spanned by the two factors (for example media investments), one would simply be left to find the maximum of the surface over that particular area. Such a causal surface would, of course, transfer the problem

of interaction effects with other variables (not considered or influenced by a policymaker). Still, such a solution method would probably be superior to the current one and in given situations, we might want to assume that interaction effects are averaged out and can be neglected. Therefore, we propose the concept of SHMEP which stands for Shapley Multiplicative Explanation.

With regard to the time dimension, the current MMM methodology is critically depending on time consistency. This is because ROI-curves are traced out over a given time horizon. Optimally, we would observe the whole domain of the ROI-curve (media spend) in an as short time range as possible. On the other hand, MMM is prone to data availability limitations and larger sample sizes would potentially lead to more precise estimates. Therefore, the potential to pool data across time was examined and a suitable modeling approach capable of dealing with dynamic change was scrutinized. TVEM is though not a foolproof tool because it still requires constant ROI-curves (for the actual ROI-curve computation, but not the model estimation) and hence imposes the necessity to identify static marketing windows correctly. However, TVEM can be estimated leveraging a pooled data sample. Further, TVEM allows the marketer to derive valuable insights from the coefficient functions which in turn helps to identify such stable marketing environments. Therefore, TVEM alleviates the problem of limited data availability, gives additional insights to the marketers, and helps to identify relevant modeling windows. Taken these propositions together with the simulation results, TVEM should enter the toolbox of MMM.

To derive coefficient functions for various marketing channels with the help of TVEM could be its own research interest. For example, coefficient functions could be estimated using different product, brand or category segments. Are time patterns shared across these different segments? Do marketing efficiencies depict considerable evolving patterns or fluctuations? Can local extreme points be predicted such that marketers could time advertising efforts?

This thesis was very much written from a methodological and exploratory perspective with the example of MMM. However, the here described modeling approaches can easily be applied to other relevant subjects. In economics, empirical discussions are mostly reduced to a marginal perspective. All else equal, what happens with y (say poverty) if we marginally increase x (schooling)? Of course, such a question is very legitimate given that we often can not pick optimal allocations in real-life problems, but are forced to gradually move in one or the other direction from the actual circumstance. On the other hand, one might be interested in how factors contribute to the final outcome. To take up the above example, we might be interested to find out whether or not schooling is an important factor when alleviating poverty (from a global and not marginal perspective). Manna et al. (2012) performed global factor decomposition leveraging the Shapley value approach. Applying the insights of this thesis, the researchers could go one step further and locally decompose poverty in its contributing factors thereby getting a more precise picture of input-response functions and more detailed views on how certain variables shape poverty (or whatever other response). As a consequence or even necessity, it might prove relevant to depart from the predominantly linear modeling approaches and approximations in economics. This gives rise to the notion of causal machine learning which might conquer policymaking in the near future.

Recall, that the specification of media transition matrices is the only unknown input in the AMSS model. Given such knowledge would exist the AMSS framework could be leveraged as a micro-founded media mix model. Such a model would allow for direct counterfactual experiments after successful calibration. In particular Nepa's experience in consumer tracking, transition matrices should be recoverable (at least for digital marketing campaigns). The AMSS could be extended to allow for more complex environments, such as non-linear demand schedules and a more realistic competitive market behaviour should be incorporated. Such an approach to MMM would clearly

be both unique and powerful. Given all the encountered difficulties of the current methodology, a novel micro-founded approach has the potential to disrupt the industry.

Yet another critique concerning MMM is the inability of conventional models to distinguish between short-term and long-term marketing effects (Cain, 2008). Cain (2008) argues that short-sighted price campaigns while boosting demand in the short-term most likely erode long-term profitability by shifting the supply curve down to sub-optimal levels. Therefore, a modeling approach paying tribute to both short- and long-term consequences might be more suitable. To mention one possible approach, Vector Error Correction Models (VECM) could be leveraged to understand the long-term equilibrium relation between sales and marketing efforts (base sales). The incremental sales can be computed following standard procedures as currently employed in MMM and this thesis (incremental sales). But now, the marketer can compare the short-term gains to long-run implications. Of course these long-run implications (output from the VECM) need to be appropriately discounted such that the marketer can compare the present value of base sales to incremental sales (Cain, 2008). Clearly, this thesis is concerning the short-run. However, short- and long-run analysis is not exclusive but can complement one another and provide a more holistic viewpoint of ROI and hence a more nuanced understanding of marketing.

We are convinced that people claiming the death of MMM do so because they became aware of one or the other problem and assumption as encountered in this thesis. Of course, the aim of MMM seems in light of these complexities ambitious. Yet we believe that this is also the misconception of statistics being a precise science such as mathematics or physics. However, marketers should not allow trust in MMM to erode. Therefore, it might be advisable to clearly communicate the complexities and simplifying assumptions. Further, MMM should potentially reduce its ambitions and mainly provide directional guidance and restrict its output to the derived empirical facts rather than the exact final media allocation where biases potentially get aggregated. This could lead to more frequent exchange between businesses and marketers and gradual convergence to the optimal allocation after several reevaluations.

10 Conclusion

MMM tries to disentangle and understand the drivers behind KPIs such as sales. The aim is to measure media effectiveness and infer an optimal media allocation. Ultimately, MMM is concerned with causal inference. The usual metric to measure media impact is known as return on investment (ROI) which pins down the expected incremental sales for each level of media spend. Theory suggests that these ROI-curves are either linear, concave or S-shaped where the curvature can be linked to constant, increasing or decreasing returns to scale. An S-shaped pattern might arise if initial low media spending gets drowned in noise and has no effect at all whereas potential customers get saturated at higher exposures. The exact shape of the ROI-curve matters for optimal media allocation since at the optimum marginal returns should equal marginal costs for each media channel.

The usual approach to MMM is a multi-step procedure, where, as a first step, the sales-generating process is modeled. It is important to understand that the modeling approach can constrain the potential shape pattern because ROI-curves are the implied model response curves. To trace out these curves, one has to decompose the model. Decomposition is non-trivial when variables interact with each other and we employ modern model agnostic techniques to fairly attribute the contribution to the different factors. Concretely, three static modeling approaches and three

decomposition methods are investigated. The models are Ordinary Least Squares (OLS), Shapley Value Regression (SVR) and Generalized Additive Models (GAM). The decomposition methods are Weighted Factor Decomposition (WFD), Accumulated Local Effects (ALE) and Shapley Additive Explanations (SHAP). Models requiring a functional form are either specified as a polynomial or a multiplicative regression where the latter is known as the multiplicative power model. The power model is expected to capture the highest order of interaction between the variables.

We further investigated the potential to pool data thereby increasing sample size. Pooling data is accompanied by the potential threat of structural change and thus parameter instability. The modeling technique known as Time-Varying Effect Modeling (TVEM) is expected to handle smooth dynamic changes well and should account for changing media efficiencies.

These methodologies were selected because MMM is prone to some inherent complexities: The framework needs to be able to allow for causal modeling and for S-shaped response functional forms. Optimally, the model can account for a high degree of multicollinearity (MC) in the feature space as otherwise, estimated ROI-curves can be very unstable. SVR alleviates the issue of MC utilizing a coalition game-theoretic approach. Response curves are also biased if the modeler does not correctly specify interactions when synergies between media channels (also known as funnel effects) are present. We therefore included interactions explicitly in our model specifications. On the other hand and as mentioned above, multiplicative models should be able to account for complex synergistic media behavior. Such multiplicative models might be highly relevant in practice since the explicit interaction specification becomes messy when marketers employ many channels at the same time.

Potentially the biggest difficulty is imposed by limited data availability. The typical sample size incorporates three years of weekly data which leaves 154 data points. There are several strategies to alleviate these data limitations, one of them being hierarchical Bayesian methods as proposed by scholars. We followed the approach to pool data across time and identified TVEM to be a promising method because of its theoretical properties. In particular, the modeler is not required to know the exact time dynamic process and does not need to specify a transition equation.

On the other hand, decomposition methods were selected mainly because of their theoretical properties where WFD can be perceived as an intuitive benchmark. SHAP is the only approach fulfilling the desirable properties of local accuracy, missingness, dummy, symmetry and linearity whereas ALE takes conditional densities into account. The latter consideration might proof highly relevant when features are correlated.

To build a realistic testing environment, we leveraged a micro-founded demand model known as aggregate marketing system simulator (AMSS), a model developed in a google research project. The AMSS allows to generate aggregate sales time-series data by simultaneously controlling for media behavior. It further decouples the simulation specification from the model specification which enables a fair comparison of different methodologies. The AMSS sales model was calibrated to real data in an effort to match certain key metrics and therefore build a realistic marketing environment. This virtual environment is characterised by probability distributions and hence probabilistic in nature. Therefore we conducted our analysis by means of a Monte Carlo simulation study with 500 repetitions.

Our research question read *How can shape effects in MMM be captured most accurately?* This being a thesis in economics, we should give an economic answer: Well, it depends. Given all the complexities characterizing a typical MMM environment, it would be naive to believe that there exists a one-size-fits-all model. Our focus therefore lied on understanding the (model specific) trade-

offs imposed by the different complexities. The following insights were derived: An information criteria (such as R2) can be used as a proxy for the goodness of ROI-curve fit. Models fitting the sales process more closely produce better shape estimates. The decomposition approach matters significantly for the goodness of ROI-curve fit. We recommend using SHAP for more additive models, whereas ALE is desirable for more multiplicative models. We attribute that insight to ALE sampling from the conditional densities and recommend extending the SHAP to sample from the conditional densities when building coalition vectors. All modeling approaches produce biased estimates which might be caused by small sample size. There is a bias-variance trade-off with additive models producing lower bias but higher variance than their multiplicative counterparts. Yet, marginally lower bias comes at the cost of a considerable increase in variance. Strong funnel effects (synergies) between media channels impose a challenge for all models, even the multiplicative models featuring a high degree of interaction between variables. Considering the media channels separately, SVR is among the best-performing methodologies. Yet, there is a trade-off between media channels, where response patterns for one channel are matched closely by a model but missed by wide margins for the other channel. GAM is the most balanced approach in that respect. The multiplicative power model produces the most consistent estimates but in particular, is not able to fit the Asym parameter correctly. This might not be too decisive for optimal allocation. TVEM can compete with the best performing models and is not too sensitive to the degree of dynamic change, as the robustness checks hint. The results are robust to alternative simulation specifications.

We believe that the AMSS is a valid tool to understand MMM and the trade-offs implied by different complexities. Our approach allows marketers to isolate and vary complicating forces in isolation. Therefore, we are convinced of the realistic virtual testing environment and recommend exploring the approach in further research. Lastly, by breaking down the multi-dimensional task of MMM in smaller tractable sub-problems and highlighting inherent complexities, we believe contributing to an understanding of MMM. Optimal media allocation remains an ambitious task but faith should not be lost. Still, both businesses and marketers should understand crucial assumptions and the different complexities thereby updating the expectations they place on MMM: Media Mix Modeling is a fact-based policy tool very much needed in the gut feel guided world of marketing.

References

Altmann, T., Bodensteiner, J., Dankers, C., Dassen, T., Fritz, N., Gruber, S., Kopper, P., Kronseder, V., Wagner, M., and Renkel, E. (2020). *Limitations of Interpretable Machine Learning*. `https://compstat-lmu.github.io/iml_methods_limitations/`.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22.

Cain, P. M. (2008). Limitations of conventional marketing mix modelling. *Admap Magazine*, 493:48–51.

Casas, I. and Fernandez-Casal, R. (2020). *tvReg: Time-Varying Coefficients Linear Regression for Single and Multi-Equations*. R package version 0.5.1.

Chan, D. and Perry, M. (2017). Challenges and opportunities in media mix modeling. Technical report, Google Inc. `https://research.google/pubs/pub45998/`.

Chen, A., Chan, D., Perry, M., Jin, Y., Sun, Y., Wang, Y., and Koehler, J. (2018). Bias correction for paid search in media mix modeling. *arXiv preprint arXiv:1807.03292*.

Demidenko, E. and Mandel, I. (2005). Yield analysis and mixed model. *Stress*, 1:2.

Greene, M. (2014). Modeling the dynamics on the effectiveness of marketing mix elements.

Grömping, U. (2006). Relative importance for linear regression in r: The package relaimpo. *Journal of Statistical Software*, 17(1):1–27.

Hanssens, D. M., Parsons, L. J., and Schultz, R. L. (2003). *Market response models: Econometric and time series analysis*, volume 12. Springer Science & Business Media.

Hastie, T. (2019). *gam: Generalized Additive Models*. R package version 1.16.1.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, pages 297–310.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

Inc., G. (2017). *amss: Agreggate Marketing System Simulator*. R package version 1.0.1.

Jin, Y., Wang, Y., Sun, Y., Chan, D., and Koehler, J. (2017). Bayesian methods for media mix modeling with carryover and shape effects. Technical report, Google Inc. `https://research.google/pubs/pub46001/`.

Larsen, K. (2015). Gam: the predictive modeling silver bullet. *Multithreaded. Stitch Fix*, 30.

Lewis, R. A. and Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, 130(4):1941–1973.

Lipovetsky, S. and Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330.

Liu, Y., Laguna, J., Wright, M., and He, H. (2014). Media mix modeling–a Monte Carlo simulation study. *Journal of Marketing Analytics*, 2(3):173–186.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Maksymiuk, S., Gosiewska, A., and Biecek, P. (2019). *shapper: Wrapper of Python Library 'shap'*. R package version 0.1.2.

Manna, R., Regoli, A., et al. (2012). Regression-based approaches for the decomposition of income inequality in Italy, 1998-2008. *Rivista di Statistica Ufficiale*, 14(1):5–18.

Mishra, S. K. (2016). Shapley value regression and the resolution of multicollinearity. *Available at SSRN 2797224*.

Molnar, C. (2019). *Interpretable Machine Learning*. `https://christophm.github.io/interpretable-ml-book/` [Accessed 10 May 2020].

Molnar, C., Bischl, B., and Casalicchio, G. (2018). iml: An r package for interpretable machine learning. *JOSS*, 3(26):786.

Pauwels, K., Currim, I., Dekimpe, M. G., Hanssens, D. M., Mizik, N., Ghysels, E., and Naik, P. (2004). Modeling marketing dynamics by time series econometrics. *Marketing Letters*, 15(4):167–183.

Pauwels, K. and Hanssens, D. M. (2007). Performance regimes and marketing policy shifts. *Marketing Science*, 26(3):293–311.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Suarab, B., Deepen, G., et al. (2014). Multiplicative marketing mix modeling. `http://learn.fractalanalytics.com/rs/fractalanalytics/images/white%20paper-%20multiplicative%20mmm%20simplified.pdf` [Accessed 10 May 2020].

Sun, Y., Wang, Y., Jin, Y., Chan, D., and Koehler, J. (2017). Geo-level bayesian hierarchical media mix modeling. Technical report, Google Inc. `https://research.google/pubs/pub46000/`.

Tan, X., Shiyko, M. P., Li, R., Li, Y., and Dierker, L. (2012). A time-varying effect model for intensive longitudinal data. *Psychological Methods*, 17(1):61.

Taylor, S. and Letham, B. (2020). *prophet: Automatic Forecasting Procedure*. R package version 0.6.

Tellis, G. J. (2006). Modeling marketing mix. *Handbook of Marketing Research*, pages 506–522.

Tucci, M. P. (1995). Time-varying parameters: a critical introduction. *Structural Change and Economic Dynamics*, 6(2):237–260.

Van Heerde, H. J., Dekimpe, M. G., and Putsis Jr, W. P. (2005). Marketing models and the Lucas critique. *Journal of Marketing Research*, 42(1):15–21.

Vaver, J. and Zhang, S. S.-H. (2017). Introduction to the aggregate marketing system simulator. Technical report, Google, Inc. `https://research.google/pubs/pub45996/`.

Wang, Y., Jin, Y., Sun, Y., Chan, D., and Koehler, J. (2017). A hierarchical Bayesian approach to improve media mix models using category data. Technical report, Google Inc. `https://research.google/pubs/pub45999/`.

A Complementary Graphs

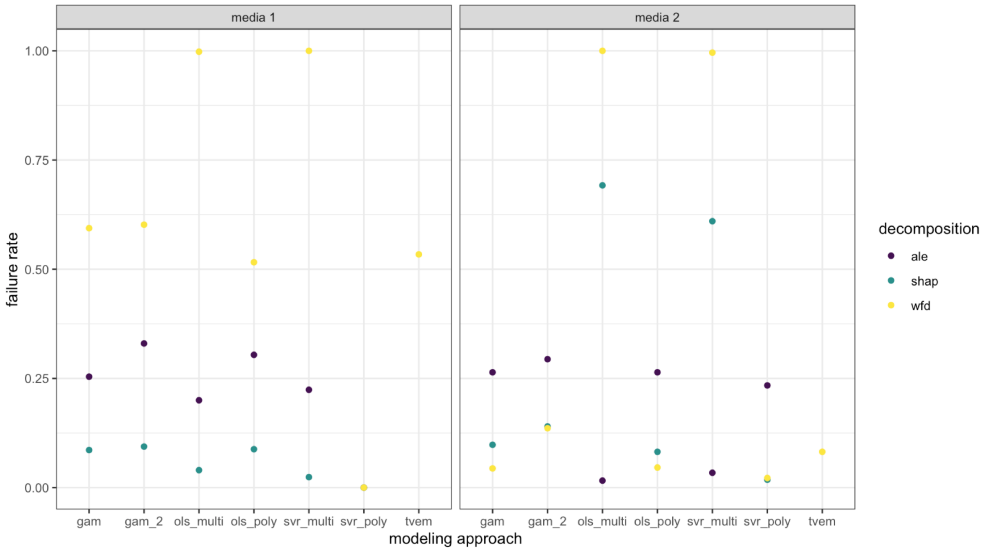

Figure 19: The failure rate is the ratio between unsuccessful curve fitting and total number of iterations (500). Multiplicative models in combination with WFD did never lead to convergence in the curve fitting algorithm.

B Alternative Simulation Specification

Marketing interventions. The transition matrices for media 1 under the alternative scenario (compare with equations 46, 47 and 48) read

$$Q^{3,\text{'activity'}} = \begin{pmatrix} 0.30 & 0.50 & 0.30 \\ 0.00 & 0.60 & 0.40 \\ 0.00 & 0.00 & 1.00 \end{pmatrix} \tag{54}$$

$$Q^{3,\text{'favorability'}} = \begin{pmatrix} 0.00 & 0.10 & 0.50 & 0.30 & 0.10 \\ 0.00 & 0.70 & 0.20 & 0.10 & 0.00 \\ 0.00 & 0.00 & 0.40 & 0.40 & 0.20 \\ 0.00 & 0.00 & 0.00 & 0.60 & 0.40 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix} \tag{55}$$

$$Q^{3,\text{'loyalty'}} = \begin{pmatrix} 0.40 & 0.35 & 0.25 \\ 0.00 & 1.00 & 0.00 \\ 0.60 & 0.00 & 0.40 \end{pmatrix} \tag{56}$$

where k was set equal to 3, indicating the chronological order of the events. The transition matrices for media 2 (compare with equations 49 and 50) read

$$Q^{4,\text{'favorability'}} = \begin{pmatrix} 0.00 & 0.10 & 0.50 & 0.30 & 0.10 \\ 0.00 & 0.70 & 0.20 & 0.10 & 0.00 \\ 0.00 & 0.00 & 0.40 & 0.40 & 0.20 \\ 0.00 & 0.00 & 0.00 & 0.60 & 0.40 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix} \tag{57}$$

$$Q^{4,\text{'availability'}} = \begin{pmatrix} 0.10 & 0.60 & 0.30 \\ 0.00 & 0.30 & 0.70 \\ 0.00 & 0.00 & 1.00 \end{pmatrix} \tag{58}$$

where k was set equal to 4, indicating the chronological order of the events.

Sales event. The segment specific intercept of the demand schedule determines the purchase likelihood of that segment. These likelihoods (compare with equations 51, 52 and 53) read

$$\alpha_{\text{'favorability'}} = (0.00, 0.00, 0.40, 0.60, 0.90) \tag{59}$$

$$\alpha_{\text{'loyalty'}} = (0.60, 0.90, 0.00) \tag{60}$$

$$\alpha_{\text{'availability'}} = (0.00, 0.70, 0.90) \tag{61}$$

The hill parameters ($hill.ec, hill.slope$) were set to $(2, 0.3)$ for media 1 and to $(3, 0.8)$ for media 2. In the original setting, we chose $(0.8, 0.3)$ for media 1 and $(3, 6)$ for media 2 which led to the shape pattern as depicted in figure 11.

C Alternative Results

Model	*ols_poly*	*ols_multi*	*svr_poly*	*svr_multi*	*gam*	*tvem*
R2	0.91	0.86	0.85	0.86	0.92	0.86

Table 3: Model validation under the alternative specification, R2 averaged over all 500 iterations.

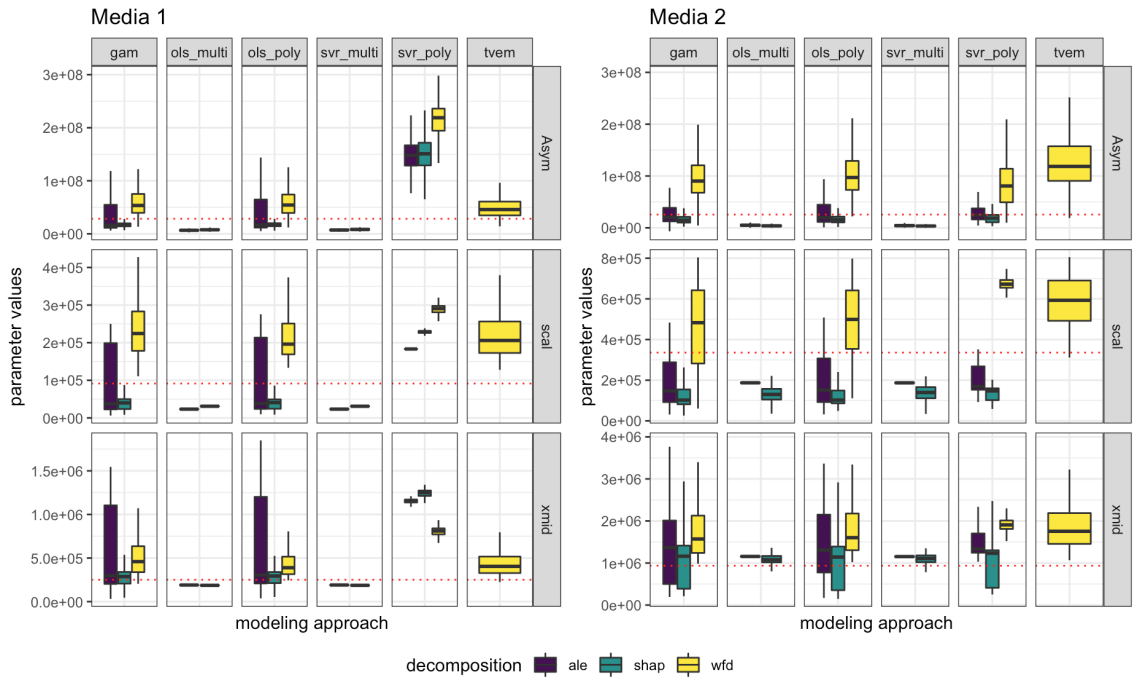

Figure 20: Parameter fit reported by boxplots under the alternative specification. Each column reflects one modeling approach. The top three rows report on media 1.

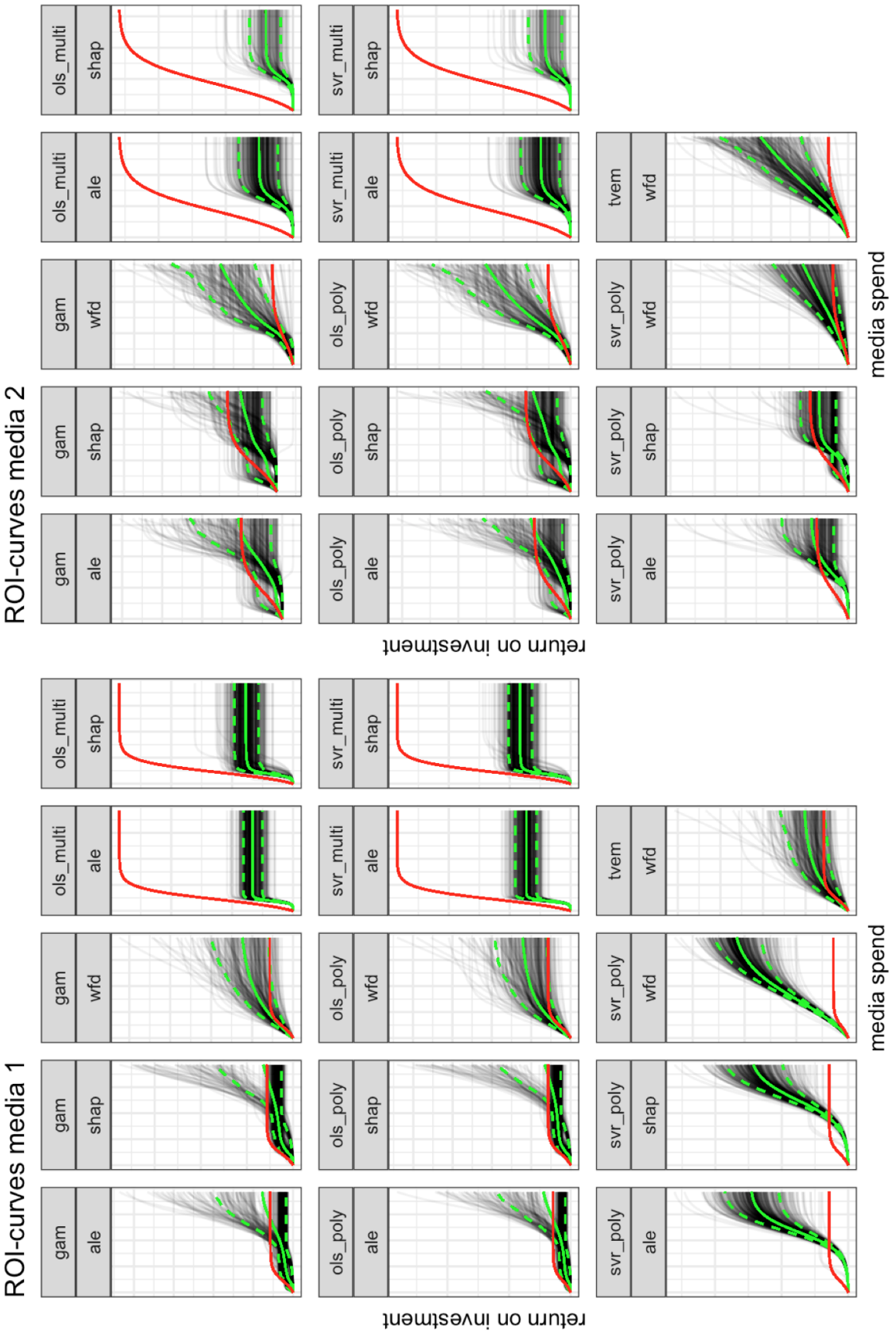

Figure 21: ROI-curves for media 1 and media 2 under the alternative specification. The red line represents the ground truth. The green line is the mean of all estimated curves and the dashed lines represent the 90% confidence interval.

D List of Resources

As mentioned throughout the text, the programming language R (R Core Team, 2013) was used and the below-mentioned packages. For further information, the reference section can be consulted.

- **amss** (Inc., 2017)
- **relaimpo** (Grömping, 2006)
- **gam** (Hastie, 2019)
- **tvReg** (Casas and Fernandez-Casal, 2020)
- **iml** (Molnar et al., 2018)
- **shapper** (Maksymiuk et al., 2019)
- **prophet** (Taylor and Letham, 2020)