



Hierarchical marketing mix models with sign constraints

Hao Chen, Minguang Zhang, Lanshan Han & Alvin Lim

To cite this article: Hao Chen, Minguang Zhang, Lanshan Han & Alvin Lim (2021): Hierarchical marketing mix models with sign constraints, Journal of Applied Statistics, DOI: [10.1080/02664763.2021.1946020](https://doi.org/10.1080/02664763.2021.1946020)

To link to this article: <https://doi.org/10.1080/02664763.2021.1946020>



View supplementary material [↗](#)



Published online: 29 Jun 2021.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



Hierarchical marketing mix models with sign constraints

Hao Chen, Minguang Zhang, Lanshan Han and Alvin Lim

Research & Development, NielsenIQ, Chicago, IL, USA

ABSTRACT

Marketing mix models (MMMs) are statistical models for measuring the effectiveness of various marketing activities such as promotion, media advertisement, etc. In this research, we propose a comprehensive marketing mix model that captures the hierarchical structure and the carryover, shape and scale effects of certain marketing activities, as well as sign restrictions on certain coefficients that are consistent with common business sense. In contrast to commonly adopted approaches in practice, which estimate parameters in a multi-stage process, the proposed approach estimates all the unknown parameters simultaneously using a constrained maximum likelihood approach and a Hamiltonian Monte Carlo algorithm. We present results on real datasets to illustrate the use of the proposed solution algorithms.

ARTICLE HISTORY

Received 11 October 2020

Accepted 16 June 2021

KEYWORDS


Marketing mix model; hierarchical models; constrained regression analysis; Hamiltonian Monte Carlo

1. Introduction

Marketing activities, such as TV advertisement, discounting, direct mail, etc., are prevailing approaches for consumer packaged goods manufactures and service providers to enhance their brand awareness and product/service messaging to consumers in order to increase sales. It is therefore of tremendous interest to measure the return of investment (ROI) of those marketing activities. However, this is by no means an easy task, especially since it is very difficult, if not impossible at all, to conduct a controlled experiment. In fact, in practice, we usually collect sales, marketing, as well as other related data, often at weekly level, and then conduct statistical analysis to relate sales quantity to various marketing activities as well as other non-marketing factors. The statistical models constructed for this purpose are known as marketing mix models (MMMs).

There are often many complications in building a MMM. First, besides being affected by marketing activities, sales volume is also affected by many non-marketing factors, such as prices, holidays, seasonality, etc. These factors, while are not of interest themselves for the purpose of understanding effectiveness of marketing activities, need to be taken into consideration to properly measure the effects of marketing activities. Secondly, different marketing activities induce very different responses, which is technically more challenging. For instance, some marketing activities, such as promotional discounting, typically

CONTACT Hao Chen  hao.chen@stat.ubc.ca  Research & Development, NielsenIQ, Chicago, IL 60606, USA

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/02664763.2021.1946020>

prompt an instant consumer response that vanishes as the activities end. Other marketing activities, such as TV advertising, may not elicit an immediate consumers response, but the carryover effect of the marketing activities might last beyond the active marketing period. These differences need to be captured in any applicable statistical models. Thirdly, the responses to any marketing activities are intrinsically heterogeneous along dimensions such as geography and product. For instance, the effect of a national TV advertisement may vary from one region to another due to geographical and demographic differences. It is essential to capture such heterogeneity in any applicable MMMs. Fourth, there often exists some sort of prior belief regarding the coefficients to be estimated. For example, while some marketing activities may not be effective, seldom do they have a negative impact on sales. Mathematically, these kinds of belief are typically translated to linear inequality constraints on the coefficients, with sign constraints being probably the most common ones. In this paper, we present a comprehensive MMM that incorporates all the aforementioned considerations.

With all the complications discussed above, the resulting MMM often features nonlinear transformations with unknown parameters as well as inequality constraints on the parameters. Such a statistical model is certainly challenging to estimate. In practice, the estimation is often accomplished in multiple steps. For example, the practitioners often first estimate the parameters involved in the nonlinear transformations and then estimate the coefficients, followed by an adjustment process to ensure that the coefficients satisfying the required constraints. These multi-step process is not only complicated to implement and automate but could also lead to inaccurate estimation of the coefficients resulting in incorrectly measuring the effects of marketing activities on sales. Therefore, in this paper, we present a more systematized approach that allows us to estimate all the unknown parameters simultaneously, while ensuring that all the constraints are satisfied.

The rest of this paper is organized as follows. In Section 2, we lay out a detailed discussion on the features of marketing mix models and then provide a literature review. In Section 3, we present model specifications, including details on how to capture carryover, shape and scale effects. In Section 4, we present our estimation approaches. Results from some numerical studies and analyses on a real dataset are reported in Sections 5 and 6, respectively, followed by concluding remarks in Section 7.

2. Marketing mix models

As was mentioned earlier, different marketing activities often generate different responses from consumers. Among all the marketing activities, advertisements are the ones that introduce unique challenges. The reasons are twofold. First, advertisements typically generate long lasting but decaying effects that go beyond the time period of active advertisement. Therefore, when we study the response from the advertisements from week to week, it looks as if a portion of the investment from previous weeks still generate response in the current week. This carryover phenomenon is known as ‘adstocking’ in marketing practice [1]. More specifically, we typically use targeted rating points (TRPs) [22] to measure the level of activity for advertisements. We study a period of w consecutive weeks, labeled by $t = 1, \dots, w$. We denote the TRP of an advertisement in week t by x_t . Due to

the carryover effect, the *effective* TRP in week t is given by:

$$\tilde{x}_t = c(x_1, \dots, x_t; \theta),$$

where θ is an unknown parameter. Note that the carryover effect from weeks earlier than the study period can be considered similarly, but we ignore such effects for simplicity of demonstrating our approach. In this paper, we consider a specific format of $c(x_1, \dots, x_t; \theta)$ given by

$$c(x_1, \dots, x_t; \theta) = \sum_{\tau=0}^{\ell-1} \alpha^\tau x_{t-\tau}, \quad \forall t = \ell, \dots, w; \quad (1)$$

with $\alpha \in (0, 1)$, referred to as the decay rate hereafter. This is to say that the carryover effect becomes negligible after ℓ weeks, decays by an unknown constant factor α each week, and is additive. In practice, one either determines the decay rate using rule-of-thumb based on experience, or estimate α in a pre-processing step before the effectiveness of the marketing activities are estimated. Ideally, we should let the data speak for itself and estimate the decay rates together with marketing effectiveness simultaneously.

Another level of complexity regarding advertisements is that the effects are in general nonlinear. Specifically, it is widely recognized that all advertisements are subject to a so-called ‘saturation’ phenomenon. Generally speaking, saturation refers to the fact that while the response still increases as the TRP increases, the rate slows down as the advertisement TRPs continue to increase. This is because of the fact that the targeted population exposed to the advertisement is finite. We use response functions to mathematically link the effectiveness of advertisements and TRPs. Due to the saturation phenomenon, a response function is typically either a C-shape (concave increasing) or S-shape (non-concave increasing) as illustrated in Figure 1. We propose to use the cumulative distribution function (CDF) of Weibull distribution to capture these two different shapes. The Weibull CDF, taking two parameters, is given as follows,

$$s(\tilde{x}; \lambda, k) = 1 - \exp\left(-\left(\frac{\tilde{x}}{\lambda}\right)^k\right). \quad (2)$$

The response function of the advertisement is therefore

$$r(\tilde{x}; \beta, \lambda, k) = \beta s(\tilde{x}; \lambda, k), \quad (3)$$

with β being the unknown coefficient of effectiveness, and λ, k being unknown nonlinear transformation parameters. We refer to λ and k as the shape and scale parameters, respectively, hereafter. Note that all α, k, λ require estimation in practice.

We use a Weibull distribution CDF in (2) to capture the ‘saturation’ phenomenon in advertisements due to its flexibility in representing both C-shape (when $0 < k \leq 1$) and S-shape (when $k > 1$) curves. Mathematically speaking, Equation (2) is a functional transformation, and it is not intended to be interpreted as a probability distribution. Neither has it any connection to prior and posterior distributions that will be discussed in Section 4. In fact, researchers have used Weibull CDF function to model ‘saturation’ characteristics in the literature, for example [3] used it to model wind turbine power, and [18] applied it in biology to study the growth of *Xanthobacter* (a type of bacteria) over time.

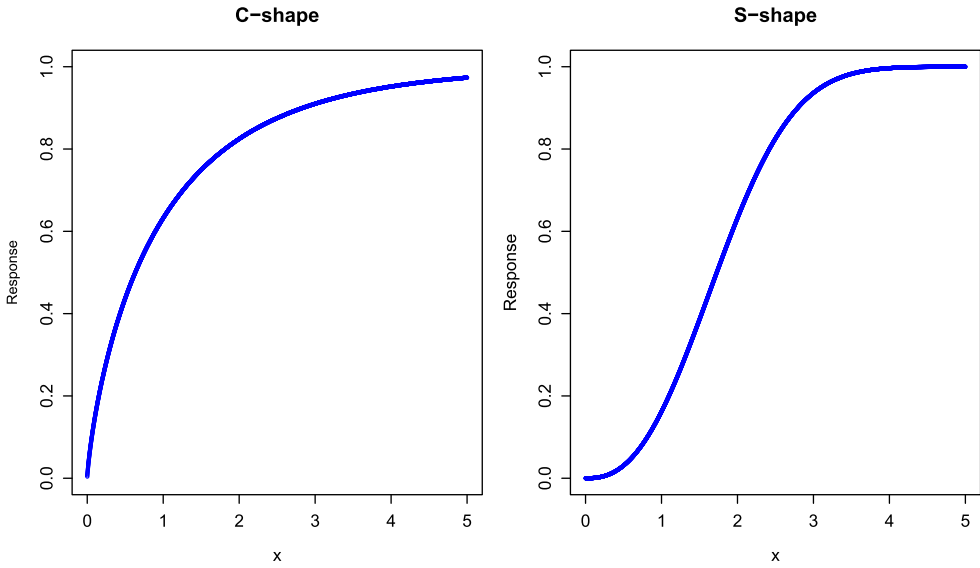


Figure 1. Illustrating plots of C-shape and S-shape.

The third layer of complications comes from the common belief that the advertisements, while may be completely ineffective, at least will not affect sales negatively. Mathematically, this can be translated to an inequality constraint, i.e. $\beta \geq 0$. Traditionally, statistical estimations are often unconstrained or under only equality constraints. The inclusion of inequality constraints impose significant challenges, especially under a hierarchical structure, which we will elaborate in the next paragraphs. In fact, most commercially available statistical software packages do not allow us to explicitly impose inequality constraints on the parameters to be estimated. Therefore, practitioners often need to apply some heuristics to ‘correct’ the signs after the coefficients are estimated. In the propose approach, we explicitly impose these constraints, and therefore no ad-hoc ‘corrections’ are needed after estimation.

The fourth layer of complications lies in the fact that there is intrinsic heterogeneity along geography and product dimensions. For example, the response to an advertisement can vary from one geographical region to another, and hence so do the coefficients of effectiveness. In the meantime, we often believe that those coefficients, while different from each other, behave like having a common coefficient adjusted by a random coefficient following a Normal distribution with 0 mean and unknown variance. This comes under the framework of mixed effect models, which will be discussed in the following sections. Mathematically, let ν denote different geographical regions indexed by $\nu = 1, \dots, g$. For each region ν , the coefficient of effectiveness $\beta_\nu \stackrel{iid}{\sim} \mathcal{N}(\beta, \eta^2)$, with $\mathcal{N}(\beta, \eta^2)$ referring to a Normal distribution with mean β and variance η^2 .

With all the added tiers of complexity, the MMM is a highly challenging statistical model to estimate. In this research, we discuss learning the unknown parameters from both a frequentist perspective via maximum likelihood estimation (MLE), and Bayesian viewpoint using the Hamiltonian Monte Carlo (HMC) approach [16]. HMC is a variant of the traditional Metropolis-Hastings algorithm [7], which belongs to the family of Markov chain

Monte Carlo (MCMC) algorithms. The benefits of HMC over the Metropolis-Hastings algorithm will be discussed in Section 4.

The research on MMMs dates back to the 1960's. Neil H. Borden coined the term 'marketing mix' in the late 1940's and laid out a conceptual framework in Borden [4] by extending the original 4Ps (Product, Price, Place, Promotion) of marketing to the 12 elements of marketing mix. Some early developments in the 1970's can be found in Lambin [12] and Little [13]. Marketing mix models became widely known after being included in the classical textbook *Basic Marketing: A Managerial Approach* [14]. Traditionally, marketing mix modeling is implemented with regression analysis using a frequentist paradigm via maximum likelihood estimation.

Since the 21st century, marketing mix modeling has received renewed interest due to the emergence of advertising channels such as paid search, digital coupons, etc., as well as progresses made in statistical methods such as Bayesian inference using MCMC approach and computation facilities allowing large-scale parallelization such as the use of graphical processing units (GPUs). The research and challenges are summarized in a recent survey [6]. This paper largely adopts the framework by Jin *et al.* [11], while expanding it to incorporate heterogeneity in marketing response [21] as well as allowing sign constraints on the coefficients to be estimated. In more recent years, in addition to the study of the marketing mix model itself, researchers have successfully applied it to different aspects of business and politics. Salman *et al.* [20] utilized it to study Egypt's hospitality industry after the Arab spring. Insights from the model can be helpful for the Egyptian hospitality industry to maintain its competitive position. Chowdhury and Naheed [8] applied it to an empirical investigation on how rural and urban voters differ in their preference of political candidates in a developing country. In addition, it was modified by Pantano *et al.* [17] to include social media participation to study how social media platforms influence customers' purchasing behavior.

3. Hierarchical marketing mix model

In this section, we provide detailed statistical models for marketing mix modeling. We will begin with the base model and then introduce hierarchical structure as well as constraints.

3.1. Base model

In this section, we present the base marketing mix model without hierarchical structure. Without loss of generality, we assume there are d independent variables in total, and the first m variables, denoted as $x_i, i = 1, \dots, m$, have carryover, shape and scale effects. The remaining $n = d - m$ variables are nuisance variables (representing non-marketing factors), denoted as $z_j, j = 1, \dots, n$. The dependent variable is the sales quantity (possibly transformed) denoted as y . Observations have been collected from w consecutive weeks, ordered chronologically, and indexed by $t = 1, \dots, w$. The available dataset is depicted in Table 1. In the dataset, $x_{t,i}$ is the organic value of independent variable x_i in week t without considering the carryover, shape and scale effects at week t . As we have discussed in the previous section, we let α_i , k_i , and λ_i be the decay, shape, and scale parameters, respectively, of variable x_i . For simplicity, we assume the maximum carryover period ℓ is known and the same for all different marketing campaigns. In practice, ℓ can be chosen to be large

Table 1. Available data from w weeks for the base model.

y_1	$x_{1,1}$	$x_{1,2}$	\cdots	$x_{1,m}$	$z_{1,1}$	$z_{1,2}$	\cdots	$z_{1,n}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
y_w	$x_{w,1}$	$x_{w,2}$	\cdots	$x_{w,m}$	$z_{w,1}$	$z_{w,2}$	\cdots	$z_{w,n}$

enough so that carryover effects beyond ℓ weeks are negligible. After taking into account the net effect of carryover, shape and scale transformations on each independent variable x_i , $i = 1, \dots, m$ in week $t = \ell, \dots, w$, the response function becomes:

$$r(x_{t-\ell+1,i}, \dots, x_{t,i}; \beta_i, \lambda_i, k_i, \alpha_i) = \beta_i s_i(c_i(x_{t-\ell+1,i}, \dots, x_{t,i}; \alpha_i); k_i, \lambda_i), \quad (4)$$

where $c(\cdot)$ and $s(\cdot)$ are defined in (1) and (2), respectively. Therefore, the overall base model is given by:

$$y_t = \sum_{i=1}^m \beta_i s_i(c_i(x_{t-\ell+1,i}, \dots, x_{t,i}; \alpha_i); k_i, \lambda_i) + \sum_{j=1}^n \gamma_j z_{t,j} + \epsilon_t, \quad (5)$$

for all $t = \ell, \ell + 1, \dots, w$, where $\epsilon_t \sim N(0, \sigma^2)$ and is independent for all $t = \ell, \ell + 1, \dots, w$. The unknown parameters α_i 's, k_i 's, λ_i 's, β_i 's, γ_j 's as well as σ^2 require estimation in practice. Note that one of the γ_j 's can be an intercept.

3.2. Extension to the hierarchical model

The model in (5) represents a linear model after the carryover, shape and scale effects are considered. To account for heterogeneity along different geographical dimensions, it is often necessary to incorporate hierarchical structures, which leads to general linear hierarchical models. We are particularly interested in hierarchical models with mixed effects, in which some or all of the independent variables have a hierarchy to account for the heterogeneity across sub-populations such as different regions using random coefficients. Moreover, we also propose to have sign constraints on some of the coefficients to be consistent with our business knowledge and common sense. For example, the coefficients for marketing activities should, in general, be non-negative. Compared to Table 1, we further assume that the data contain an additional layer of regions, indexed by $v = 1, 2, \dots, g$. The available data for hierarchical modeling is given in Table 2. We let \mathcal{H}_β and \mathcal{H}_γ be the sets of indices of β -variables with sign constraints and indices of γ -variables with sign constraints, respectively. We define $\overline{\mathcal{H}}_\beta$ and $\overline{\mathcal{H}}_\gamma$ be the complements of \mathcal{H}_β and \mathcal{H}_γ in $\{1, \dots, m\}$ and $\{1, \dots, n\}$, respectively. Without loss of generality, we assume all the sign constraints are nonnegative constraints. The hierarchical model is given below.

$$\begin{aligned}
y_{t,v} = & \sum_{i \in \mathcal{H}_\beta} \beta_{i,v} s_i(c_i(x_{t-\ell+1,i,v}, \dots, x_{t,i,v}; \alpha_i); k_i, \lambda_i) \\
& + \sum_{i \in \overline{\mathcal{H}}_\beta} \beta_{i,v} s_i(c_i(x_{t-\ell+1,i,v}, \dots, x_{t,i,v}; \alpha_i); k_i, \lambda_i) \\
& + \sum_{j \in \mathcal{H}_\gamma} \gamma_{j,v} z_{t,j,v} + \sum_{j \in \overline{\mathcal{H}}_\gamma} \gamma_{j,v} z_{t,j,v} + \epsilon_{t,v}, \quad \forall t = \ell, \dots, w; v = 1, \dots, g
\end{aligned}$$

Table 2. Available data from g regions. Each has w weeks of marketing data.

$v = 1$	$y_{1,1}$	$x_{1,1,1}$	$x_{1,2,1}$	\cdots	$x_{1,m,1}$	$z_{1,1,1}$	$z_{1,2,1}$	\cdots	$z_{1,n,1}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$v = 1$	$y_{w,1}$	$x_{w,1,1}$	$x_{w,2,1}$	\cdots	$x_{w,m,1}$	$z_{w,1,1}$	$z_{w,2,1}$	\cdots	$z_{w,n,1}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$v = g$	$y_{1,g}$	$x_{1,1,g}$	$x_{1,2,g}$	\cdots	$x_{1,m,g}$	$z_{1,1,g}$	$z_{1,2,g}$	\cdots	$z_{1,n,g}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$v = g$	$y_{w,g}$	$x_{w,1,g}$	$x_{w,2,g}$	\cdots	$x_{w,m,g}$	$z_{w,1,g}$	$z_{w,2,g}$	\cdots	$z_{w,n,g}$

$$\begin{aligned}
\beta_{i,v} &\stackrel{iid}{\sim} N(\beta_i, \eta_i^2), \quad \forall i = 1, \dots, m \\
\gamma_{j,v} &\stackrel{iid}{\sim} N(\gamma_j, \xi_j^2), \quad \forall j = 1, \dots, n \\
\epsilon_{t,v} &\stackrel{iid}{\sim} N(0, \sigma^2) \\
\beta_{i,v} &\geq 0, \quad \forall i \in \mathcal{H}_\beta \\
\gamma_{j,v} &\geq 0, \quad \forall j \in \mathcal{H}_\gamma
\end{aligned} \tag{6}$$

In this model, we assume the carryover, shape and scale parameters are the same across different regions, while they can vary across different marketing activities. In theory, we could also allow them to vary across sub-populations. However, this may lead to significantly enlarged parameter space and lead to identifiability issues.

Given the hierarchical model in (6), it is obvious that we have the following parameters that need to be estimated: (1) Carryover parameters: $\alpha_1, \dots, \alpha_m$; (2) Shape parameters: k_1, \dots, k_m ; (3) Scale parameters: $\lambda_1, \dots, \lambda_m$; (4) Means of fixed regression parameters: β_1, \dots, β_m ; $\gamma_1, \dots, \gamma_n$; (5) Variances of fixed regression parameters: $\eta_1^2, \dots, \eta_m^2$; ξ_1^2, \dots, ξ_n^2 ; (6) random regression parameters: $\beta_{i,v}, \gamma_{j,v}$, where $i = 1, \dots, m$, $j = 1, \dots, n$ and $v = 1, \dots, g$. In addition, the parameters are constrained such that $0 \leq \alpha_i < 1$, $k_i > 0$, $\lambda_i > 0$, $\beta_{i,v} \geq 0$, and $\gamma_{j,v} \geq 0$, for all $i \in \mathcal{H}_\beta, j \in \mathcal{H}_\gamma$. With the sign constraints, the model estimation becomes more challenging no matter which estimation approach we take. When maximum likelihood estimation is applied, the maximization problem is an inequality constrained nonlinear nonconvex optimization problem. When we adopt a Bayesian inference paradigm, the major challenge is to manage the computation efficiency as well as handling the constraints. We present details regarding parameter estimation of the proposed MMM in the next section.

4. Parameter estimation of MMMs

4.1. Maximum likelihood estimation of MMMs

To facilitate the MLE approach [5], we first examine the likelihood function. We notice that, due to the existence of sign constraints on $\beta_{i,v}$ for all $i \in \mathcal{H}_\beta$, the probability density function (PDF) of $\beta_{i,v}$ given β_i, η_i^2 should be considered as a one-sided truncated Normal

distribution. That is, for all $i \in \mathcal{H}_\beta$, it is given by

$$f(\beta_{i,v} | \beta_i, \eta_i^2) = \zeta_{\beta,i}(\beta_i) \frac{1}{\sqrt{2\pi}\eta_i} \exp\left(-\frac{1}{2} \left(\frac{\beta_{i,v} - \beta_i}{\eta_i}\right)^2\right),$$

which is the PDF of a Normal distribution $N(\beta_i, \eta_i^2)$ with a β_i -dependent scaling factor

$$\zeta_{\beta,i}(\beta_i) = \frac{1}{(1 - \Phi(-\frac{\beta_i}{\eta_i}))},$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of standard Normal distribution, i.e.

$$\Phi(\omega) = \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{\omega}{\sqrt{2}}\right)\right).$$

The dependence of the scaling factor, in fact, has profound implications. In particular, it is not correct to ignore this scaling factor when maximizing the likelihood function, and hence the likelihood function is fundamentally different to the one for traditional unconstrained linear hierarchical models. Similarly, the PDF of $\gamma_{j,v}$ given γ_j, ξ_j^2 , for all $j \in \mathcal{H}_\gamma$ is given by

$$f(\gamma_{j,v} | \gamma_j, \xi_j^2) = \zeta_{\gamma,j}(\gamma_j) \frac{1}{\sqrt{2\pi}\xi_j} \exp\left(-\frac{1}{2} \left(\frac{\gamma_{j,v} - \gamma_j}{\xi_j}\right)^2\right),$$

with

$$\zeta_{\gamma,j}(\gamma_j) = \frac{1}{(1 - \Phi(-\frac{\gamma_j}{\xi_j}))}.$$

Let Θ denoted all the unknown parameters to be estimated. With r_i analogously defined as in (4), we have

$$f(y_{t,v} | \Theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \left(\frac{y_{t,v} - (\sum_{i=1}^m r_i(x_{t,i,v}; \beta_{i,v}, \alpha_i, \lambda_i, k_i) + \sum_{j=1}^n \gamma_{j,v} z_{t,j,v})}{\sigma}\right)^2\right),$$

for $t = \ell, \dots, w$. Therefore, the joint likelihood function is given by:

$$L(\Theta) = \left(\prod_{t=\ell}^w \prod_{v=1}^g f(y_{t,v} | \Theta)\right) \times \left(\prod_{i=1}^m \prod_{v=1}^g f(\beta_{i,v} | \beta_i, \eta_i^2)\right) \times \left(\prod_{j=1}^n \prod_{v=1}^g f(\gamma_{j,v} | \gamma_j, \xi_j^2)\right). \quad (7)$$

The ML approach hence leads to the following constrained optimization problem:

$$\begin{aligned} & \max_{\Theta} \quad \ln(L(\Theta)) \\ & \text{s.t} \quad \beta_{i,v} \geq 0 \quad i \in \mathcal{H}_\beta, v = 1, \dots, g \\ & \quad \gamma_{j,v} \geq 0 \quad j \in \mathcal{H}_\gamma, v = 1, \dots, g. \end{aligned} \quad (8)$$

As we can see, the objective function in (8) is highly nonlinear and non-convex. On the other hand, the constraints are relatively simple. We can apply different optimization algorithms to solve this problem, although it is typically impossible to find a global

optimal solution of (8). Generally speaking, optimization algorithms can be grouped into three categories. In the first category, only first-order gradients of the objective function and/or the constraints are utilized. These algorithms are typically inexpensive in terms of computational time and space complexity at each iteration, but often require a lot of iterations for producing reasonable estimates. In the second category of the algorithms, the second-order gradient (Hessian) of the objective function is utilized. These algorithms are more expensive at each iteration, but often require less iterations in total. The third category is somewhere in between the first two. They aim to approximate the Hessian matrix in a less expensive way compared to obtaining the exact Hessian matrix. These algorithms typically require a reasonable number of iterations. Due to the complexity of the objective in (8) and the potential large number of variables, we opt for the third class of algorithms. In particular, we apply two different algorithms: the limited memory version of bounded Broyden-Fletcher-Goldfarb-Shanno (L-BFGS-B) [23] and the sequential quadratic programming (SQP) [2] algorithm. Numerical results will be provided in Section 5.

4.2. Hamiltonian monte carlo approach

Hamiltonian Monte Carlo is a type of MCMC approach which uses Hamiltonian dynamics to propose new random samples. Traditional Gaussian random walk Metropolis-Hastings algorithms typically use a one-dimensional Normal proposal. More specifically, let $\Theta = (\theta_1, \dots, \theta_d)^T$ be the vector of all parameters. A random walk proposed for the j -th parameter is drawn from a Normal distribution

$$\theta_j^* \sim N(\theta_j, \sigma_j^2 | \theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_d),$$

where σ_j^2 is tuned to ensure the acceptance rate is about 20–40% [19]. As we can see, at each iteration, it needs to independently propose a candidate for each variable. The drawbacks are obvious: first of all, one needs independent proposals for each variable and then combine and evaluate Θ^* collectively, usually leading to a high rejection rate and inefficiency for high-dimensional problems. Secondly, it is mathematically tedious as one needs to analytically derive the conditional posterior distribution for each unknown parameter. The remedy proposed is the HMC approach, which is able to propose a multi-dimensional candidate in one shot. To do so, we follow the so-called Hamiltonian dynamics, which was first studied by physicists, and was later borrowed by statisticians. Hamiltonian dynamics [9] describes a frictionless puck that slides over a surface of varying height. The state of the system consists of the position (given by a vector ψ) of the puck and the momentum of the puck (given by a vector ν). The potential energy, $U(\psi)$ is viewed as a function of ψ and the kinetic energy, $K(\nu) = |\nu|^2/(2s)$, where s is the mass of the puck. If the puck encounters a rising slope, the puck's momentum allows it to continue, with its kinetic energy decreasing and its potential energy increasing, until the kinetic energy is zero when it starts to slide back. Let the Hamiltonian be defined as $H(\psi, \nu) = U(\psi) + K(\nu)$. This dynamics is described by the following differential equations:

$$\frac{\partial \psi_i}{\partial t_i} = \frac{\partial H}{\partial \nu_i},$$

$$\frac{\partial v_i}{\partial t_i} = -\frac{\partial H}{\partial \psi_i}.$$

When applying the HMC algorithm, we let ψ be the vector of unknowns, i.e. $\psi = \Theta$, and let v be an auxiliary vector of the same dimension as Θ . Let $K(v) = \frac{1}{2}v^T v$. Let $U(\Theta) = -\ln(P(\Theta))$ with $P(\Theta)$ being the posterior PDF of the unknowns up to a multiplicative constant. Typically, $P(\Theta)$ is the product of prior distribution and likelihood function. We use a leapfrog procedure [16], which is an enhancement to the explicit Euler's method [10]. At a given time τ , the leapfrog method compute $\Theta(\tau + \Delta\tau)$ and $v(\tau + \Delta\tau)$ by

$$\begin{aligned} v(\tau + \Delta\tau/2) &= v(\tau) - \frac{\Delta\tau}{2} \nabla P(\Theta)|_{\Theta=\Theta(\tau)} \\ \Theta(\tau + \Delta\tau) &= \Theta(\tau) + \Delta\tau v(\tau + \Delta\tau/2) \\ v(\tau + \Delta\tau) &= v(\tau + \Delta\tau/2) - \nabla P(\Theta)|_{\Theta=\Theta(\tau+\Delta\tau)} \end{aligned} \quad (9)$$

As we can see, we start from the current v and ψ and then first updating v a half step, then the position ψ a whole step, and then finish by updating v the other half of the step. The magnitude of $\Delta\tau$ is called the step size. Note that equations (9) can be repeated for κ times, to obtain $\Theta(\tau + \kappa\Delta\tau)$ and $v(\tau + \kappa\Delta\tau)$. We then let $\Theta^* = \Theta(\tau + \kappa\Delta\tau)$ be the proposal. It is worth pointing out that $\Delta\tau$ and κ are two important user-defined parameters that one needs to carefully tune them to make the overall acceptance rate close to HMC's optimal acceptance rate 0.65 [16].

Let $\Theta^{(i)}$ collectively denote the unknown parameters at iteration i . $\Theta^* = \Theta(\tau + \kappa\Delta\tau)$ is the proposal generated by repeating the leapfrog process κ times with $\Theta(\tau) = \Theta^{(i)}$ and $v(\tau)$ being a random sample from multivariate normal distribution $N(\mathbf{0}_d, \mathbf{I}_d)$, where $\mathbf{0}_d$ is the d -dimensional all 0 vector and \mathbf{I}_d is the $d \times d$ identity matrix. The next iteration $\Theta^{(i+1)}$ is given by

$$\Theta^{(i+1)} = \begin{cases} \Theta^* & \text{with probability } p \\ \Theta^{(i)} & \text{with probability } 1 - p \end{cases},$$

where

$$p = \min \left(1, \frac{P(\Theta^*) \exp \left(-\frac{1}{2} v(\tau + \kappa\Delta\tau)^T v(\tau + \kappa\Delta\tau) \right)}{P(\Theta^{(i)}) \exp \left(-\frac{1}{2} v(\tau)^T v(\tau) \right)} \right). \quad (10)$$

We next discuss how to handle the constraints under the Bayesian framework. Under this framework, a major difference to the maximum likelihood framework lies in the fact that we need to specify prior distributions for all the unknown parameters. Those prior distributions encode our prior belief on the unknown parameters. As discussed earlier, the sign constraints on the unknowns are mathematical representations of our business knowledge regarding those parameters. Therefore, conceptually, it is natural to include sign constraints in the prior. For example, for any $i \in \mathcal{H}_\beta$ and any $v = 1, \dots, g$, let the priors of $\beta_{i,v}$ be $\pi_{\beta_{i,v}}(\beta_{i,v})$, we make sure that

$$\pi_{\beta_{i,v}}(\omega) = 0, \quad \forall \omega < 0.$$

Similarly, we can specify priors for $\gamma_{j,v}$'s for all $j \in \mathcal{H}_\gamma$ and $v = 1, \dots, g$. We also assume that the priors of the unknowns are independent. For simplicity of notations, we omit

the subscript in the prior distribution. The joint prior distributions of all the unknown parameters is given by:

$$\begin{aligned} \pi(\Theta) = & \left(\prod_{i=1}^m \left(\pi(\alpha_i) \pi(k_i) \pi(\lambda_i) \pi(\beta_i) \pi(\eta_i^2) \prod_{v=1}^g \pi(\beta_{i,v}) \right) \right) \\ & \times \left(\prod_{j=1}^n \left(\pi(\gamma_j) \pi(\xi_j^2) \prod_{v=1}^g \pi(\gamma_{j,v}) \right) \right) \pi(\sigma^2). \end{aligned} \quad (11)$$

Under Bayesian framework, the posterior distribution of the parameters is proportional to the product of the prior distribution and the likelihood function. Since we have encoded the sign constraints in the prior distribution, we will not include them in the likelihood function anymore. Therefore, we have a likelihood function different from (7) with the β and γ dependent on the scaling factor due to the truncation of the Normal distribution removed. We have

$$\mathcal{L}(\Theta) = \left(\prod_{t=\ell}^w \prod_{v=1}^g f(y_{t,v} | \Theta) \right) \times \left(\prod_{i=1}^m \prod_{v=1}^g f_N(\beta_{i,v} | \beta_i, \eta_i^2) \right) \times \left(\prod_{j=1}^n \prod_{v=1}^g f_N(\gamma_{j,v} | \gamma_j, \xi_j^2) \right). \quad (12)$$

where

$$f_N(\beta_{i,v} | \beta_i, \eta_i^2) = \frac{1}{\sqrt{2\pi} \eta_i} \exp \left(-\frac{1}{2} \left(\frac{\beta_{i,v} - \beta_i}{\eta_i} \right)^2 \right).$$

and

$$f_N(\gamma_{j,v} | \gamma_j, \xi_j^2) = \frac{1}{\sqrt{2\pi} \xi_j} \exp \left(-\frac{1}{2} \left(\frac{\gamma_{j,v} - \gamma_j}{\xi_j} \right)^2 \right).$$

And therefore we let

$$P(\Theta) = \pi(\Theta) \mathcal{L}(\Theta).$$

As we can see, $P(\Theta)$ is always 0 outside the feasible region of the optimization problem (8), and therefore any proposal that falls outside the region is not accepted according to Equation (10). On the other hand, $\mathcal{L}(\Theta)$, not involving evaluation of Normal CDF, is less complex than $L(\Theta)$ in terms of numerically evaluating its value and gradient, which is one of the most time consuming parts in the leapfrog procedure. The above treatment of the constraints guarantees that the HMC algorithm always takes legitimate samples and also mitigates the computational load.

5. Simulated examples

5.1. Example bundle 1 – base model

We first work with example bundles concerning the performance under the base model. Following the same notations, the bundle consists of the following 4 cases. (1) Case 1 : $m = 2, w = 52$; (2) Case 2 : $m = 2, w = 104$; (3) Case 3 : $m = 4, w = 104$; (4) Case 4 : $m = 4, w = 208$. $n = 1$ and $\ell = 5$ are fixed for all examples. The true model parameters

Table 3. RMSE of HMC, L-BFGS-B and SQP for the base model.

	HMC	L-BFGS-B	SQP
Case 1	0.051	0.500	0.433
Case 2	0.047	0.501	0.555
Case 3	0.077	0.532	0.521
Case 4	0.088	0.545	0.516

and prior distributions are given in Section 1 of the *supplemental document*. In addition, we apply a kernel trick on the carryover parameters, and after the logistic transformation, it works at the unbounded α_i^* scale. The scale will then be converted back to the original one before outputting the final estimates. We kicked off a run with the number of HMC iterations equaling 30,000 and the first 6,000 iterations are treated as burn-in. The thinning parameter is fixed at 300 after burn-in to ensure no significant autocorrelation is observed in the collected samples. A convergence diagnostics is conducted and reported in Section 5.3. We also include results from two constrained optimization methods: limited memory version of bounded Broyden-Fletcher-Goldfarb-Shanno (L-BFGS-B) [23] and sequential quadratic programming (SQP) [2] as comparisons. Each optimization method is repeated 30 times with different initial values, and the estimates with the largest log-likelihood is recorded.

The root mean squared error (RMSE) of the example bundles are reported in Table 3. The estimated parameters for all the 4 cases are reported in Tables 1 through 4 in the *supplemental document*. From Table 3, we observe that HMC has the smallest RMSE: Taking Case 1 as an illustrating example, the RMSE of HMC is 0.051 compared with 0.500 and 0.433 for the two optimization methods. In general, RMSEs of HMC are about 7 times smaller than those of L-BFGS-B and SQP. The performance of L-BFGS-B and SQP are similar to each other, although none is able to obtain estimates that are as accurate as HMC does.

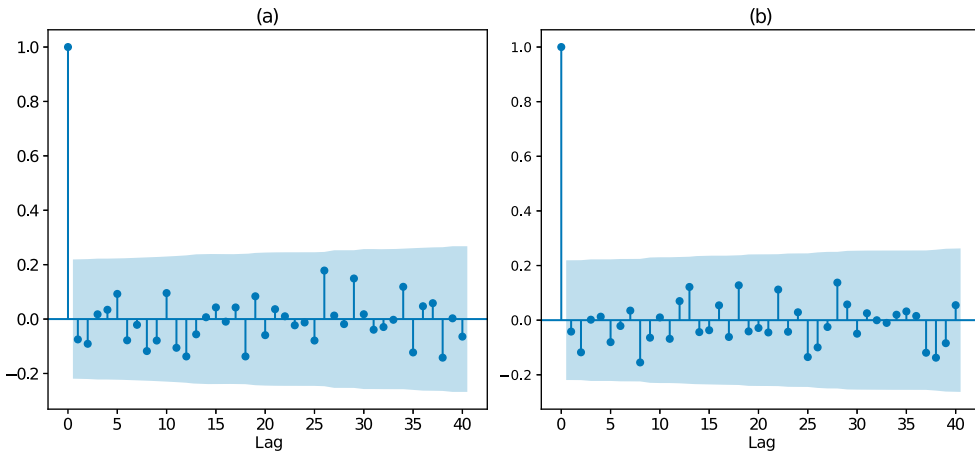
5.2. Example bundle 2 – hierarchical model

In this part, we assess the performance of the three methods under the hierarchical model. The proposed example bundle includes 4 examples below. (1) Case 5 : $m = 2, w = 52, g = 2$; (2) Case 6 : $m = 2, w = 104, g = 2$; (3) Case 7 : $m = 4, w = 104, g = 2$; (4) Case 8 : $m = 4, w = 208, g = 2$. $n = 1$ and $\ell = 5$ are fixed for all examples as before. All of the true parameters are same as in Section 5.1 except as $g = 2$, we will need to consider both the fixed effects and random effects for regression parameters. The true model parameters and priors are same as the ones used for the base model except the fixed effects and random effects are new under the hierarchical model, and they are reported in Section 3 of the *supplemental document*.

For HMC, we further increase the thinning parameter to 400 to ensure no significant autocorrelation exists. All other settings are kept the same as those in Section 5.1, we report the RMSE of the three methods in Table 4. The estimated parameters are reported in Tables 5 through 8 in the *supplemental document*. Table 4 is consistent with Table 3 that the HMC has the smallest RMSE value, indicating its superior performance over the other two optimization methods. The conclusions we draw from the base model extends to the

Table 4. RMSE of HMC, L-BFGS-B and SQP for the hierarchical model.

	HMC	L-BFGS-B	SQP
Case 5	0.061	0.252	0.425
Case 6	0.066	0.274	0.351
Case 7	0.081	0.250	0.491
Case 8	0.090	0.369	0.554

**Figure 2.** ACF plots for (a) β_1 and (b) β_2 for collected samples after applying burn-in and thinning of Case 1. The light blue areas represent 95% confidence intervals.

hierarchical model: compared to the other two optimization methods, HMC is better at recovering the ‘true’ parameters in the simulated examples.

5.3. Convergence diagnostics

We present two examples to illustrate MCMC diagnostics for justifying the results reported in Tables 3 and 4. The first example is Case 1 representing the performance of the base model. The second example is Case 5 illustrating the performance of the hierarchical model. Three different diagnostics tools are considered: the autocorrelation (ACF) plot, the traceplot and the histogram. Since there are so many parameters, we present the results of β_1 and β_2 which are the most important parameters that can be interpreted, in practice, as the effect of an advertisement on sales volume of a product. The ACF plot, traceplot and histogram are reported in Figures 2, 3 and 4, respectively.

From Figure 2, it is observed that after applying burn-in and thinning, there is no significant autocorrelation among the collected samples. In addition, Figure 3 and Figure 4 confirm that the mixing is satisfactory, and the histogram of both parameters center at the true value, 1, which further strengthen the validity of the results.

The same plots are drawn for Case 5 of the hierarchical model as well, and they are presented in Figures 1, 2, 3 in the *supplemental document* to save space of the main article.

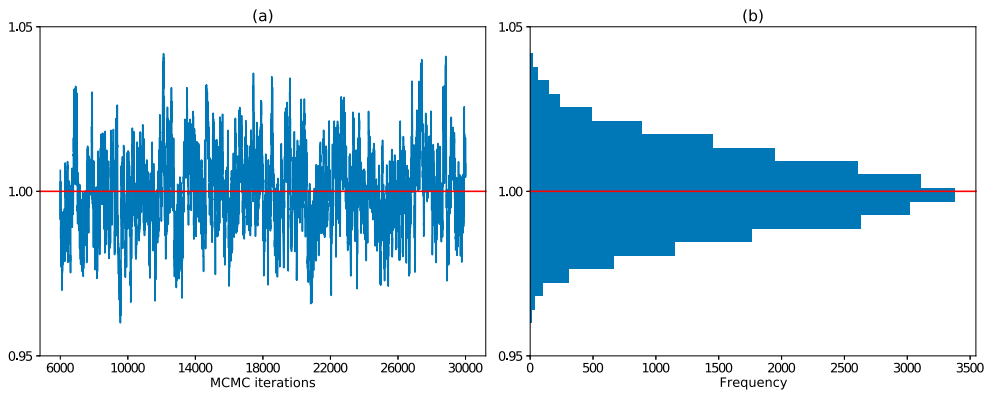


Figure 3. (a) is the traceplot of all samples after burn-in for β_1 ; (b) is the histogram of all samples after burn-in for β_1 of Case 1. The red horizontal line is the true value, 1.

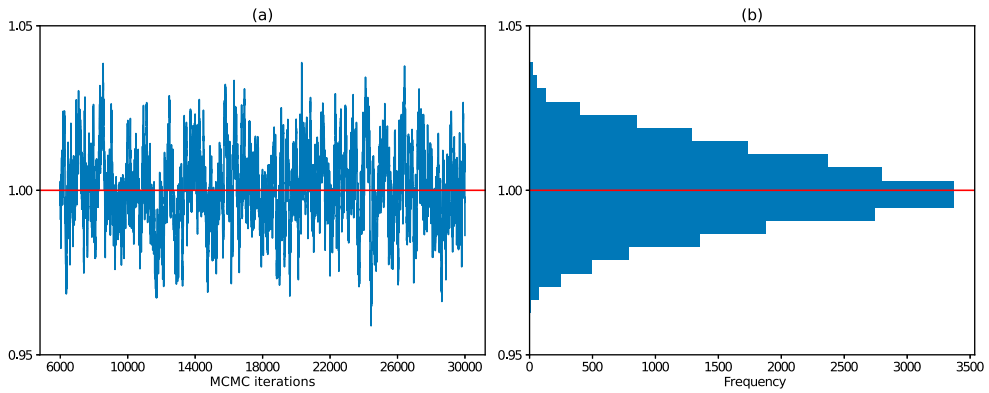


Figure 4. (a) is the traceplot of all samples after burn-in for β_2 ; (b) is the histogram of all samples after burn-in for β_2 of Case 1. The red horizontal line is the true value, 1.

6. Analysis on a real-world dataset

In this section, we consider a real-world application. Data has been collected from a clothing retailer that contains weekly sales information for the most recent 104 weeks. The descriptions of variables are given in Table 5. In order to further compare the performance, in addition to the proposed marketing mix model, we also implemented an ad hoc procedure that is currently used in practice. This ad hoc procedure is described in detail in Section 6.1 of the *supplemental document*. Abbreviated headings are used in Table 5 as follows: *Ind Var* is independent variable; *Dep Var* is dependent variable; *Pos Sign Cons* is positive sign constraint. The marketing mix model will be estimated by HMC since it has better performance than the other two optimization methods shown in the simulation study. The model performance is measured by the marginal R^2 and conditional R^2 [15], and the calculation details are provided in Section 6.2 of the *supplemental document*.

The model performance is reported in Table 6. For both metrics, the proposed model is better. In addition, we also observe that some of the estimates of the ad hoc process do not comply with the sign constraints, which agrees with the observations made in Section 5. In

Table 5. Descriptions of variables for the real dataset.

Variable	Adstocking Effect	Ind Var	Dep Var	Pos Sign Cons
Television TRP	Yes	Yes	No	Yes
Outdoor Impression	Yes	Yes	No	Yes
Coupon Distributed Quantity	No	Yes	No	Yes
Digital Marketing Distributed Quantity	No	Yes	No	Yes
Digital Display Impression	No	Yes	No	Yes
Digital Facebook Impression	No	Yes	No	Yes
Digital Instagram Impression	No	Yes	No	Yes
Digital Pinterest Impression	No	Yes	No	Yes
Digital Paidsearch Impression	No	Yes	No	Yes
Digital Youtube Impression	No	Yes	No	Yes
Seasonality Index	No	Yes	No	No
Unemployment Rate	No	Yes	No	No
Sales Quantity	No	No	Yes	N.A.

Table 6. Model performance of real application.

	Marginal R^2	Conditional R^2
ad hoc process	0.414	0.656
Marketing mix model	0.618	0.787

practice, additional heuristics will be employed to adjust the input dataset and/or arbitrarily ‘correct’ the estimated parameters, which leads to further deterioration of performance. With such disadvantages in the ad hoc process, the proposed model provides an attractive alternative for practitioners.

7. Concluding remarks

Marketing mix models have been widely used among practitioners as a standard way to quantify the effectiveness of advertising activities. However, the process is largely ad hoc and some parameters are set based on experience rather than derived from the data itself. In this research, we attempt to reduce ad hoc influence as much as possible by systematizing the whole process and making it more data-driven: we introduce nonlinear functions with unknown parameters to capture the carryover, shape and scale effects. In addition, we propose two models: the first is the base model where only the fixed effects are considered. The second model is one with hierarchical effects utilizing both the fixed effects and random effects to account for heterogeneity. All of the unknown parameters are simultaneously learned by two optimization methods as well as by the HMC, which is a novel Bayesian method originating from the study of Hamiltonian dynamics in physics. Moreover, sign constraints are also taken into account via proper specification of prior distributions as well as the enhancement to the leapfrog algorithm discussed in Section 4. With the sign constraints ensuring intuitive outcome of marketing activities, the resulting marketing mix models are more realistic from the perspective of practitioners.

The proposed marketing mix model represents an attractive alternative over an ad hoc process currently used in practice. It not only streamlines the modeling process entirely, but also incorporates the sign constraints automatically through the model specification. The ad hoc process is considered for the real application in Section 6, but it is unable to provide estimates that comply with all sign constraints in any of the examples. In practice, heuristics

to correct the signs of the parameters will be employed further introducing subjectiveness in the measures.

Admittedly, there are still some areas where we can continue to address. For example, we assume a constant carryover effect for each advertisement. However, a more sophisticated way of quantifying the carryover, shape and scale effects is that one could allow all of the three effects to vary across different regions, although it might dramatically increase the number of unknown parameters. All in all, we believe that systematizing and standardizing marketing mix model and letting the data speak through the model is crucial to the success of any marketing analytics application, especially in this big data era.

Acknowledgments

The authors would like to thank Dr. Hao Chen (no relation to the first author) of the NielsenIQ Precima Merchandising Analytics team for introducing us to the Hamiltonian Monte Carlo approach. The authors also thank the editor-in-chief, the associate editor, and the two anonymous reviewers for their helpful suggestions that have helped to improve the manuscript substantially.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

- [1] B.A. Bickart, *Carryover and backfire effects in marketing research*, J. Market. Res. 30 (1993), pp. 52–62.
- [2] P.T. Boggs and J.W. Tolle, *Sequential quadratic programming*, Acta Numer. 4 (1995), pp. 1–51.
- [3] N. Bokde, A. Feijóo, and D. Villanueva, *Wind turbine power curves based on the weibull cumulative distribution function*, Appl. Sci. 8 (2018), p. 1757.
- [4] N.H. Borden, *The concept of the marketing mix*, J. Advert. Res. 4 (1964), pp. 7–12.
- [5] G. Casella and R.L. Berger, *Statistical inference*, in *Duxbury Advanced Series*, 2001.
- [6] D. Chan and M. Perry, *Challenges and opportunities in media mix modeling* (2017). Available at <https://ai.google/research/pubs/pub45998>.
- [7] S. Chib and E. Greenberg, *Understanding the metropolis-hastings algorithm*, Am. Stat. 49 (1995), pp. 327–335.
- [8] T.A. Chowdhury and S. Naheed, *Multidimensional political marketing mix model for developing countries: An empirical investigation*, J. Political Mark. (2019). doi:10.1080/15377857.2019.1577323.
- [9] P.A.M. Dirac, *Generalized Hamiltonian dynamics*, Can. J. Math. 2 (1950), pp. 129–148.
- [10] M.K. Jain, *Numerical Solution of Differential Equations*, Wiley Eastern, New Delhi, 1979.
- [11] Y. Jin, Y. Wang, Y. Sun, D. Chan, and J. Koehler, *Bayesian methods for media mix modeling with carryover and shape effects* (2017). Available at <https://ai.google/research/pubs/pub46001>.
- [12] J.J. Lambin, *A computer on-line marketing mix model*, J. Market. Res. 9 (1972), pp. 119–126.
- [13] J.D. Little, *Brandaidd: A marketing-mix model, part 1: Structure*, Oper. Res. 23 (1975), pp. 628–655.
- [14] E.J. McCarthy, *Basic Marketing: A Managerial Approach*, Irwin McGraw-Hill, Homewood, 1978.
- [15] S. Nakagawa and H. Schielzeth, *A general and simple method for obtaining r^2 from generalized linear mixed-effects models*, Methods Ecol. Evol. 4 (2013), pp. 133–142.
- [16] R.M. Neal, *MCMC using Hamiltonian dynamics*, in *Handbook of Markov Chain Monte Carlo*, Chapter 5, S. Brooks, A. Gelman, G. Jones, and X.L. Meng, eds., CRC Press, 2011.
- [17] E. Pantano, C.V. Priporas, and G. Migliano, *Reshaping traditional marketing mix to include social media participation*, Eur. Bus. Rev. 31 (2019), pp. 162–178.

- [18] O. Rahneva, H. Kiskinov, I. Dimitrov, and V. Matanski, *Application of a weibull cumulative distribution function based on m existing ones to population dynamics*, Int. Electron. J. Pure Appl. Math. 12 (2018), pp. 111–121.
- [19] J.S. Rosenthal, *Optimal proposal distributions and adaptive MCMC*, in *Handbook of Markov Chain Monte Carlo* 4, 2011.
- [20] D. Salman, Y. Tawfik, M. Samy, and A. Artal-Tur, *A new marketing mix model to rescue the hospitality industry: Evidence from egypt after the arab spring*, Future Bus. J. 3 (2017), pp. 47–69. <https://www.sciencedirect.com/science/article/pii/S2314721017300105>.
- [21] Y. Sun, Y. Wang, Y. Jin, D. Chan, and J. Koehler, *Geo-level Bayesian hierarchical media mix modeling* (2017). Available at <https://ai.google/research/pubs/pub46000>.
- [22] J. Surmanek, *Media Planning: A Practical Guide*, McGraw Hill Professional, Lincolnwood, 1996.
- [23] C. Zhu, R.H. Byrd, P. Lu, and J. Nocedal, *Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization*, ACM Trans. Math. Softw. (TOMS) 23 (1997), pp. 550–560.