Decomposing Sales via Bayesian Regression

Drue Jaramillo
Department of Applied Mathematics & Statistics
Johns Hopkins University
Baltimore, MD 21218
djarami3@jh.edu

Abstract

Marketing mix modeling (MMM) focuses on determining the relationship between sales and marketing efforts in order to optimize marketing budgets and maximize return on ad spend (ROAS). We consider the approach of a simple Bayesian normal regression model, but with an added first-order autoregressive structure determined by an unknown parameter. After fitting the model to a simulated marketing mix dataset using MCMC methods, we discuss the applications of the model and further improvements that may be made.

1 Introduction

In the modern age, marketing is a critical component of every business's long-term growth and success. Naturally, questions arise as to how a company could spend their capital with the greatest effectiveness. However, this is complicated by the sheer number of possible predictors of marketing success (often measured by sales). Between the dozens of available marketing channels such as direct mail, social media, radio, and search engine marketing (SEM); the hundreds of confounding variables such as holidays, macroeconomic indicators, and competitor's marketing efforts; and the interactions between all of the above, the set of all possible models cannot be searched effectively.

Further complications arise in MMM due to the lack of available data. In order to collect the necessary data, a company must frequently track all of the predictors it wishes to use, as well as its measure of success, such as sales. While the lack of data may be solved by a few years' worth of daily data, this often comes at the cost of a high variance in the data. Hence, many companies opt for less-frequent-but-less-variable data.

Considering all of the above, it becomes clear that a marketing mix model should have three goals: simplicity, interpretability, and a posterior distribution. Simplicity is desired since we may easily have far more predictors than data points, so we should aim to avoid overfitting by using only a manageable subset of the possible predictors. Interpretability is desired since this type of model is highly valuable to key decision-makers and investors, so it should be relatively easy to explain the results to those with little mathematical background. Lastly, a posterior distribution is desired so that we can deal with a wider, more practical variety of point estimates, credible intervals, and probabilities. All of these conditions are well-satisfied by a Bayesian normal regression model.

1.1 The data

Marketing spending and sales tend to be closely-guarded secrets among businesses since they reveal sensitive information that a competitor could easily use to their advantage. Accordingly, companies do not tend to publish this information to the public. Therefore, in order to study this type of problem without access to private information, we must study a simulated dataset. In this paper, we will be looking at the dataset created for He [2020]. It contains weekly data over 4 years consisting of spending for 13 different marketing channels, the Consumer Price Index, the average price of gas, the

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

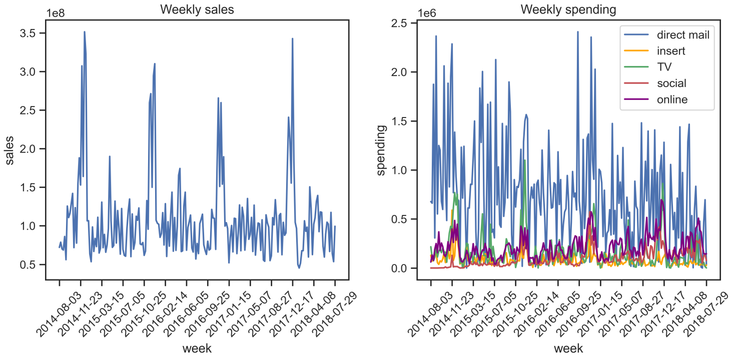

Figure 1: Weekly sales and spending

number of stores our fictional company has, and indicator variables for all major holidays. A plot of the weekly sales and weekly spending for some of the channels can be seen in 1.

1.2 Related work

Fong and DeSarbo [2007] used a Bayesian multiple regression model to determine regime changes and thereby determine the effects of marketing promotions. On the other hand, Ng et al. [2021] formulated a hierarchical Bayesian time-varying coefficient model to try and address the issues seen in many MMM approaches. Outside the Bayesian paradigm, several traditional optimization methods have been attempted, including: a Quadratic Knapsack Problem formulation by Pferschy et al. [2015]; a dynamic, hierarchical allocation model by Fischer et al. [2011]; and a mixed-integer nonlinear program by Huang et al. [2021]. However, the most popular approach appears to be the use of Bayesian structural time series, which allow for time series data to be represented as the sum of multiple different time series that represent various factors, such as trend, seasonality, and regression components (see Scott and Varian [2014], Brodersen et al. [2015], Scott and Varian [2019]).

2 The model

We assume that sales is some linear function of the predictors and has Gaussian noise. By itself, this could be represented in matrix form by the sampling model:

$$Y \sim \mathcal{N}_n(X\beta, \sigma^2 I) \tag{1}$$

where Y is the $n \times 1$ vector of observations, X is the $n \times p$ matrix of regressors, β is the $p \times 1$ vector of regression coefficients, and σ^2 is the common variance of the data. However, our intuition indicates that Y_i and Y_j are not necessarily uncorrelated for all $i \neq j$. In fact, we would expect sales to have a bit of momentum, so that next week's sales are somewhat similar to this week's sales. We an incorporate this into our model by using a modified version of (1):

$$Y \sim \mathcal{N}_n(X\beta, \sigma^2 C_\rho) \tag{2}$$

where C_ρ is the matrix with first-order autoregressive structure determined by the parameter ρ:

$$C_\rho = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & \cdots & 1 \end{bmatrix}$$

Hence, the variance of Y_i is σ^2 for each i, but now the correlation betweeen Y_i and Y_j is $\rho^{|j-i|}$, which diminishes to 0 as $|j-i|$ grows.

2.1 Prior distributions

We will use standard conjugate priors for β and σ:

$$\beta \sim \mathcal{N}_p(\beta_0, \Sigma_0) \tag{3}$$

$$\sigma^2 \sim \text{inverse-gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right) \tag{4}$$

In order to make these priors weakly-informative, we will set $\beta_0 = 0$ and $\nu_0 = 3$, as well as make Σ_0 a diagonal matrix with large entries so that is relatively diffuse. In addition, we will set σ_0^2 equal to our prior expectation of σ^2, which is around $(1,000,000)^2$. On the other hand, there is no conjugate prior for ρ since it has no closed-form full posterior conditional. Thus, we will choose the weakly-informative uniform prior given by

$$\rho \sim \text{beta}(1, 1) \tag{5}$$

2.2 Posterior distributions

Adapting the results of Hoff [2009], we find that the full conditional posteriors of β and σ^2 are

$$\{\beta | X, y, \sigma^2, \rho\} \sim \mathcal{N}_p(\beta_n, \Sigma_n) \tag{6}$$

$$\{\sigma^2 | X, y, \beta, \rho\} \sim \text{inverse-gamma}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + SSR_\rho}{2}\right) \tag{7}$$

where y is the $n \times 1$ vector of observed data and

$$\Sigma_n = (X^T C_\rho^{-1} X/\sigma^2 + \Sigma_0^{-1})^{-1}$$

$$\beta_n = \Sigma_n(X^T C_\rho^{-1} y/\sigma^2 + \Sigma_0^{-1}\beta_0)$$

$$SSR_\rho = (y - X\beta)^T C_\rho^{-1}(y - X\beta)$$

On the other hand, there is no closed-form for the full conditional posterior of ρ, but as we will see in section (3), we don't require one.

3 MCMC sampling

We will approximate the posterior distribution using a Metropolis-Hastings algorithm that performs Gibbs sampling to update β and σ^2 and performs a Metropolis step to update ρ. Our MCMC algorithm proceeds as follows:

1. Sample $\beta^{(s+1)} \sim p(\beta | X, y, \sigma^{2(s)}, \rho^{(s)})$.
2. Sample $\sigma^{2(s+1)} \sim p(\sigma^2 | X, y, \beta^{(s+1)}, \rho^{(s)})$.
3. Sample $\rho^{(s+1)}$:
   (a) Propose $\rho^* \sim \text{uniform}(\rho^{(s)} - \delta, \rho^{(s)} + \delta)$.
   (b) If $\rho^* < 0$, then set it equal to $|\rho^*|$. If $\rho^* > 1$, then set it equal to $2 - \rho^*$.
   (c) Compute the acceptance ratio:
   $$r = \frac{p(y|X, \beta^{(s+1)}, \sigma^{2(s+1)}, \rho^*)p(\rho^*)}{p(y|X, \beta^{(s+1)}, \sigma^{2(s+1)}, \rho^{(s)})p(\rho^{(s)})}$$
   and sample $u \sim \text{uniform}(0, 1)$.
   (d) If $u < r$, then set $\rho^{(s+1)} = \rho^*$. Otherwise, set $\rho^{(s+1)} = \rho^{(s)}$.

4 Simulation study

In order to verify that our algorithm works and does a decent job of approximating the true parameter values, it is necessary to simulate data according to our model and then run our algorithm on the simulated data. Our simulated regressors consisted of the same non-spending variables as the real data, but the spend data was replaced with multivariate Gaussian noise with means ranging from 25,000 to 125,000 and a correlation between successive timesteps of 0.7. Then, we created an arbitrary β vector and simulated new data y according to (2) with $\rho = 0.6$. The resulting weekly sales and weekly spending for some of the marketing channels can be seen in 2

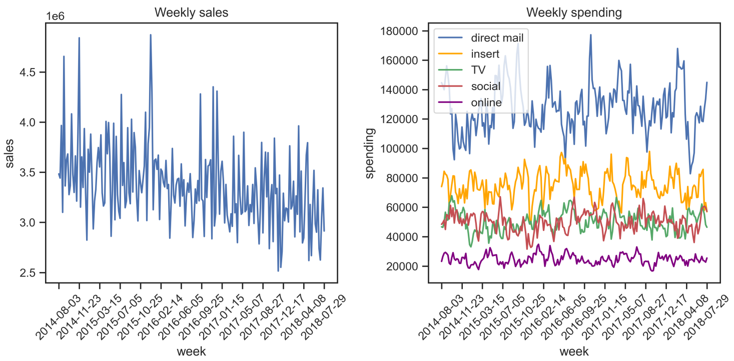

Figure 2: Weekly sales and spending (simulated)

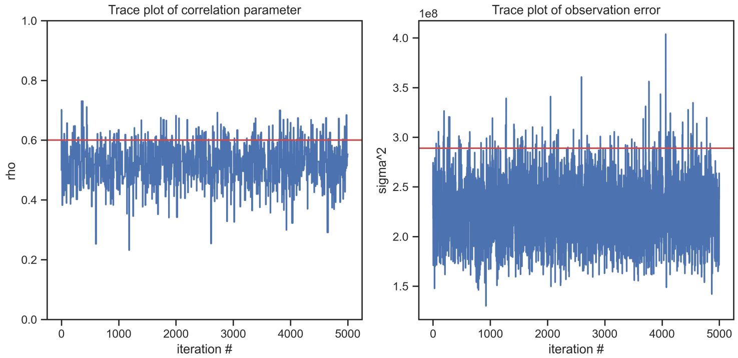

Figure 3: Simulation results

4.1 Simulation results

Beginning with arbitrary initial values, we ran the MCMC algorithm given in section (3) for 5,000 iterations. Furthermore, in accordance with the research presented by Sherlock and Roberts [2009] and Gelman et al. [1997], we tuned δ so that the acceptance rate for $\rho^{(s+1)}$ would be close to 0.234. The simulation ran for 121 seconds with an acceptance rate for $\rho^{(s+1)}$ of 0.2268.

The trace plots in 3 seem to indicate that the Markov chain has converged, and it appears to approximate both parameters decently well. In fact, given that our priors had lower values for ρ and σ^2 than the simulated data, the underestimation we see is well within reason. Altogether, it would appear that our algorithm works, so we will proceed.

5 Empirical results

Beginning with arbitrary initial values again, we ran the same MCMC algorithm for 100,000. We discarded a burn-in period of 5,000 iterations and then kept every 10th iteration from then on, leaving us with 9,500 samples in the end. We kept the same δ as we had for the simulated data, but it did not provide the same near-optimal acceptance rate. The simulation ran for 2,807 seconds with an acceptance rate for $\rho^{(s+1)}$ of 0.41377.

The trace plots in 4 also seem to indicate that the overall Markov chain has converged. This is further evidenced by the autocorrelation plots in 5, which show that the individual Markov chains have nearly 0 autocorrelation for all lags. Lastly, the effective samples sizes in 1 for all parameters are above 8,000, which should be more than enough to approximate the posterior distribution(s).

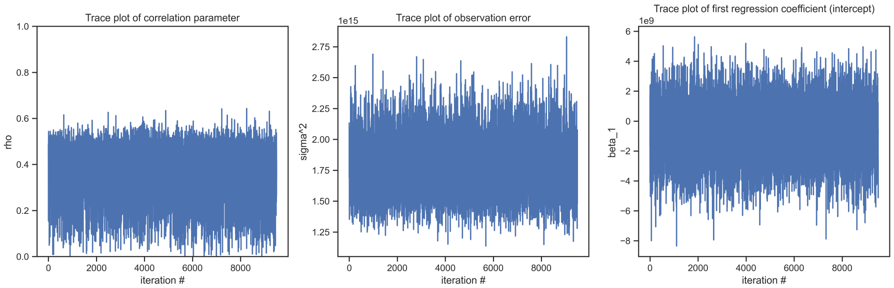

Figure 4: Trace plots for ρ, σ^2, and β_2

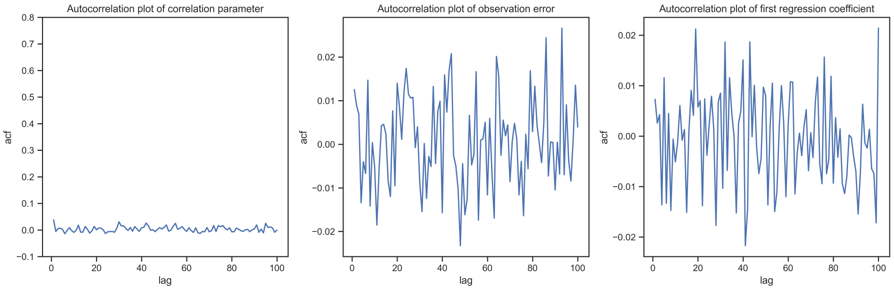

Figure 5: Autocorrelation plots for ρ, σ^2, and β_2

5.1 Interpretation

The sales data appears to have some serial correlation, with ρ most almost certainly between 0.2 and 0.5. Furthermore, the sales data appears to have quite large variance, although this may be due to the model misintepreting the large spikes as not being due to the regression components. Between the two of these facts, it would appear that the data is likely generated by some sort of temporally-correlated process (such as a serially correlated regression), but that it does not match the type of our model. Indeed, this is the case when we reference He [2020].

5.2 Posterior predictive distribution

As a quick note: If we have a Markov chain of parameter values $\phi_1, \ldots, \phi_S$, and we have a new set of regressors $\tilde{X}$, then we can easily generate a posterior predictive distribution by plugging ϕ_i and $\tilde{X}$ into (2) for each i and generating a Markov chain of predicted values $\tilde{y}_1, \ldots, \tilde{y}_S$. This would be particularly useful for generating credible intervals around predicted sales, rather than just giving point estimates.

6 Conclusion

Now that we have the regression coefficients, we could estimate the return on ad spend (ROAS) and marginal return on ad spend (mROAS) of each marketing channel. The ROAS of the channels tell us which channels are more effective than others at producing sales, and thereby give us a way to

Table 1: Effective sample sizes

Parameter	ESS
ρ	8,808
σ^2	9,500
β	8,393 (smallest)

evaluate the performance of our current marketing efforts. The mROAS of the channels tell us which channels will increase (decrease) sales the most (least) if we increase (decrease) spending in those channels. This is the key to the ultimate goal of MMM: optimizing our marketing budget.

However, while the Bayesian normal regression model appears to do approximate the sales impact of marketing efforts decently well, there is still plenty of room for improvement. It is a rather simplistic model that could benefit from a bit more complexity. In particular, it could benefit from utilizing non-linearity, which is something we would expect to see given the diminishing marginal returns of investments, especially those in marketing. This could be accomplished in many ways, including via Gaussian processes or Bayesian structural time series.

References

Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using bayesian structural time-series models. *The annals of applied statistics*, 9(1):247–274, Mar 1, 2015. URL `https://www.jstor.org/stable/24522418`.

Marc Fischer, Sonke Albers, Nils Wagner, and Monika Frie. Dynamic marketing budget allocation across countries, products, and marketing activities. *Marketing science (Providence, R.I.)*, 30(4):568–585, Jul 1, 2011. URL `https://www.jstor.org/stable/23012011`.

Duncan Fong and Wayne DeSarbo. A bayesian methodology for simultaneously detecting and estimating regime change points and variable selection in multiple regression models for marketing research. *Quantitative marketing and economics*, 5(4):427–453, 2007. URL `http://econpapers.repec.org/article/kapqmktec/v_3a5_3ay_3a2007_3ai_3a4_3ap_3a427-453.htm`.

A. Gelman, W. R. Gilks, and G. O. Roberts. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of applied probability*, 7(1):110–120, Feb 1, 1997. URL `https://www.jstor.org/stable/2245134`.

Sibyl He. Python/stan implementation of multiplicative marketing mix model, Dec 1, 2020. URL `https://towardsdatascience.com/python-stan-implementation-of-multiplicative-marketing-mix-model-with-deep-dive-into-adstock-a7320865b33`

Peter D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. Springer New York, New York, NY, 1. aufl. edition, 2009. ISBN 9780387922997. URL `https://library.biblioboard.com/viewer/f661bed6-c05c-11ea-8686-0a28bb48d135`.

Hsin-Chan Huang, Jiefeng Xu, and Alvin Lim. Marketing mix optimization with practical constraints. Jan 10, 2021. URL `https://arxiv.org/abs/2101.03663`.

Edwin Ng, Zhishi Wang, and Athena Dai. Bayesian time varying coefficient model with applications to marketing mix modeling. Technical report, Cornell University Library, arXiv.org, Sep 5, 2021. URL `https://search.proquest.com/docview/2538895522`.

Ulrich Pferschy, Joachim Schauer, and Gerhild Maier. *A Quadratic Knapsack Model for Optimizing the Media Mix of a Promotional Campaign*, pages 251–264. Operations Research and Enterprise Systems. Springer International Publishing, Cham, Apr 17, 2015. ISBN 3319175084. URL `http://link.springer.com/10.1007/978-3-319-17509-6_17`.

Steven L. Scott and Hal R. Varian. Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2):4–23, Jan 1, 2014. URL `https://www.inderscienceonline.com/doi/10.1504/IJMMNO.2014.059942`.

Steven L. Scott and Hal R. Varian. *Bayesian Variable Selection for Nowcasting Economic Time Series*, pages 119–136. Economic Analysis of the Digital Economy. University of Chicago Press, Chicago, 2019. URL `http://www.degruyter.com/doi/10.7208/9780226206981-007`.

Chris Sherlock and Gareth Roberts. Optimal scaling of the random walk metropolis on elliptically symmetric unimodal targets. *Bernoulli : official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 15(3):774–798, Aug 1, 2009. URL `https://www.jstor.org/stable/20680177`.