

# Tackling Multicollinearity in Marketing Mix Models: A Bayesian Hierarchical Shrinkage Approach

Chiheb Ben Hammouda<sup>1</sup>, Yanfei Chen<sup>2</sup>, Anda Denic<sup>3</sup>, Tijn Jacobs<sup>\*4</sup>,  
and Paul Sanders<sup>1</sup>

<sup>1</sup>Mathematical Institute, Utrecht University, Utrecht, The Netherlands

<sup>2</sup>Faculty of Economics and Business, KU Leuven, Belgium

<sup>3</sup>Department of Computer Science, University of Nis, Serbia

<sup>4</sup>Department of Mathematics, Vrije Universiteit, Amsterdam, The Netherlands

## Abstract

Multicollinearity presents a fundamental challenge in Marketing Mix Modeling (MMM), particularly when estimating the individual effects of correlated media channels. In this work, we propose a Bayesian hierarchical shrinkage framework to address this issue within the context of Bayesian MMM. Our approach introduces structured priors that exploit group-level similarities among media channels while allowing for channel-specific deviations. We evaluate the method through synthetic experiments and showcase how it could be applied to real-world data from Opella, demonstrating improved parameter stability, interpretability, and robustness. The proposed framework offers an efficient solution for inference in multichannel marketing environments.

KEYWORDS: Bayesian Inference, Marketing Mix Modeling, Multicollinearity, Hierarchical Models, Shrinkage Priors, LightweightMMM, Media Attribution, Causal Inference

## 1 Introduction

In today's digital age, businesses have access to a wide variety of marketing channels, including TV, radio, print, online, and social media. With so many platforms

available, it becomes crucial for companies to understand the impact of their marketing efforts on sales and customer behavior. This is where marketing mix modeling (MMM) comes into play.

MMM is a statistical analysis technique used to measure the effectiveness of different marketing channels and determine the optimal allocation of resources across these channels (Chan and Perry, 2017). By analyzing historical data, MMM quantifies the impact of various marketing activities on key performance indicators (KPIs) such as sales, customer acquisition, and brand awareness.

MMM provides businesses with insights to make data-driven decisions about their marketing strategies. It helps companies understand the return on investment (ROI) of each marketing channel and optimize resource allocation to maximize overall marketing impact. The model can be used to develop forecasts of expected returns for different marketing scenarios, enabling businesses to set realistic goals and make informed decisions about future campaigns. Furthermore, MMM reveals how different marketing channels interact and complement each other, aiding in the creation of more cohesive and integrated marketing strategies.

One significant challenge in MMM is dealing with multicollinearity among different media channels (Hanssens et al., 2020; Chan and Perry, 2017). Multicollinearity can occur both in the data, which can be highly correlated due to business decisions to fund different media during the same periods, and in the model, making it difficult to distinguish between different media. This can lead to unreliable estimates and hinder the model's ability to provide actionable insights.

The primary constraint in this project is the requirement to use Bayesian modeling. Consequently, our improvements will be confined to methods applicable within the Bayesian framework. This constraint arises from the findings that traditional linear regression methods are insufficient for accurately estimating the required ROI (Jin et al., 2016). Additionally, the methods should integrate seamlessly with the current workflow of Google's **LightweightMMM** (Chan and Perry, 2017), the program and model on which Opella's research is based. Extending our approach to fit within this framework will be most beneficial for Opella. Furthermore, the focus of this project should be on the causal interpretation of the effect estimates, enabling Opella to provide the best possible advice to their customers regarding media spending. This means that our methods should prioritize reducing the variance of the solutions over achieving the perfect bias for optimal spending. Lastly, we should be able to provide solutions for a limited dataset.

To address these challenges, researchers have explored advanced statistical techniques and machine learning algorithms. For instance, Bayesian methods have gained popularity due to their ability to incorporate prior knowledge and handle uncertainty more effectively (Hanssens et al., 2020; Chan and Perry, 2017).

State-of-the-art Bayesian modeling techniques have significantly advanced the field of MMM by enhancing model performance and interpretability (MMA, 2024). Bayesian methods offer several advantages over traditional approaches, including the ability to incorporate prior knowledge, handle uncertainty, and provide probabilistic estimates. Recent developments in Bayesian MMM have focused on integrating ma-

---

chine learning techniques, such as deep learning and Gaussian processes, to capture complex relationships between marketing activities and business outcomes (Wiecki, 2022). These models can handle large datasets and uncover hidden patterns that traditional statistical models might miss. Additionally, Bayesian MMM can be calibrated to ensure consistency with incrementality measurements, improving the accuracy of marketing attribution (Chan and Perry, 2017). The use of Bayesian methods in MMM has revolutionized business strategies and decision-making processes, providing marketers with more robust and actionable insights.

Furthermore, hierarchical modeling, which often utilizes the Bayesian framework, has become an essential approach in MMM to address the complexities of multi-level data structures (Khandelwal, 2023). Hierarchical models allow for the analysis of data organized at different levels, such as regions and media categories, providing more granular insights into marketing effectiveness. Traditional MMM approaches often struggle to capture the relationships within hierarchical data, leading to biased estimates and suboptimal marketing strategies. Bayesian hierarchical models, in particular, have shown promise in overcoming these limitations by incorporating prior knowledge and allowing for more flexible modeling of hierarchical structures (Chen et al., 2020; Khandelwal, 2023). These models can account for the carryover, shape, and scale effects of marketing activities, providing a more comprehensive understanding of their impact.

Despite these advancements, several challenges remain in the field of MMM. One significant issue is the need for transparency and interpretability in models. As MMM becomes more complex, it is crucial for marketers to understand the underlying mechanisms driving the results. Efforts to develop more interpretable models, such as those based on explainable AI, are ongoing (Wiecki, 2022). Furthermore, the deprecation of third-party cookies and increasing privacy concerns have prompted researchers to explore privacy-preserving techniques in MMM (Roggio, 2025).

In this report, we explore an extension of the hierarchical modeling approach within a Bayesian framework, focusing specifically on categorical media spending. Our aim is to develop a structured method that captures distinctions between different media categories, which we believe is a meaningful step toward improving interpretability and attribution in MMM.

To set the stage, we begin by revisiting the baseline Bayesian model underlying Google’s LightweightMMM framework, providing the necessary background for our proposed enhancements. We then introduce an improved hierarchical model, presenting results under various prior specifications to highlight its flexibility and robustness. The report concludes with an application to real-world data provided by Opella, enabling us to test our hypotheses and assess the empirical performance of the proposed framework.

To set the stage, we begin by revisiting the baseline Bayesian model underlying Google’s™s LightweightMMM framework, providing the necessary background for our proposed enhancements. We then introduce an improved hierarchical model, presenting results under various prior specifications to highlight its flexibility and robustness. While our motivation stems from a real-world marketing dataset provided

by Opella – characterized by strong multicollinearity among media channels – our current analysis focuses on synthetic data to serve as a proof of concept. The real data exploration is summarized in Appendix ?? and illustrates the practical relevance of the problem. Due to the computational challenges associated with extending the hierarchical model to constrained parameter spaces, applying the method to the real data is left for future work.

## 2 Bayesian Approach for Marketing Mix Modeling

We begin by describing the full marketing mix model. In our modeling, we follow a Bayesian approach. More details are provided in Section 2.1. Let  $X_{t,m}$  represent the expenditure on the media channel  $m = 1, 2, \dots, M$  during week  $t \geq 0$ , and let  $Z_{t,j}$  denote the control variables  $j = 1, 2, \dots, J$  measured in the same period. The dependent variable,  $Y_t$ , represents the key performance indicator (KPI) at week  $t$ . The relationship between the dependent and independent variables is formalized through the following model, adapted from Jin et al. (2017):

$$Y_t = \alpha + \nu t^\kappa + \sum_{d=1}^2 \left( \gamma_{1,d} \cos\left(\frac{2\pi dt}{52}\right) + \gamma_{2,d} \sin\left(\frac{2\pi dt}{52}\right) \right) + \sum_{m=1}^M \beta_m \text{Hill}(X_{t,m}^*; \mathcal{K}_m, S_m) + \sum_{j=1}^J \gamma_j Z_{t,j} + \epsilon, \quad (1)$$

where  $\epsilon \sim \mathcal{N}(0, \sigma^2)$  represents independent Gaussian noise.

Each term in the model (1) plays a specific role in capturing different effects:

- **Baseline level:** The parameter  $\alpha$  represents the inherent level of the KPI in the absence of external influences.
- **Trend effect:** The term  $\nu t^\kappa$  models long-term trends, capturing either a growing or declining trajectory over time.
- **Seasonality:** The seasonal component,  $\sum_{d=1}^2 \left( \gamma_{1,d} \cos\left(\frac{2\pi dt}{52}\right) + \gamma_{2,d} \sin\left(\frac{2\pi dt}{52}\right) \right)$ , accounts for periodic fluctuations within a year, such as seasonal demand variations.
- **Carryover and diminishing returns:** The term modeled by the Hill function,

$$\text{Hill}(X_{t,m}^*; \mathcal{K}_m, S_m) := \frac{X_{t,m}^*}{1 + (X_{t,m}^*/\mathcal{K}_m)^{-S_m}},$$

captures the carryover and saturation effects of media expenditures, ensuring that increasing investments do not necessarily result in proportional KPI increases, where in (1),  $X_{t,m}^* := X_{t,m} + \eta_m X_{t-1,m}^*$ , is the adjusted media spend with carryover effect and  $X_{1,m}^* = X_{1,m}$ .

- **Effectiveness of media spending:** The coefficient  $\beta_m \geq 0$  quantifies the effectiveness of spending in media channel  $m$ , representing its contribution to the KPI.
- **Control variables:** The term  $\sum_{j=1}^J \gamma_j Z_{t,j}$  adjusts for external factors that influence the KPI, such as economic indicators, holidays, or competitor activity.

A complete specification of the priors assigned to the parameters in equation (1) is provided in Appendix B.

This model provides a comprehensive framework for capturing the full effects of marketing activities and external factors on the KPI. However, estimating the full model introduces several challenges, particularly in handling heterogeneity across media channels, incorporating prior knowledge, and ensuring identifiability.

## 2.1 Bayesian Approach

In the Bayesian framework, parameters are treated as random variables rather than fixed values. This approach begins with specifying a likelihood function, denoted as  $\mathcal{L}(y | \theta)$ , which describes the probability of observing the data  $y$  conditional on a set of parameters  $\theta$ . To incorporate prior knowledge or uncertainty about these parameters, we introduce a prior distribution,  $p(\theta)$ , which represents our beliefs about  $\theta$  before observing the data.

Bayes' theorem then allows us to update our prior beliefs using the observed data, resulting in a posterior distribution of the parameters:

$$p(\theta | y) = \frac{p(\theta)\mathcal{L}(y | \theta)}{p(y)} \propto p(\theta)\mathcal{L}(y | \theta), \quad (2)$$

where  $p(y)$ , known as the marginal likelihood, serves as a normalizing constant independent of  $\theta$ . Since this term does not affect inference directly, it is often omitted, leaving the unnormalized posterior  $p(\theta)\mathcal{L}(y | \theta)$ . This formulation highlights the core idea of Bayesian inference: the prior distribution  $p(\theta)$  is updated with information from the data through the likelihood function  $\mathcal{L}(y | \theta)$  to obtain the posterior distribution  $p(\theta | y)$ .

## 2.2 Bayesian Hierarchical Modeling

In many applications, parameter estimation extends beyond a single stage of priors. Instead, the model introduces additional latent variables or unobservables, which govern the prior distributions of the original parameters. This leads to hierarchical Bayesian models, where parameters at one level are themselves assigned probability distributions conditioned on higher-level parameters. If  $\theta$  depends on a hyperparameter  $\kappa$ , and  $\kappa$  follows a prior distribution conditioned on another parameter  $\alpha$ , this hierarchical structure can be expressed as:

$$\begin{aligned} \theta &\sim p(\theta | \kappa) \\ \kappa &\sim p(\kappa | \alpha), \end{aligned} \quad (3)$$

where  $\alpha$  is assumed to be known, or a fixed hyperparameter. This formulation allows for a more flexible encoding of dependencies and structural uncertainty within the model.

Bayesian modeling, in a broader sense, describes the generative mechanism underlying both observed data and unobserved quantities. The unobserved components in the hierarchy may represent either latent variables or parameters, but regardless of their role, their distributions are updated in light of observed data. The posterior distribution of any unobservable, whether it is a model parameter or a latent effect, represents the updated knowledge about that quantity given the data.

A hierarchical Bayesian model can be viewed as a multi-step generative process:

- First, the parameter vector  $\theta$  is drawn from a prior distribution  $\Pi$ .
- Then, the observed data  $x$  is generated from a probability distribution  $P_\theta$ , conditioned on  $\theta$ .

In this structure, the distinction between parameters and latent variables is conceptual rather than fundamental. The key difference lies in observability: parameters remain unobserved, while data are directly measurable. The goal of Bayesian inference is to estimate the posterior distributions of all unobserved quantities given the observed data.

### 2.2.1 Advantages of Hierarchical Bayesian Modeling in Marketing Mix Models

Hierarchical Bayesian modeling offers a principled framework for incorporating structural dependencies in parameter estimation. Unlike traditional approaches that treat parameters independently, hierarchical models introduce higher-level distributions that can govern groups of parameters, enabling regularization and improved robustness.

This is particularly useful in marketing mix models, where media effectiveness varies across channels, data are often limited, and latent sources of variation – such as consumer behavior or macroeconomic trends – are difficult to observe directly. By modeling these dependencies probabilistically, hierarchical models enable estimates that remain well-behaved even in the presence of sparsity or noise.

Moreover, these models are flexible enough to incorporate important features of advertising dynamics such as carryover and saturation effects, enhancing both interpretability and inference quality.

In Section 3, we build on these ideas to develop a hierarchical shrinkage approach specifically tailored to handle the challenges posed by multicollinearity in marketing data.

## 2.3 A Base Model

To better illustrate the core aspects of our methodology, we consider a reduced version of model (1) that focuses exclusively on media expenditures and control variables.

This base model abstracts away nuisance components such as baseline effects, long-term trends, and seasonality, which, while important in applied settings, are not the primary focus of our analysis. By isolating the marginal posterior effects of media spending, this simplification enables clearer interpretation, facilitates methodological comparisons, and more directly highlights the impact of hierarchical priors in the estimation process. Moreover, it allows for a more efficient sampling strategy and straightforward implementation, making it well-suited for initial methodological evaluation.

The base model is given by:

$$Y_t = \sum_{m=1}^M \beta_m \text{Hill}(X_{t,m}^*; \mathcal{K}_m, S_m) + \gamma Z_t + \epsilon, \quad (4)$$

where  $\epsilon \sim \mathcal{N}(0, \sigma^2)$  represents independent Gaussian noise.

This model retains the core response function of media spending while accounting for the carryover, shape, and scale effects through the Hill transformation. Additionally, a single control variable,  $Z_t$ , is included to adjust for external influences.

### 2.3.1 Prior Specification

In the full model, the effectiveness parameters  $\beta_m$  are traditionally assumed to be a priori independent with a non-negativity constraint. This constraint is often imposed in Bayesian Marketing Mix Modeling under the assumption that media expenditures should not have a negative impact on the KPI. Specifically, a common prior formulation is:

$$\begin{aligned} \beta_m &= |\beta_m^*|, \\ \beta_m^* &\sim \mathcal{N}(0, \sigma_{\beta^*}^2), \end{aligned} \quad (5)$$

for  $m = 1, 2, \dots, M$ , where  $\sigma_{\beta^*}^2 > 0$  is a fixed variance parameter. This prior structure ensures that the effectiveness parameters remain non-negative while allowing for a degree of variability.

This approach has been widely adopted in the Bayesian analysis of Marketing Mix Models, as studied by [Chen et al. \(2021\)](#). Also, a log-transform could be performed to make the effectiveness parameters non-negative. However, in our base model, we choose to relax the positivity constraint and directly model the effectiveness parameters as:

$$\beta_m \sim \mathcal{N}(0, \sigma_{\beta}^2). \quad (6)$$

This alternative formulation simplifies the inference process while still allowing the data to determine the directionality of the effect. By removing the strict non-negativity assumption, we enable greater flexibility in capturing potential media inefficiencies or saturation effects that may result in non-positive influences on the KPI.

### 3 A Hierarchical Shrinkage Approach

In this section, we present our solution to the problem by introducing a Bayesian hierarchical model that mitigates the challenges posed by multicollinearity. The presence of strong correlations among covariates complicates the analysis, makes it difficult to isolate the individual effects of explanatory variables. Standard regression approaches struggle in such cases, often leading to unstable estimates and inflated parameter variances. Hierarchical Bayesian modeling provides a structured approach to addressing these challenges through both regularization and information sharing across related parameters.

One of the primary mechanisms for handling multicollinearity is the use of shrinkage priors, which serve as Bayesian counterparts to regularization techniques like Lasso or Ridge regression. By introducing structured priors on parameters, hierarchical models shrink the coefficients of correlated variables toward zero or toward each other, reducing overfitting and improving interpretability. This regularization effect ensures that the estimates remain stable even when the design matrix contains highly correlated predictors.

Additionally, hierarchical structuring allows for pooling of information across related components, such as media channels categorized by type. Instead of estimating each parameter independently, the model introduces shared higher-level distributions that constrain individual parameter estimates based on common patterns. This pooling effect reduces variance in the estimates, helping to disentangle collinear effects by focusing on broader trends rather than purely individual-level relationships. By incorporating both shrinkage and hierarchical structure, our proposed method effectively addresses multicollinearity while preserving the flexibility needed for nuanced media effectiveness estimation. In the next Section, we introduce our extended hierarchical model, demonstrating how these principles are applied in the marketing mix context.

#### 3.1 Mathematical formulation

Recall the base model as presented in equation (4). In the standard formulation, the effectiveness parameters  $\{\beta_m\}_{m=1}^M$  are typically modeled as independent across media channels  $m = 1, 2, \dots, M$ . However, this assumption overlooks the underlying similarities between certain channels. To address this, we introduce a hierarchical structure where media effectiveness parameters are grouped according to their pre-defined *media component*, incorporating prior dependencies among the effectiveness parameters. This grouping can be informed by domain knowledge, such as categorizing media channels into digital and offline groups, or it can be determined empirically through clustering techniques, as proposed by [Wu et al. \(2023\)](#).

Assume there are  $C \geq 1$  media components, each representing a group of media channels. Within each component  $c = 1, 2, \dots, C$ , the mean effectiveness is denoted by  $\mu_c$ , representing the shared impact of media spending within that component. However, media channels within the same component may still exhibit individual differences in effectiveness. To capture this heterogeneity, we introduce two key pa-

parameters: the within-component variance  $\tau_c$ , which quantifies the extent to which individual media channel effects are centered around the shared component mean, and the channel-specific scaling parameter  $\lambda_m$ , which reflects how much the effectiveness of the  $m$ th channel deviates from the component mean.

To formalize this structure, we specify a hierarchical prior on the effectiveness parameters  $\beta_m$ , where each media channel’s effect is modeled as a deviation from its component mean  $\mu_c$ . This deviation is scaled by a channel-specific factor  $\lambda_m$  and the within-component variance  $\tau_c$ , allowing flexibility in capturing both group-level and channel-specific variations. The model is formulated as follows:

$$\begin{aligned}\beta_m \mid \mu_c, \tau_c^2, \lambda_m^2, \sigma^2 &\sim \mathcal{N}(\mu_c, \lambda_m^2 \tau_c^2 \sigma^2), \\ \mu_c \mid \sigma_\mu^2 &\sim \mathcal{N}(0, \sigma_\mu^2), \\ \lambda_m^2 &\sim p_\lambda, \\ \tau_c^2 &\sim p_\tau,\end{aligned}\tag{7}$$

for  $m = 1, 2, \dots, M$ , where media channel  $m$  belongs to component  $c$  and  $\sigma_\mu^2$  is a fixed hyperparameter.

The *global shrinkage parameter*  $\tau_c$  controls how tightly individual media effects  $\beta_m$  cluster around the component mean  $\mu_c$ . A small  $\tau_c$  indicates strong within-component similarity, while a larger  $\tau_c$  allows for greater deviation. The *local shrinkage parameter*  $\lambda_m$  allows individual media channels to deviate from the component mean to varying degrees, ensuring that outliers or particularly influential channels are not overly shrunk towards the component average. The mixing densities  $p_\lambda$  and  $p_\tau$  allow for flexible modeling of the shrinkage behavior. This formulation aligns with the broader class of global-local shrinkage priors (Polson and Scott, 2011), which have been widely studied in Bayesian statistics.

Several well-known shrinkage priors fit within this framework, including the Horseshoe prior (Carvalho et al., 2010), the Regularized Horseshoe (Bhadra et al., 2017), the Bayesian Lasso (Park and Casella, 2008), the Normal-Gamma prior (Brown and Griffin, 2010), and the Dirichlet-Laplace prior (Bhattacharya et al., 2015).

We will specify two prior formulations that can be useful in this setting.

### 3.2 The Normal-Inverse Gamma Model

The Normal-Inverse Gamma model provides a simple and conjugate framework, often used in Bayesian regression settings where variance uncertainty needs to be modeled explicitly. It allows for robust parameter estimation, particularly in hierarchical models where variance components require adaptive shrinkage.

The prior is formulated as:

$$\begin{aligned}
 \beta_m \mid \mu_c, \tau_c^2, \lambda_m^2, \sigma^2 &\sim \mathcal{N}(\mu_c, \tau_c^2 \lambda_m^2 \sigma^2), \\
 \mu_c \mid \sigma_\mu^2 &\sim \mathcal{N}(0, \sigma_\mu^2), \\
 \tau_c^2 \mid a_\tau, b_\tau &\sim \mathcal{IG}(a_\tau, b_\tau), \\
 \lambda_m^2 \mid a_\lambda, b_\lambda &\sim \mathcal{IG}(a_\lambda, b_\lambda),
 \end{aligned} \tag{8}$$

where  $a_\tau, b_\tau, a_\lambda, b_\lambda > 0$  are fixed hyperparameters. The inverse gamma  $\mathcal{IG}$  density is given by:

$$p(x \mid a, b) \propto x^{-(a+1)} \exp\left\{-\frac{b}{x}\right\}, \tag{9}$$

for  $x > 0$ .

These hyperparameters can be chosen based on prior knowledge about the effectiveness of a certain media component. For instance, setting  $a_\tau = a_\lambda = b_\tau = b_\lambda = 0.1$  induces a weakly informative prior, reflecting prior ambivalence about the variability in media effectiveness. On the other hand, setting  $a_\tau = a_\lambda = 3$  and  $b_\tau = b_\lambda = 1.5$  results in a more informative prior, suggesting that the effectiveness of spending within this component is expected to exhibit limited variability.

The tails of the inverse gamma distribution allow for substantial within-component variability in effect estimates, enabling the model to accommodate both small and large variance values flexibly. Moreover, the inverse gamma prior can be finely adjusted to reflect prior knowledge about the underlying data-generating mechanism. By carefully selecting the hyperparameters  $a$  and  $b$ , one can control the degree of shrinkage applied to variance estimates, ensuring that the prior remains informative while still allowing sufficient flexibility in hierarchical modeling.

Figure 1 illustrates the density of this prior, conditional on  $\mu_c = 0$ , overlaid on a Gaussian distribution with fixed variance. This highlights the prior's ability to capture a wide range of potential effect sizes while maintaining hierarchical structure.

### 3.3 The Horseshoe prior

The horseshoe prior [Carvalho et al. \(2010\)](#) is a sparsity-inducing prior that has gained popularity in Bayesian modeling, particularly for high-dimensional regression and hierarchical models. It is designed to strongly shrink small signals while allowing large signals to remain unshrunk, making it particularly well-suited for sparse models where only a subset of predictors have substantial effects. Recent work has provided theoretical guarantees for the horseshoe prior, demonstrating its ability to achieve optimal posterior contraction rates in high-dimensional settings [Van der Pas et al. \(2014, 2017\)](#).

The Horseshoe prior introduces a global-local shrinkage structure that adaptively regulates the amount of regularization applied to each coefficient. It is defined hier-

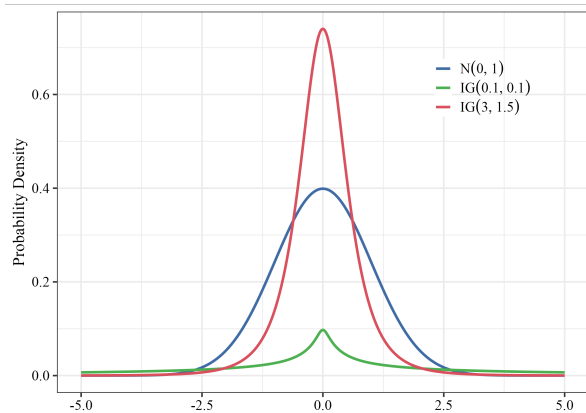


Figure 1: Density comparison of Normal-Inverse Gamma priors with varying informativeness. The standard normal distribution  $\mathcal{N}(0, 1)$  is shown as a reference, while the weakly informative  $\mathcal{IG}(0.1, 0.1)$  prior allows for greater variance, and the moderately informative  $\mathcal{IG}(3, 1.5)$  prior concentrates more strongly around zero.

archically as:

$$\begin{aligned}
 \beta_m \mid \mu_c, \tau_c^2, \lambda_m^2, \sigma^2 &\sim \mathcal{N}(\mu_c, \tau_c^2 \lambda_m^2 \sigma^2), \\
 \mu_c \mid \sigma_\mu^2 &\sim \mathcal{N}(0, \sigma_\mu^2), \\
 \tau_c^2 &\sim \mathcal{C}^+(0, 1), \\
 \lambda_m^2 &\sim \mathcal{C}^+(0, 1),
 \end{aligned} \tag{10}$$

where each media channel  $m$  is associated with a group  $c$ , and  $\mathcal{C}^+(0, 1)$  denotes the standard half-Cauchy distribution. The density of a half-Cauchy distribution with scale  $\alpha > 0$  is given by:

$$p(x \mid \alpha) \propto \frac{1}{1 + (x/\alpha)^2}, \quad x > 0. \tag{11}$$

Due to its heavy tails and strong peak at zero, the Horseshoe prior is well-suited for sparse settings: it aggressively shrinks noise-dominated coefficients while leaving signals with strong evidence relatively unshrunk. This adaptivity makes it robust to both small and large effects.

The global shrinkage parameter  $\tau_c$  controls the overall level of sparsity, while the local shrinkage parameters  $\lambda_m$  allow individual coefficients to either be strongly regularized or remain large. The half-Cauchy priors on  $\tau_c$  and  $\lambda_m$  ensure heavy-tailed behavior, leading to strong shrinkage of small coefficients while allowing larger ones to remain unshrunk. This adaptivity enhances separation between true signals and noise by aggressively regularizing small effects while preserving significant ones.

In Figure 2, the density of the horseshoe prior is plotted conditional on  $\mu_c = 0$ . Compared to Figure 1, it exhibits stronger shrinkage near zero while maintaining heavier tails, allowing for better separation between small and large effects. These properties make the horseshoe prior particularly effective in sparse models, where only a subset of predictors have substantial effects.

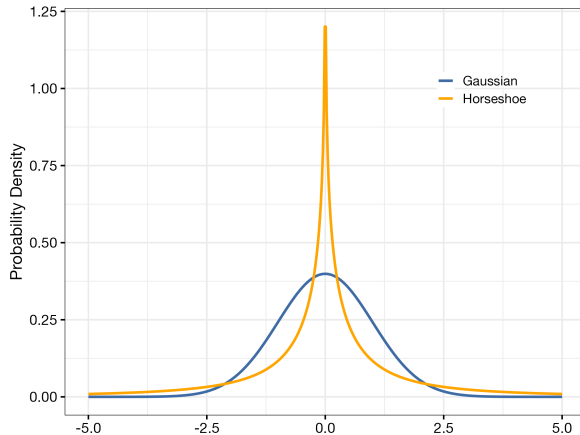


Figure 2: Density of the horseshoe prior with a standard normal as reference.

Unlike traditional Normal-Inverse Gamma priors or Laplace priors (Bayesian LASSO), the horseshoe prior offers adaptive shrinkage, applying aggressive regularization to small coefficients while leaving large ones relatively unaffected [Carvalho et al. \(2009\)](#); [Polson and Scott \(2012\)](#). This property makes it highly beneficial in sparse regression, variable selection, and hierarchical Bayesian modeling, where a small number of predictors contribute significantly to the response. Furthermore, the horseshoe prior has been shown to provide asymptotic uncertainty quantification, making it a principled choice for inference in high-dimensional settings [Van der Pas et al. \(2016\)](#).

### 3.4 Posterior Computation

The choice of priors in our model ensures that posterior sampling can be performed efficiently. We employ Hamiltonian Monte Carlo (HMC) ([Neal, 2011](#); [Betancourt, 2017](#)), a gradient-based Markov Chain Monte Carlo (MCMC) method. Unlike traditional random-walk Metropolis-Hastings algorithms, which often struggle with correlated or ill-conditioned posteriors, HMC uses Hamiltonian dynamics to explore the parameter space more effectively.

HMC introduces an auxiliary momentum variable and models the posterior distribution as a physical system, where kinetic and potential energy define its movement.

The algorithm follows Hamilton’s equations:

$$\begin{aligned}\frac{d\psi}{dt} &= \frac{\partial H}{\partial v}, \\ \frac{dv}{dt} &= -\frac{\partial H}{\partial \psi},\end{aligned}\tag{12}$$

where  $\psi$  represents the parameters of interest,  $v$  is an auxiliary momentum, and  $H(\psi, v)$  is the Hamiltonian function combining the potential energy  $U(\psi) = -\log P(\psi)$  (from the posterior) and kinetic energy  $K(v) = \frac{1}{2}v^T v$ . New proposals are generated by numerically solving these equations using the leapfrog integrator, which updates both position and momentum in multiple small steps.

Stan optimizes HMC using the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014), which adaptively tunes the step size and trajectory length, improving convergence and reducing the need for manual tuning. To further enhance computational efficiency and sampling robustness, we adopt a non-centered parameterization strategy (Betancourt and Girolami, 2015), which is particularly beneficial in hierarchical models with weakly identified parameters. The Stan implementation of the Horseshoe model is based on the auxiliary variables representation from Makalic and Schmidt (2016). It can be found in Appendix D along with the other implementations.

## 4 Numerical Examples

To evaluate the effectiveness and suitability of the proposed hierarchical priors, we conduct a series of simulated examples. These serve as a proof of concept, illustrating how the priors behave under controlled conditions. A key focus is to assess their performance in the presence of multicollinearity, particularly in settings with a large number of parameters and limited data. By simulating covariates with varying degrees of correlation, we examine the priors’ ability to produce stable and well-regularized parameter estimates of the effectiveness parameters  $\beta$ .

While these simulations offer valuable insight into the practical performance of the priors, they are exploratory in nature. A more rigorous empirical study and theoretical analysis are needed to fully characterize their properties and assess robustness across broader data-generating processes.

We begin with a description of the data generation process common to both simulations. Following that, we present two proof-of-concept examples. In the first, data is generated according to the base model. In the second, data is simulated under the extended hierarchical model. For both scenarios, we fit the base model with three different priors on the effectiveness parameters: an independent prior, a Normal-Inverse Gamma prior, and a Horseshoe prior.

### 4.1 Simulation set-up

To mimic media spending behavior, we generate synthetic time series data with  $n = 104$  time points (representing two years of weekly data) and  $m = m_1 + m_2$

covariates (spendings in media channels), grouped into two blocks. The first block ( $m_1 = 4$ ) consists of highly correlated covariates, constructed using a Toeplitz correlation structure with parameter  $\rho = 0.9$ . The second block ( $m_2 = 6$ ) is independent of the first. These form a block-diagonal covariance matrix  $\Sigma$ , from which multivariate Gaussian noise is drawn with within group correlation  $\rho = 0.9$ . Temporal dependence is added by applying an AR(1) process with autoregressive coefficient  $\phi = 0.8$ , yielding smoother time series. To reflect non-negative spending behavior, the series are exponentiated.

A single control variable  $Z_t \sim \mathcal{N}(0, 1)$  and Gaussian noise  $\varepsilon_t \sim \mathcal{N}(0, 1)$  are added. The outcome variable is computed according to (4) where  $\gamma = 1$ . Negative values of  $Y_t$  are truncated at zero. This setup creates a realistic yet controlled environment for evaluating prior performance in the presence of block-wise correlation, temporal dependence, and nonlinear effects.

To quantify performance, we repeat the simulation  $M = 100$  times and compute the following metrics:

- Root mean squared error (RMSE) of the estimated  $\beta$  coefficients.
- Average length of 95% credible intervals for the  $\beta$  coefficients.
- Average posterior mean of  $\sigma$ .

For each simulation, we draw 2,000 posterior samples following a burn-in of 1,000 iterations. To reduce autocorrelation, the samples are thinned by a factor of 2, meaning every second sample is retained.

## 4.2 Synthetic Example 1 – Base Model

In the first synthetic example, data is generated according to the base model described earlier. This setup assumes that the effectiveness parameters  $\beta$  are independent across media channels. We sample these coefficients from a uniform distribution  $\beta \sim \mathcal{U}[0.8, 1.4]$ . The specific prior settings and hyperparameters used for model fitting are detailed in Appendix C.

Figure 3 shows an example of the simulated media inputs and the corresponding KPI over time, where the media inputs are color-coded by group. The resulting posterior distributions of the  $\beta$  parameters under the three priors (independent, Normal-Inverse Gamma, and Horseshoe) are presented in Figure 4 together with the traceplots in Figure 5.

The results from  $M = 100$  replications are summarized in Table 1 and Figure 6, where we report the RMSE of the estimated  $\beta$  coefficients, the average length of their 95% credible intervals, and the average posterior mean of  $\sigma^2$ .

In terms of RMSE, the Normal-Inverse Gamma prior achieves the best performance, slightly outperforming both the base model and the Horseshoe prior. This indicates that even in the absence of true hierarchical structure, mild regularization can help stabilize estimates and improve predictive accuracy.

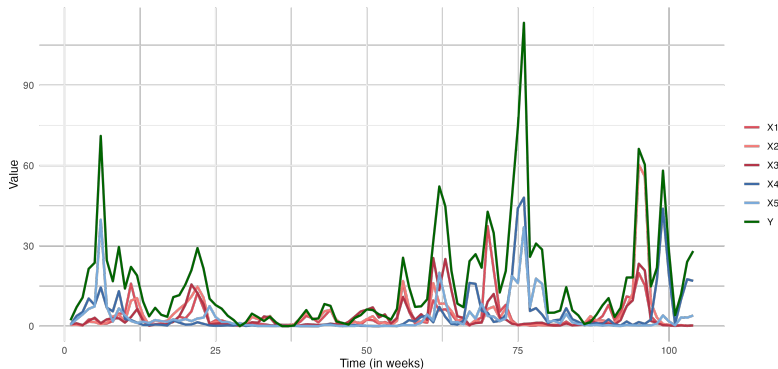


Figure 3: Simulated media covariates ( $X$ ) and KPI ( $Y$ ) over time for  $m_1 = 3$  and  $m_2 = 2$ . Group 1 channels are shown in red, Group 2 in blue, and the KPI in green.

Table 1: Synthetic example 1: performance metrics

Metric	Base	Normal-Inverse Gamma	Horseshoe
RMSE	0.1121	<b>0.0967</b>	0.1450
Avg. 95% CI length	0.8896	0.7791	<b>0.5364</b>
Average $\sigma^2$	1.0089	1.0080	1.0106

Both hierarchical priors result in tighter posterior distributions compared to the base model, as seen from the reduced average length of the 95% credible intervals. The Horseshoe prior in particular exhibits the strongest shrinkage effect, resulting in substantially narrower intervals than both alternatives. However, this comes at the cost of a slightly higher RMSE, reflecting a bias-variance trade-off. In scenarios with more media channels that are highly correlated, such aggressive regularization may be preferable, as it can help stabilize estimation and improve robustness.

Finally, all models produce comparable posterior estimates for the residual variance  $\sigma^2$ , suggesting that differences in uncertainty are primarily driven by the priors on  $\beta$ . Together, these results reinforce the utility of regularizing priors, even under model misspecification, while highlighting the need for cautious application of highly concentrated shrinkage like the Horseshoe.

### 4.3 Synthetic Example 2 – Hierarchical Model

In the second simulation, we generate data from the extended hierarchical model, where each media channel belongs to one of two predefined components. Effectiveness parameters  $\beta_m$  are drawn conditionally on their component-specific mean  $\mu_g$ , reflecting group-level structure and allowing for within-group variability.

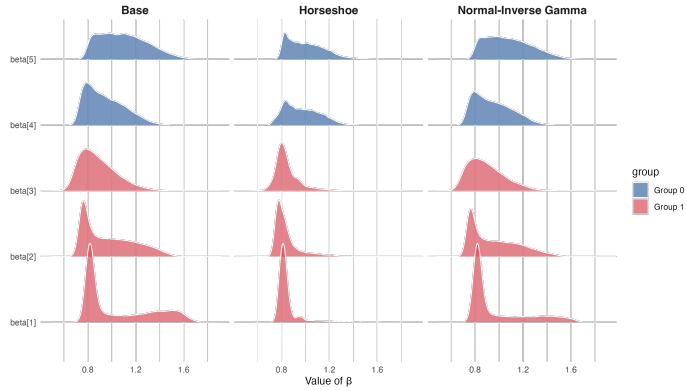


Figure 4: Posterior distributions of the  $\beta$  coefficients under the three priors: independent, Normal-Inverse Gamma, and Horseshoe.

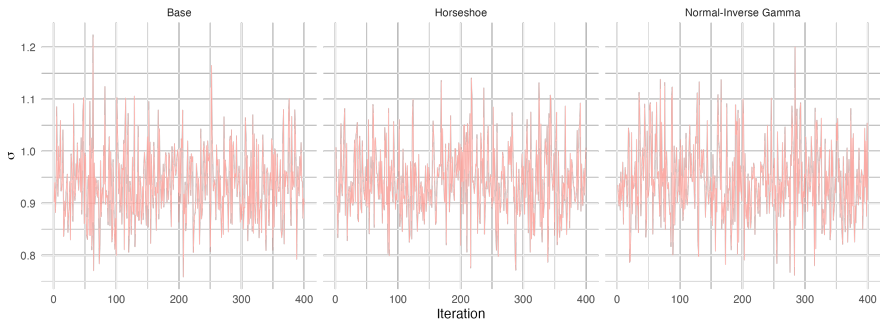


Figure 5: Traceplots of  $\sigma$  corresponding to the three sampled models.

First, the group mean effectiveness parameters are sampled as:

$$\mu_g \sim \begin{cases} \mathcal{U}[0.7, 0.9], & \text{if } g = 1 \text{ (Group 1),} \\ \mathcal{U}[1.4, 1.8], & \text{if } g = 2 \text{ (Group 2),} \end{cases} \quad (13)$$

where  $g \in \{1, 2\}$  indexes the component.

Subsequently, individual channel effects are drawn from:

$$\beta_m \sim \begin{cases} \mathcal{N}(\mu_1, 1/4), & \text{if } m \in \text{Group 1,} \\ \mathcal{N}(\mu_2, 1/2), & \text{if } m \in \text{Group 2.} \end{cases} \quad (14)$$

This design introduces heterogeneity within groups while preserving a clear separation between group-level means. It allows us to assess whether the hierarchical priors

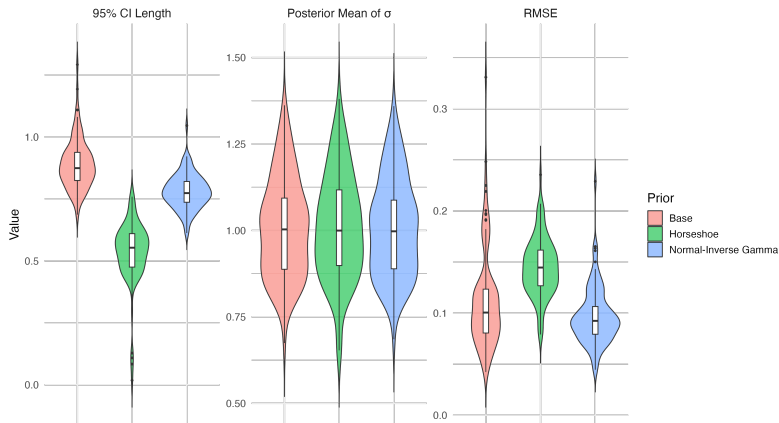


Figure 6: Synthetic example 1: evaluation metrics for all three models.

can accurately recover the underlying component structure under varying levels of within-group variation.

The results from  $M = 100$  replications are summarized in Table 2 and Figure 7, where we report the RMSE of the estimated  $\beta$  coefficients, the average length of their 95% credible intervals, and the average posterior mean of  $\sigma^2$ .

Table 2: Synthetic example 2: performance metrics

Metric	Base	Normal-Inverse Gamma	Horseshoe
RMSE	0.1203	<b>0.0868</b>	0.1130
Avg. 95% CI length	0.9646	0.8268	<b>0.5718</b>
Posterior mean of $\sigma^2$	1.0273	1.0249	1.0369

Across all three priors, the root mean squared error (RMSE) values are largely comparable, with the Normal-Inverse Gamma model showing a slight edge in accuracy. Although performance is generally stable, we observe a few outliers across the methods. These deviations may be attributed to sampling variability and could likely be reduced through increased computational effort, such as using more chains, longer warm-up periods, or more aggressive thinning to reduce autocorrelation.

Beyond point estimates, the hierarchical priors yield notable improvements in posterior uncertainty. Both the Normal-Inverse Gamma and Horseshoe models substantially reduce the average length of the 95% credible intervals relative to the base model. This indicates that hierarchical regularization introduces more concentrated posterior mass around the estimated coefficients, reflecting greater certainty in the estimates.

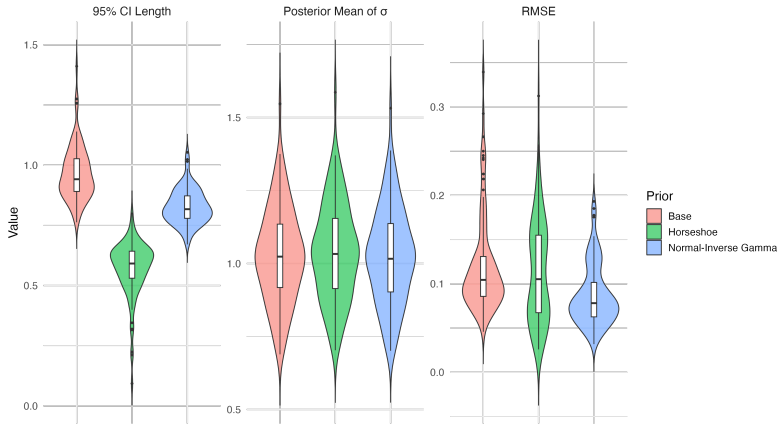


Figure 7: Synthetic example 2: evaluation metrics for all three models.

Among the models, the Horseshoe prior produces the narrowest intervals, reflecting its strong global-local shrinkage behavior. This pattern suggests that the Horseshoe is especially effective in suppressing noise while retaining signal, consistent with its design for sparse settings.

It was observed during preliminary runs that the standard Horseshoe prior may exhibit some instability in posterior computation. A natural extension would be to explore the Regularized Horseshoe prior (Piiironen and Vehtari, 2017), which improves the posterior geometry and often results in more stable and efficient sampling

## 5 Discussion

This paper presents a Bayesian hierarchical modeling framework for addressing multicollinearity in Marketing Mix Modeling (MMM). Using structured shrinkage priors – particularly the Normal-Inverse Gamma and Horseshoe distributions – the approach offers a principled way to improve inference in high-dimensional settings with correlated predictors. Through simulated experiments, the methodology demonstrates its ability to yield tighter credible intervals and more stable parameter estimates compared to baseline models with independent priors.

Importantly, this work should be seen as a proof of concept. The current implementation is restricted to a base version of the MMM and has not yet been applied to real-world marketing data. Extending the method to the full model – incorporating seasonal trends, long-term dynamics, and auxiliary control variables – remains an important direction for future research. A key challenge in this extension involves the posterior computation under positivity constraints for media effectiveness parameters. These constraints induce a nontrivial truncation of the posterior distribution, making

Hamiltonian Monte Carlo sampling more difficult and potentially less efficient.

Further research is needed to develop computational strategies that can handle such constraints in hierarchical shrinkage settings. Possible directions include reparameterizations, gradient-based constrained inference methods, or introducing auxiliary variable samplers tailored to truncated priors. Ultimately, the aim is to develop a scalable and interpretable framework that is robust to multicollinearity and applicable in real-world marketing applications.

## References

- Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.
- Michael Betancourt and Mark Girolami. Hamiltonian monte carlo for hierarchical models. In *Current Trends in Bayesian Methodology with Applications*. CRC Press, 2015. URL <https://arxiv.org/abs/1312.0906>.
- Anindya Bhadra, Jyotishka Datta, Nicholas G. Polson, and Brandon Willard. The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131, December 2017. doi: 10.1214/17-BA1075. URL <https://projecteuclid.org/euclid.ba/1516093226>.
- Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai, and David B. Dunson. Dirichlet-laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015. doi: 10.1080/01621459.2014.960967. URL <https://doi.org/10.1080/01621459.2014.960967>.
- Philip J. Brown and Jim E. Griffin. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, March 2010. doi: 10.1214/10-BA507. URL <https://projecteuclid.org/euclid.ba/1339616728>.
- Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. *Journal of Machine Learning Research*, 5:73–80, 2009.
- Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, June 2010. doi: 10.1093/biomet/asq017. URL <https://www.jstor.org/stable/25734098>.
- David Chan and Michael Perry. Challenges and opportunities in media mix modeling. *Google Research*, 2017. URL [https://services.google.com/fh/files/misc/challenges\\_and\\_opportunities\\_in\\_media\\_mix\\_modeling.pdf](https://services.google.com/fh/files/misc/challenges_and_opportunities_in_media_mix_modeling.pdf).
- Hao Chen, Minguang Zhang, Lanshan Han, and Alvin Lim. Hierarchical marketing mix models with sign constraints. *arXiv*, 2020. URL <https://arxiv.org/abs/2008.12802>.

- Hao Chen, Minguang Zhang, Lanshan Han, and Alvin Lim. Hierarchical marketing mix models with sign constraints. *Journal of Applied Statistics*, 48(13-15):2944–2960, 2021. doi: 10.1080/02664763.2021.1946020. URL <https://doi.org/10.1080/02664763.2021.1946020>.
- Dominique M. Hanssens et al. Market response models and marketing practice. *UCLA Anderson School of Management*, 2020. URL <https://www.anderson.ucla.edu/sites/default/files/documents/areas/fac/marketing/ASMBIPractice%280%29.pdf>.
- Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.
- Yuxue Jin, Yueqing Wang, Yunting Sun, David Chan, and Jim Koehle. Bayesian methods for media mix modeling with carryover and shape effects, 2016. URL <http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/46001.pdf>.
- Yuxue Jin, Yueqing Wang, Yunting Sun, David Chan, and Jim Koehler. Bayesian methods for media mix modeling with carryover and shape effects. Technical report, Google Inc., 2017.
- Shekhar Khandelwal. How to use bayesian hierarchical marketing mix modeling (bhmmm) to redefine marketing strategies at a regional level. *Mercury Media Technology*, 2023. URL <https://www.mercurymediatechnology.com/en/blog/bayesian-hierarchical-marketing-mix-modeling/>.
- Russell N Laczniak and Darrel D Muehling. Delayed effects of advertising moderated by involvement. *Journal of Business Research*, 20(3):263–277, 1990.
- Enes Makalic and Daniel F Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2016.
- Ipsos MMA. The current state of marketing mix modeling, 2024. URL <https://mma.com/blog/the-current-state-of-marketing-mix-modeling/>.
- Radford M Neal. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2, 2011.
- Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. doi: 10.1198/016214508000000337. URL <https://doi.org/10.1198/016214508000000337>.
- Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11:5018–5051, 2017. ISSN 1935-7524. doi: 10.1214/17-EJS1337SI. URL <https://doi.org/10.1214/17-EJS1337SI>.

---

Nicholas G. Polson and James G. Scott. Shrink globally, act locally: Sparse bayesian regularization and prediction. In José M. Bernardo et al., editors, *Bayesian Statistics 9*. Oxford University Press, 2011. doi: 10.1093/acprof:oso/9780199694587.003.0017. URL <https://doi.org/10.1093/acprof:oso/9780199694587.003.0017>. Online edition, accessed 27 Feb. 2025.

Nicholas G Polson and James G Scott. The half-cauchy prior for a global scale parameter in hierarchical models. *Bayesian Analysis*, 7(4):887–902, 2012.

A. Roggio. The rebirth of marketing mix modeling. <https://www.practicaledge.com/the-rebirth-of-marketing-mix-modeling>, 2025. Accessed: 2025-03-25.

S Van der Pas, B J Kleijn, and A W Van der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2): 2585–2618, 2014.

S Van der Pas, B Szabó, and A W Van der Vaart. Adaptive bayesian estimation using a hierarchical shrinkage prior. *Bernoulli*, 23(2):1031–1057, 2017.

S L Van der Pas, B Szabó, and A W Van der Vaart. Uncertainty quantification for the horseshoe. *Bayesian Analysis*, 11(4):1221–1248, 2016.

Thomas Wiecki. Bayesian marketing mix models: State of the art and their future, 2022. URL <https://www.pymc-labs.com/blog-posts/2022-11-11-HelloFresh/>.

Yufei Wu, Zhiying Gu, Alex Deng, Jacob Zhu, and Linsha Chen. Hierarchical clustering as a novel solution to the notorious multicollinearity problem in observational causal inference. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2023. URL <https://airbnb.tech/wp-content/uploads/sites/19/2023/12/31.KDD-Paper-Hierarchical-Clustering-As-a-Solution-to-Multicollinearity-%E2%80%93-Marketing-Application-as-an-Example.pdf>.

## A Real World Example: Sanofi Data

### A.1 Exploratory data analysis

The dataset provided by Opella includes weekly sales figures, media investments, and supplementary information for Allegra, covering the period from January 2022 to December 2023. In this section, we use tables and figures to analyze the dataset, uncover underlying patterns, and derive preliminary insights.



Figure 8: Allegra

### A.1.1 Sales information

The primary focus of our analysis is the sales volume of Allegra, with trends and frequencies depicted in Figures 9 and 10, respectively. From Figure 9, we observe distinct peaks in May and June of both 2022 and 2023, indicating a surge in demand during these periods. The histogram in Figure 10 reveals that the majority of weekly sales volumes fall below 60,000, with a noticeable long tail. To assess the distribution of the sales volume, we test for fits to the Poisson, exponential, and gamma distributions. The results show that none of these standard distributions adequately capture the sales volume distribution, a finding that is likely attributed to the significant tail in the sales volume data.

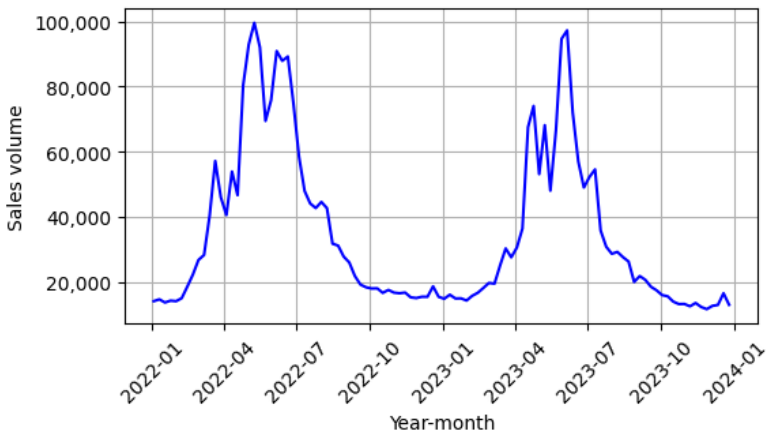


Figure 9: Trend of the sales volume

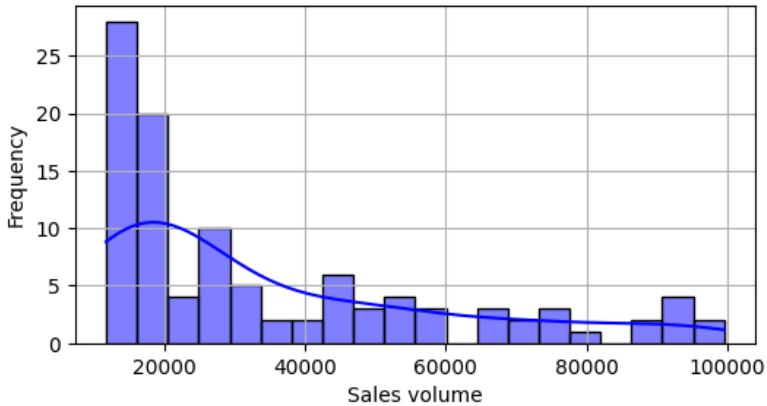


Figure 10: Frequency of the sales volume

### A.1.2 Media investments

To effectively promote Allegra, Opella invests in a total of 11 media channels: two offline, five online, and four trade. The offline channels include radio and newspapers, while the online channels encompass paid search, YouTube, social media, and others. Trade channels include customer promotions. In the dataset, each column corresponding to a media channel represents investments in euros. Figure 11 illustrates the weekly investments across all media channels. It is evident that the largest investment is directed towards Offline 2, with its investment significantly surpassing that of the other channels. Regarding the maximum weekly investment, both Offline 2 and Trade 2 exceed 400,000 euros, indicating Opella’s strong expectations for these channels.

In our analysis, we assume that the ratio of media impressions to investment is constant for each media channel. Therefore, the level of investment is proportional to the magnitude of impressions for each channel. We proceed by calculating the Spearman correlation matrix for all media channels, with the heatmap shown in Figure 12. The figure reveals a strong positive correlation among all online channels, as indicated by the dark-blue square in the center of the heatmap. Additionally, a strong positive correlation is observed among the first three trade channels. Interestingly, the fourth trade channel shows only mild correlations with the other three trade channels. While inter-category correlations are present, they are generally weaker than the intra-category correlations.

We also examine the delayed effects of advertisements, where the impact of an advertisement is not immediate but occurs with a time lag (Laczniaak and Muehling, 1990). To analyze this, we select three representative media channels—Offline 2, Online 4, and Trade 2—which receive the highest investments in their respective categories. Figure 13 juxtaposes their investment trends with sales volume. The figure

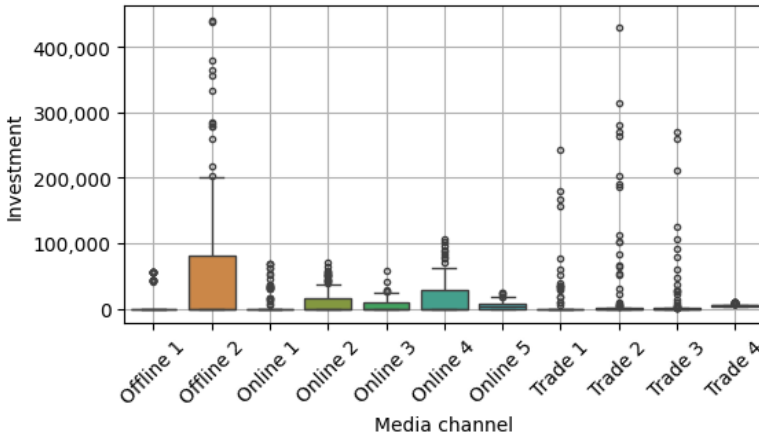


Figure 11: Investments across media channels

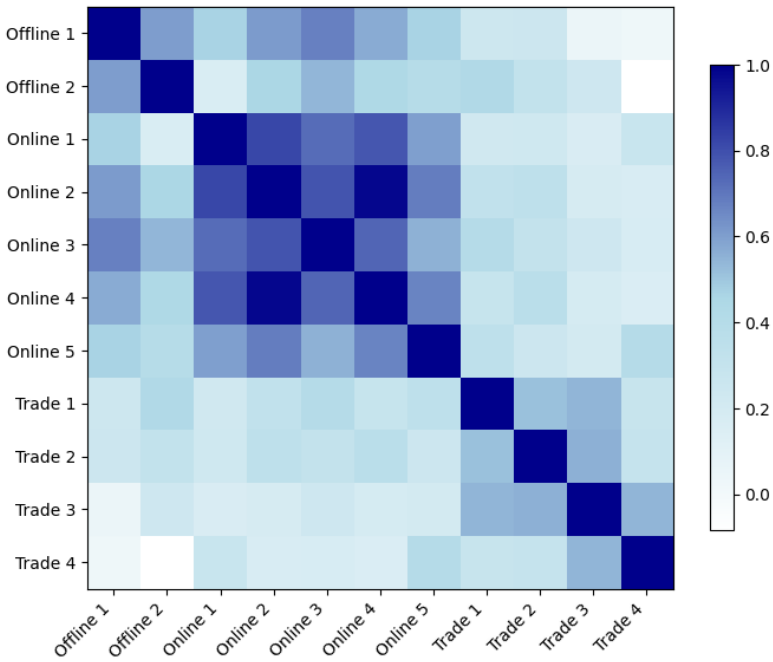


Figure 12: Spearman correlation matrix of media channels

reveals two peaks in media investment, each slightly preceding the corresponding peak in demand. Additionally, the investments appear to be executed in intervals, possibly

indicating that Opella aims to avoid customers seeing the same advertisement over consecutive weeks.

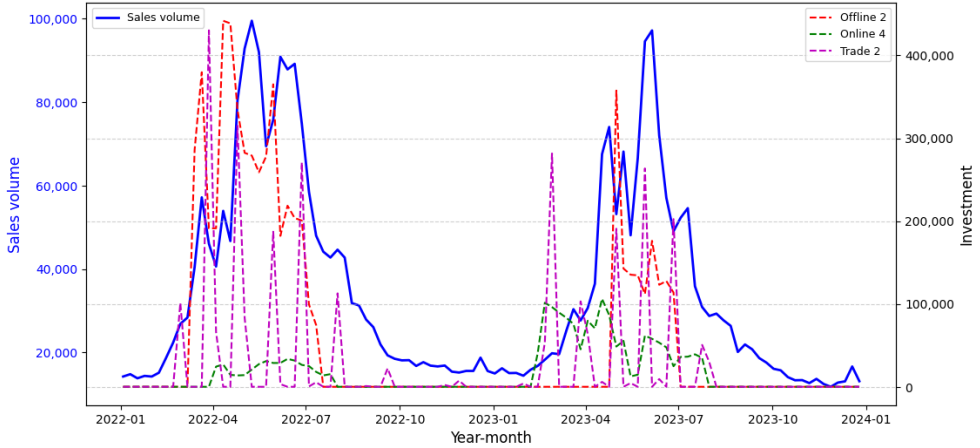


Figure 13: Delayed effects of advertisements

### A.1.3 Supplementary information

The dataset includes a column titled “allergy index”, which is calculated using Google Trends data for selected keywords as a proxy for measuring allergy cases. The chronological trend of the allergy index is visualized in Figure 14. We observe that the trend closely mirrors that of the sales volume, suggesting that demand for Allegra increases when allergic symptoms are more prevalent.

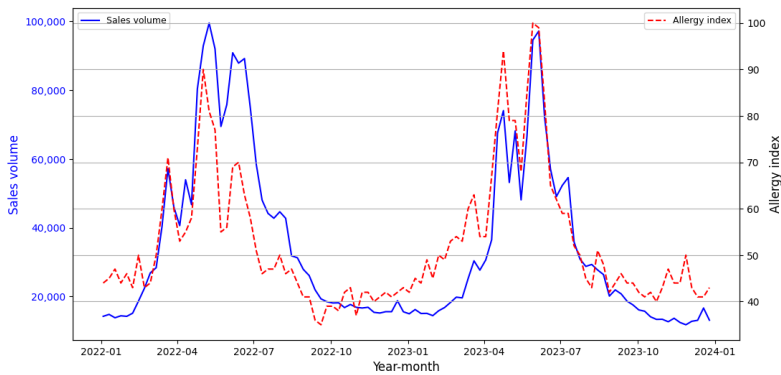


Figure 14: Trend of the allergy index

Figure 15 illustrates the number of public holidays in each week. Over the 104-week period, there were no public holidays in 82 weeks, one public holiday in 19 weeks, and two public holidays in only three weeks.

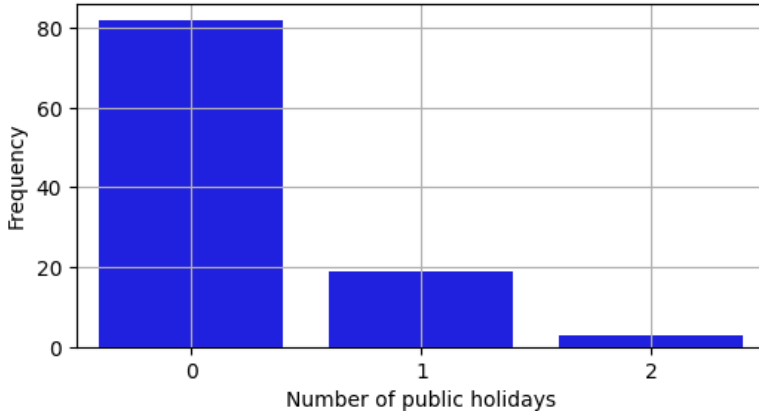


Figure 15: Number of public holidays

In a typical week, Allegra is sold at varying prices across different locations. To monitor these prices, Opella sets a reference price and expresses the weekly price as a discount relative to the 90th percentile price of all medicines sold. Figure 16 illustrates the trend of the 90th percentile price over the two-year period. Overall, the price shows an upward trend, likely reflecting the impact of inflation. While the first peak in sales volume during May and June 2022 is not reflected in the price trend, a significant price drop follows the second peak in sales volume in May and June 2023. This decline is likely due to Opella attempting to clear excess stock, as the company had overproduced Allegra during the demand peak in 2023.

## B Appendix: Detailed Model Formulation

We define the **Bayesian Marketing Mix Model (MMM)** to estimate the impact of media spend on sales while accounting for carryover, saturation effects, trend, and seasonality. Time  $t$  is measured in weeks, with data typically collected over 2 to 3 years.

**Likelihood Function** We model the observed sales volume  $Y_t$  at time  $t$  as:

$$Y_t \sim \mathcal{N}(\mu_t, \sigma^2) \quad (15)$$

where:

- $\mu_t$  is the expected sales at time  $t$ ,



Figure 16: Trend of the 90th percentile price

- $\sigma^2$  is the variance of the residual noise.

The expected sales  $\mu_t$  is given by:

$$\begin{aligned} \mu_t = & \alpha + \nu t^\kappa + \sum_{d=1}^2 \left( \gamma_{1,d} \cos\left(\frac{2\pi dt}{52}\right) + \gamma_{2,d} \sin\left(\frac{2\pi dt}{52}\right) \right) \\ & + \sum_{m=1}^M \beta_m \cdot \text{Hill}(X_{t,m}^*; \mathcal{K}_m, \mathcal{S}_m) + \sum_{c=1}^C \gamma_c X_{t,c} \end{aligned} \quad (16)$$

**Trend Component** The trend is captured as a power law:

$$\nu \sim \mathcal{N}(0, 1), \quad (17)$$

$$\kappa \sim \text{Uniform}(0.5, 1.5), \quad (18)$$

where  $\nu$  scales the trend and  $\kappa$  controls its nonlinearity.

**Seasonality Component** Seasonal variation is captured using a second-order Fourier expansion:

$$\sum_{d=1}^2 \left( \gamma_{1,d} \cos\left(\frac{2\pi dt}{52}\right) + \gamma_{2,d} \sin\left(\frac{2\pi dt}{52}\right) \right), \quad (19)$$

with:

$$\gamma_{1,d}, \gamma_{2,d} \sim \mathcal{N}(0, 1). \quad (20)$$

**Adstock Transformation (Carryover Effect)** The carryover effect is modeled recursively:

$$X_{t,m}^* = X_{t,m} + \eta_m X_{t-1,m}^*, \quad X_{1,m}^* = X_{1,m}, \quad (21)$$

where  $\eta_m \sim \text{Beta}(2, 1)$ .

**Hill Function (Saturation Effect)** The nonlinear effect of media spend is modeled using the Hill function:

$$\text{Hill}(X_{t,m}^*; \mathcal{K}_m, \mathcal{S}_m) = \frac{X_{t,m}^*}{1 + \left(\frac{X_{t,m}^*}{\mathcal{K}_m}\right)^{-\mathcal{S}_m}}, \quad (22)$$

where:

$$\mathcal{K}_m \sim \text{Gamma}(1, 1), \quad (23)$$

$$\mathcal{S}_m \sim \text{Gamma}(1, 1). \quad (24)$$

**Media Coefficients.** Each media coefficient  $\beta_m$  follows a Truncated Normal distribution:

$$\beta_m \sim \text{TruncNormal}(0, v_m; 0, \infty), \quad v_m = \sum_{t=1}^N X_{t,m}. \quad (25)$$

This prior allows higher-spend channels to exhibit greater variability in estimated effects, while shrinking low-spend channels more strongly toward zero.

**Other Priors.**

$$\alpha \sim \text{TruncNormal}(0, 2; 0, \infty), \quad (26)$$

$$\gamma_c \sim \mathcal{N}(0, 1), \quad (27)$$

$$\sigma^2 \sim \text{InvGamma}(0.05, 0.0005). \quad (28)$$

**Full Model Summary.** For completeness, we summarize the full model below:

$$Y_t \sim \mathcal{N}(\mu_t, \sigma^2), \quad (29)$$

$$\begin{aligned} \mu_t = \alpha + \nu t^\kappa + \sum_{d=1}^2 \left( \gamma_{1,d} \cos\left(\frac{2\pi dt}{52}\right) + \gamma_{2,d} \sin\left(\frac{2\pi dt}{52}\right) \right) \\ + \sum_{m=1}^M \beta_m \cdot \frac{X_{t,m}^*}{1 + \left(\frac{X_{t,m}^*}{\mathcal{K}_m}\right)^{-\mathcal{S}_m}} + \sum_{c=1}^C \gamma_c X_{t,c}, \end{aligned} \quad (30)$$

$$X_{t,m}^* = X_{t,m} + \eta_m X_{t-1,m}^*, \quad X_{1,m}^* = X_{1,m}, \quad (31)$$

$$\beta_m \sim \text{TruncNormal}(0, v_m; 0, \infty), \quad v_m = \sum_{t=1}^N X_{t,m}, \quad (32)$$

$$\alpha \sim \text{TruncNormal}(0, 2; 0, \infty), \quad \nu \sim \mathcal{N}(0, 1), \quad \kappa \sim \text{Uniform}(0.5, 1.5), \quad (33)$$

$$\gamma_{1,d}, \gamma_{2,d} \sim \mathcal{N}(0, 1), \quad \gamma_c \sim \mathcal{N}(0, 1), \quad (34)$$

$$\eta_m \sim \text{Beta}(2, 1), \quad \mathcal{K}_m \sim \text{Gamma}(1, 1), \quad \mathcal{S}_m \sim \text{Gamma}(1, 1), \quad (35)$$

$$\sigma^2 \sim \text{InvGamma}(0.05, 0.0005). \quad (36)$$

## C Detailed Simulation Settings

We describe below the specific prior settings and hyperparameters used in the simulation studies.

### Base Model

For the base model with independent priors on the effectiveness parameters, we use:

$$\beta_m \sim \mathcal{N}(1, 1). \quad (37)$$

### Normal-Inverse Gamma Prior

For the Normal-Inverse Gamma prior, we use the following hierarchical structure:

$$\begin{aligned} \beta_m \mid \mu_c, \tau_c^2, \lambda_m^2, \sigma^2 &\sim \mathcal{N}(\mu_c, \tau_c^2 \lambda_m^2 \sigma^2), \\ \mu_c \mid \sigma_\mu^2 &\sim \mathcal{N}(1, 1/2), \\ \tau_c^2 \mid a_\tau, b_\tau &\sim \mathcal{IG}(3, 3/2), \\ \lambda_m^2 \mid a_\lambda, b_\lambda &\sim \mathcal{IG}(1/2, 1/2). \end{aligned} \quad (38)$$

## Horseshoe Prior

For the Horseshoe prior, we use:

$$\begin{aligned}
 \beta_m \mid \mu_c, \tau_c^2, \lambda_m^2, \sigma^2 &\sim \mathcal{N}(\mu_c, \tau_c^2 \lambda_m^2 \sigma^2), \\
 \mu_c \mid \sigma_\mu^2 &\sim \mathcal{N}(1, 1/2), \\
 \tau_c^2 &\sim \mathcal{C}^+(0, 1), \\
 \lambda_m^2 &\sim \mathcal{C}^+(0, 1).
 \end{aligned}
 \tag{39}$$

## Other Parameters

The following hyperpriors are shared across all models:

$$\begin{aligned}
 \gamma &\sim \mathcal{N}(0, 2), \\
 \sigma &\sim \text{Gamma}(2, 1/2), \\
 \eta_m &\sim \text{Beta}(2, 1), \\
 \mathcal{K}_m &\sim \text{Gamma}(1, 1), \\
 \mathcal{S}_m &\sim \text{Gamma}(1, 1).
 \end{aligned}
 \tag{40}$$

# D Stan implementation

## D.1 Base Model (Independent Prior)

```

1 functions {
2   real hill_function(real x, real kappa, real s) {
3     return x / (1 + pow(x / kappa, -s));
4   }
5
6   real carryover_step(real x_t, real xstar_prev, real eta) {
7     return x_t + eta * xstar_prev;
8   }
9 }
10
11 data {
12   int<lower=1> n; // Number of time points
13   int<lower=1> m; // Number of media channels
14   matrix[n, m] X; // Raw media inputs
15   vector[n] Z; // Control variable
16   vector[n] Y; // Outcome variable
17 }
18
19 parameters {
20   vector[m] beta; // Media coefficients
21   real gamma; // Control coefficient
22   real<lower=0> sigma; // Noise SD
23   vector<lower=0, upper=1>[m] eta; // Carryover parameters
24   vector<lower=0>[m] kappa; // Hill half-saturation

```

```

25   vector<lower=0>[m] s;           // Hill steepness
26 }
27
28 transformed parameters {
29   matrix[n, m] X_star;
30
31   for (j in 1:m) {
32     X_star[1, j] = X[1, j];
33     for (t in 2:n) {
34       X_star[t, j] = carryover_step(X[t, j], X_star[t - 1, j], eta[j]);
35     }
36   }
37 }
38
39 model {
40
41   // Priors
42   beta ~ normal(1, 1);
43   gamma ~ normal(0, 1);
44   sigma ~ gamma(2, 1.0 / 2.0);
45   eta ~ beta(2, 1);
46   kappa ~ gamma(1, 1);
47   s ~ gamma(1, 1);
48
49   // Likelihood
50   for (t in 1:n) {
51     real mu = 0;
52     for (j in 1:m) {
53       mu += beta[j] * hill_function(X_star[t, j], kappa[j], s[j]);
54     }
55     mu += gamma * Z[t];
56     Y[t] ~ normal(mu, sqrt(sigma));
57   }
58 }

```

Listing 1: Stan code for the base model using independent priors.

## D.2 Hierarchical Model (Normal-Inverse Gamma Prior)

```

1 functions {
2   real hill_function(real x, real kappa, real s) {
3     return x / (1 + pow(x / kappa, -s));
4   }
5
6   real carryover_step(real x_t, real xstar_prev, real eta) {
7     return x_t + eta * xstar_prev;
8   }
9 }
10
11 data {
12   int<lower=1> n;           // Number of time points
13   int<lower=1> m;         // Number of media channels
14   matrix[n, m] X;        // Raw media inputs
15   vector[n] Z;           // Control variable

```

```

16   vector[n] Y; // Outcome variable
17   array[m] int<lower=0, upper=1> group_indicator; // Group assignment (0 or 1)
18 }
19
20 parameters {
21   // Non-centered parameterization for beta
22   vector[m] beta_raw;
23
24   // Hierarchical parameters
25   real mu_0; // Group 0 mean effectiveness
26   real mu_1; // Group 1 mean effectiveness
27   real<lower=0> tau_0_sq; // Group 0 global shrinkage
28   real<lower=0> tau_1_sq; // Group 1 global shrinkage
29   vector<lower=0>[m] lambda_sq; // Local shrinkage
30
31   real gamma; // Control coefficient
32   real<lower=0> sigma; // Noise SD
33   vector<lower=0, upper=1>[m] eta; // Carryover parameters
34   vector<lower=0>[m] kappa; // Hill half-saturation
35   vector<lower=0>[m] s; // Hill steepness
36 }
37
38 transformed parameters {
39   vector[m] beta;
40   matrix[n, m] X_star;
41
42   // Non-centered beta transformation
43   for (j in 1:m) {
44     if (group_indicator[j] == 0)
45       beta[j] = mu_0 + sqrt(tau_0_sq) * sqrt(lambda_sq[m]) * beta_raw[j];
46     else
47       beta[j] = mu_1 + sqrt(tau_1_sq) * sqrt(lambda_sq[m]) * beta_raw[j];
48   }
49
50   // Compute X_star recursively
51   for (j in 1:m) {
52     X_star[1, j] = X[1, j];
53     for (t in 2:n) {
54       X_star[t, j] = carryover_step(X[t, j], X_star[t - 1, j], eta[j]);
55     }
56   }
57 }
58
59 model {
60
61   // Hierarchical priors
62   mu_0 ~ normal(1, 1.0/2.0);
63   mu_1 ~ normal(1, 1.0/2.0);
64   tau_0_sq ~ inv_gamma(3, 1.5);
65   tau_1_sq ~ inv_gamma(3, 1.5);
66   lambda_sq ~ inv_gamma(0.5, 0.5);
67
68   // Standard normal prior for beta_raw
69   beta_raw ~ normal(0, 1);
70

```

```

71 // Other priors
72 gamma ~ normal(0, 1);
73 sigma ~ gamma(2, 1.0 / 2.0);
74 eta ~ beta(2, 1);
75 kappa ~ gamma(1, 1);
76 s ~ gamma(1, 1);
77
78 // Likelihood
79 for (t in 1:n) {
80   real mu = 0;
81   for (j in 1:m) {
82     mu += beta[j] * hill_function(X_star[t, j], kappa[j], s[j]);
83   }
84   mu += gamma * Z[t];
85   Y[t] ~ normal(mu, sqrt(sigma));
86 }
87 }

```

Listing 2: Stan code for the hierarchical model with Normal-Inverse Gamma prior.

### D.3 Hierarchical Model with Horseshoe Prior

```

1 functions {
2   real hill_function(real x, real kappa, real s) {
3     return x / (1 + pow(x / kappa, -s));
4   }
5
6   real carryover_step(real x_t, real xstar_prev, real eta) {
7     return x_t + eta * xstar_prev;
8   }
9 }
10
11 data {
12   int<lower=1> n; // Number of time points
13   int<lower=1> m; // Number of media channels
14   matrix[n, m] X; // Raw media inputs
15   vector[n] Z; // Control variable
16   vector[n] Y; // Outcome variable
17   array[m] int<lower=0, upper=1> group_indicator; // Group assignment (0 or 1)
18 }
19
20 parameters {
21
22   // Non-centered parameterization for beta
23   vector[m] beta_raw;
24
25   // Hierarchical parameters
26   real mu_0; // Group 0 mean effectiveness
27   real mu_1; // Group 1 mean effectiveness
28   real<lower=0> tau_0_sq; // Group 0 global shrinkage
29   real<lower=0> tau_1_sq; // Group 1 global shrinkage
30   vector<lower=0>[m] lambda_sq; // Local shrinkage
31   real<lower=0> aux_tau_0; // Auxilliary variable tau_0
32   real<lower=0> aux_tau_1; // Auxilliary variable tau_1

```

```

33   vector<lower=0>[m] aux_lambda;           // Auxilliary variable lambda
34
35   real gamma;                             // Control coefficient
36   real<lower=0> sigma;                     // Noise SD
37   vector<lower=0, upper=1>[m] eta;        // Carryover parameters
38   vector<lower=0>[m] kappa;               // Hill half-saturation
39   vector<lower=0>[m] s;                   // Hill steepness
40 }
41
42 transformed parameters {
43   vector[m] beta;
44   matrix[n, m] X_star;
45
46   // Non-centered beta transformation
47   for (j in 1:m) {
48     if (group_indicator[j] == 0)
49       beta[j] = mu_0 + sqrt(tau_0_sq) * sqrt(lambda_sq[m]) * beta_raw[j];
50     else
51       beta[j] = mu_1 + sqrt(tau_1_sq) * sqrt(lambda_sq[m]) * beta_raw[j];
52   }
53
54   // Compute X_star recursively
55   for (j in 1:m) {
56     X_star[1, j] = X[1, j];
57     for (t in 2:n) {
58       X_star[t, j] = carryover_step(X[t, j], X_star[t - 1, j], eta[j]);
59     }
60   }
61 }
62
63 model {
64
65   // Hierarchical priors
66   mu_0 ~ normal(1, 1.0/2.0);
67   mu_1 ~ normal(1, 1.0/2.0);
68   tau_0_sq ~ inv_gamma(1.0 / 2.0, 1.0 / aux_tau_0);
69   tau_1_sq ~ inv_gamma(1.0 / 2.0, 1.0 / aux_tau_1);
70   lambda_sq ~ inv_gamma(1.0 / 2.0, 1.0 / aux_lambda);
71   aux_tau_0 ~ inv_gamma(1.0 / 2.0, 1);
72   aux_tau_1 ~ inv_gamma(1.0 / 2.0, 1);
73   aux_lambda ~ inv_gamma(1.0 / 2.0, 1);
74
75   // Standard normal prior for beta_raw
76   beta_raw ~ normal(0, 1);
77
78   // Other priors
79   gamma ~ normal(0, 1);
80   sigma ~ gamma(2, 1.0 / 2.0);
81   eta ~ beta(2, 1);
82   kappa ~ gamma(1, 1);
83   s ~ gamma(1, 1);
84
85   // Likelihood
86   for (t in 1:n) {
87     real mu = 0;

```

---

```
88     for (j in 1:m) {
89         mu += beta[j] * hill_function(X_star[t, j], kappa[j], s[j]);
90     }
91     mu += gamma * Z[t];
92     Y[t] ~ normal(mu, sqrt(sigma));
93 }
94 }
```

Listing 3: Stan code for the horseshoe model using auxiliary variable representation.