Web Appendix

Your MMM is Broken: Identification of Nonlinear and Time-varying Effects in Marketing Mix Models

Ryan Dew, Nicolas Padilla, Anya Shchetkina

August 14, 2024

Contents

A Additional Derivations

A.1 Time-varying Approximates Nonlinear

Here, we present the derivation of the OLS approximation to a nonlinear DGP.

Assumption 1 (DGP).

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

We state explicitly the relationship of x_i and ε_i, which will be useful when assessing the convergence of the OLS estimator.

Remark 1 (Uncorrelated errors). $\mathbf{E}(\varepsilon_i \cdot x_i) = 0$ *(directly from the DGP).*

We require f to be differentiable for the Taylor expansion.

Assumption 2 (Differentiable). *f is differentiable up to order $K > 2$.*

Under Assumption 2 we can write f using its Taylor expansion of degree 1 around $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + R_1(x) \tag{A-1}$$

where

$$\begin{aligned} R_1(x) &= \frac{f''(\xi)}{2}(x - \bar{x})^2 && \text{(Lagrange form)} \\ &= \int_{\bar{x}}^{x} \frac{f''(t)}{1!}(x - t)dt && \text{(Integral form)} \end{aligned}$$

with ξ between x and $\bar{x}$.

Now, suppose we took data generated from this process, and approximated it with a local linear model, such that:

$$y_i = b_0 + b_1 x_i + \epsilon_i, \tag{A-2}$$

for a region around x_i. The corresponding OLS estimator is given by:

$$\begin{aligned} \hat{b}_1 &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ \hat{b}_0 &= \bar{y} - \hat{b}_1 \bar{x} \end{aligned} \tag{A-3}$$

Note that by definition

$$\epsilon_i = y_i - (b_0 + b_1 x_i) = [f(x_i) - (b_0 + b_1 x_i)] + \varepsilon_i. \tag{A-4}$$

We can rewrite (A-3) as well using (A-1)

$$\begin{aligned}
\sum_i (x_i - \bar{x})(y_i - \bar{y}) &= \sum_i (x_i - \bar{x}) \left[f(x_i) + \varepsilon_i \right] - \bar{y} \sum_i (x_i - \bar{x}) \\
&= \sum_i (x_i - \bar{x}) \left[f(\bar{x}) + f'(\bar{x})(x_i - \bar{x}) + R_1(x_i) + \varepsilon_i \right] \\
&= f'(\bar{x}) \sum_i (x_i - \bar{x})^2 + \sum_i (x_i - \bar{x}) R_1(x_i) + \sum_i (x_i - \bar{x}) \varepsilon_i \\
&= f'(\bar{x}) \sum_i (x_i - \bar{x})^2 + \frac{1}{2} \sum_i f''(\xi_i)(x_i - \bar{x})^3 + \sum_i (x_i - \bar{x}) \varepsilon_i \qquad \text{(A-5)}
\end{aligned}$$

Now replacing (A-5) in (A-3), we have

$$\begin{aligned}
\hat{b}_1 &= \frac{1}{\sum_i (x_i - \bar{x})^2} \left[f'(\bar{x}) \sum_i (x_i - \bar{x})^2 + \frac{1}{2} \sum_i f''(\xi_i)(x_i - \bar{x})^3 + \sum_i (x_i - \bar{x}) \varepsilon_i \right] \\
&= f'(\bar{x}) + \frac{1}{2} \frac{\frac{1}{n} \sum_i f''(\xi_i)(x_i - \bar{x})^3}{\frac{1}{n} \sum_i (x_i - \bar{x})^2} + \frac{\frac{1}{n} \sum_i (x_i - \bar{x}) \varepsilon_i}{\frac{1}{n} \sum_i (x_i - \bar{x})^2} \qquad \text{(A-6)}
\end{aligned}$$

A.2 Nonlinear Approximates Time-varying

Consider a true data generating process given (as before) by,

$$y_t = \beta_t \cdot x_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \qquad \text{(A-7)}$$

where we assume that the time-varying effectiveness, β_t, can be written as $B(t)$, a continuous function.

Proposition 1. *Marketing effectiveness, β_t, can be expressed as a function of spending, x_t, for all non-null spending, i.e., $B(t) = h(x_t)\ \forall t : x_t \neq 0$ for some function $h : \mathbb{R} \setminus \{0\} \to \mathbb{R}$; if and only if the DGP in Equation A-7 can be written as*

$$y_t = f(x_t) + \varepsilon_t, \ \ \forall t \qquad \text{(A-8)}$$

where $f(x) = \begin{cases} h(x) \cdot x, & x \neq 0 \\ 0 & x = 0. \end{cases}$

Note that the function doesn't have to exist for null values of spending (since β_t does not matter in that case, any value would work).

Proof. $\Rightarrow$ Consider $x_t \neq 0$. Just replacing $\beta_t = h(x_t)$, in Equation A-7 we have

$$y_t = \beta_t \cdot x_t + \varepsilon_t = h(x_t) \cdot x_t + \varepsilon_t$$

Now for $x_t = 0$, we have that $y_t = \beta_t \cdot 0 + \varepsilon_t = 0 + \varepsilon_t$. Finally, if we define

$$f(x) = \begin{cases} h(x) \cdot x, & x \neq 0 \\ 0 & x = 0. \end{cases}$$

we have that $y_t = f(x_t) + \varepsilon_t$.

$\Leftarrow$ Assume there is a function $f(x)$, such that $\forall t: \;\; y_t = \beta_t \cdot x_t + \varepsilon_t = f(x_t) + \varepsilon_t$; and that there is no function $h : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ such that $B(t) = h(x_t), \forall t : x_t \neq 0$.

Then, the second condition implies that if such a function does not exist, there must be at least two periods t and t' such that

$$\beta_t \neq \beta_{t'} \text{ and } x_t = x_{t'} \neq 0,$$

i.e., there is at least a one *one-to-many* relation between x and β. Then, the first assumption implies that

$$\begin{aligned} \beta_t \cdot x_t &= f(x_t) \\ &= f(x_{t'}) \\ &= \beta_{t'} \cdot x_{t'} \\ &= \beta_{t'} \cdot x_t \\ x_t \neq 0 \implies \beta_t &= \beta_{t'}, \end{aligned}$$

which is a contradiction.

□

Condition 1 (Strictly Monotonic Spending). *Suppose x_t can be expressed as a strictly monotonic function of t, denoted $X(t)$. Then the DGP in (A-7) can be fully expressed as a static nonlinear DGP. In such cases, $X^{-1}(x)$ exists, and each spend level x can have at most a single period t where such spend was observed, which implies that $\beta_t = B(X^{-1}(x_t))$ and*

$$f(x) = B(X^{-1}(x)) \cdot x.$$

Proof. It is easy to verify that this condition satisfies Proposition 1 with $h(x) = B(X^{-1}(x))$. □

Definition 1 (Parent Process). *p_t is called a parent process for x_t and β_t if there exist functions $\mathcal{X}(z)$ and $\mathcal{B}(z)$ such that $x_t = \mathcal{X}(p_t)$ and $\beta_t = \mathcal{B}(p_t)$.*

Condition 2 (Invertible Parent Relationship). *Assume there exists a parent process for x_t and β_t, p_t. If $\mathcal{X}(p_t)$ is invertible, then the DGP in (A-7) can be fully expressed as a static nonlinear DGP, such that $\beta_t = \mathcal{B}(\mathcal{X}^{-1}(x_t))$ and $f(x) = \mathcal{B}(\mathcal{X}^{-1}(x)) \cdot x$.*

Proof. Just replace $h(x) = B(X^{-1}(x))$ to satisfy the condition in Proposition 1. □

B Setting Hyperparameters for GP Models

In this appendix, we describe how we set hyperparameters for GP models for simulations and empirical applications.

B.1 Simulations

To illustrate a general point of the conflation between nonlinear and time-varying effects, we employ straightforward GP models with default priors. In particular, our nonlinear model is a direct implementation of the standard GP regression in Stan (Stan Development Team 2024),[1] which itself follows Williams and Rasmussen (2006). The model standardizes the data and utilizes a standard normal prior for all unconstrained and a half-normal prior for all positive coefficients.

For the time-varying model, we again implemented the model in Stan, using data standardization similar to the nonlinear GP and a weakly informative prior for the lengthscale of the GP. Specifically, following the guidance from Stan Development Team (2024), we use an Inverse Gamma distribution for its zero-avoiding property, with parameters $\text{InvGamma}(2, 30)$. The 95% interval of this distribution is from 5 to 124, which ensures that a reasonable range of lengthscales for our number of periods, 100, are likely. The amplitude and the noise standard deviation are set to have half-normal priors, consistent with ranges used in our simulation DGP.

B.2 Classic Advertising Datasets

For the classic advertising datasets — DWC and seasonally-adjusted Lydia Pinkham — we used very similar priors to the simulations. Specifically, for the nonlinear model, the priors were unchanged. For time-varying model, weakly informative normal and half-normal priors were used for all parameters except the lengthscale, which was given an $\text{InvGamma}(5, 200)$ prior. Following the reasoning above, this prior avoids very low values, and puts probability mass on smooth functions, with lengthscales between roughly 20 and 120. Note that, while this prior induces significant smoothing, the results are also robust to using less informative priors.

B.3 Modern MMM Data

For the multichannel Nielsen data, we construct a more sophisticated set of models that takes into account trend and seasonality, national holidays, and promotions. To handle varying ranges of the spending and sales across channels and brands in our sample, all data is standardized within the models.

Trend and Seasonality Both models include a time-varying intercept, that includes trend and seasonality components. We capture the trend component with a GP with a half-normal amplitude prior, and a normal lengthscale prior with parameters $\mathcal{N}(T, \frac{T}{8})$, where T is the total number of time

[1] https://avehtari.github.io/casestudies/Motorcycle/motorcycle_gpcourse.html

periods. This ensures that the trend is sufficiently smooth, and thus will not overfit the data. The seasonality component is captured by a GP with a periodic kernel, where the period is set to 52 weeks (yearly periodicity), the amplitude prior is again half-normal, and the lengthscale prior is $\mathcal{N}(0.5\pi, 0.25\pi)$. This prior has been used previously by Dew et al. (2024) as a weakly informative prior over the lengthscale of the periodic kernel.

National Holidays We include dummy variables for 11 weeks of national holidays. The coefficients are assumed to be drawn from a normal distribution $Normal(\mu_h, \sigma_h)$, where μ_h has a standard normal and σ_h has a half-normal priors.

Promotions The coefficient on promotions has a standard normal prior.

Main Nonlinear GP The main multivariate GP in the nonlinear model consists of additive, one-per-channel components, each having its own amplitude and lengthscale. Since the spending in all channels, as well as sales, is standardized, we chose a half-normal prior for the amplitudes and an inverse gamma prior for the lengthscales. The specific inverse gamma prior we choose is $\text{InvGamma}(2, 3)$, which has 95% of its mass between 0.5 and 12, resulting in a weakly informative prior for relatively smooth functions over the range of possible values of standardized spending.

Main Time-varying GP The main multivariate GP in the time-varying model consists of one-per-channel components over time that are multiplied by their respective channel spendings and added. Each component has its own amplitude and lengthscale. Similarly to the nonlinear model, we chose a half-normal prior for the amplitudes. For the lengthscales, we chose an $\text{InvGamma}(2, 30)$ prior, with 95% interval in between 5 and 124. Given the range of this GPs input (time), which spans 1-104, this results in a weakly informative prior resulting in relatively smooth functions over this interval.

C Other Methods for Nonlinear and Time-varying Effects

While our focus in the main body of the paper is on using GPs to capture nonlinear and time-varying effects, due to their ease for simulating relatively smooth functions, and their broad applicability across the model classes we are interested in, the conflation issue is not specific to GPs. Thus, in this section, we illustrate two alternative methods for capturing nonlinear and time-varying effects — B-splines and random walks — and show that the conflation issue still emerges. We first briefly describe these alternatives, then show how they can be applied in our focal context, using the introductory example from the paper as an illustration.

C.1 B-splines

To begin, we focus on a common alternative to GPs for estimating unknown functions: splines. Specifically, we use B-splines, implemented, as with our focal specification, in a Bayesian fashion using Stan (Lang and Brezger 2004; Stan Development Team 2024). In a spline specification, an unknown function $f(x)$ is modeled as a linear combination of nonlinear basis functions:

$$f(x) = \sum_{j=1}^{J} a_j B_j(x). \tag{A-9}$$

For B-splines, these basis functions take the form of piecewise polynomials of a given order k, defined over a series of knots at locations $\ell_1, \ldots, \ell_M$, where M is selected a priori. For the sake of concision, we do not exhaustively review B-splines here. We refer interested readers to Lang and Brezger (2004). B-splines can not only be used to flexibly approximate functions of interest, but can also be adapted to have special properties, like monotonicity, which may be useful in the context of MMM but are difficult to enforce in GPs (Brezger and Steiner 2008). Similar specifications have been used in marketing (e.g., Kim et al. 2007; Boughanmi and Ansari 2021; Haschka 2023).

C.2 Random Walk

For time-varying models, dynamic linear and state-space specifications have enjoyed tremendous popularity in marketing, as reviewed in the main body of the paper. The simplest such specification is a random walk, wherein a time-varying parameter β_t is modeled as:

$$\beta_t \sim \mathcal{N}(\beta_{t-1}, \sigma_\beta^2). \tag{A-10}$$

Such a specification has been used by, e.g., Winer (1979), and is a simple version of the model used in many other papers, including, e.g., Van Heerde et al. (2004). In this random walk specification, σ_β plays the role of a regularizer, controlling how much the parameter is allowed to vary, period-by-period. To employ this model within our MMM framework to estimate time-varying coefficients, we again use a Bayesian implementation in Stan, setting the prior of σ_β to be $\text{HalfNormal}(0, 1)$.

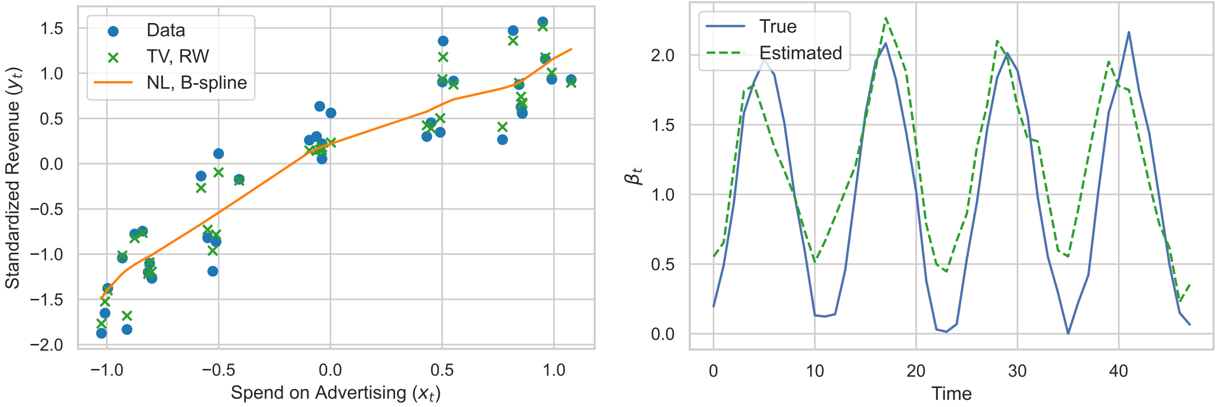

Figure A-1: Conflation Using B-splines and Random Walk
At left, the fit of the nonlinear B-splines model and the time-varying random walk model on the data from Figure 1 in the main body of the paper, illustrating the conflation. At right, the recovery of β_t by the time-varying random walk model.

C.3 Illustration

To illustrate these methods, and demonstrate the persisting conflation, we return to the introductory example, plotted in Figure 1 of the main body of the paper. The two models we implement are (1) a nonlinear model with:

$$y_t = f(x_t) + \epsilon_t,$$

where $f(x_t)$ is estimated using B-splines with order $k = 3$ (i.e., cubic) and 10 knots, and (2) a time-varying coefficients model with,

$$y_t = \alpha + \beta_t x_t + \varepsilon_t,$$

where β_t follows a random walk with the prior noted above. We estimated these two models on the introductory example data with 10 holdout observations. We plot the results in Figure Figure A-1, showing, at left, the fit of the two models, and at right, the recovered β_t from the time-varying coefficients model. From the results, we can see that, again, the two models fit very well, with the true model (the time-varying one) also able to correctly recover the data generating β_t. To evaluate the conflation issue, we examine the RMSE on the holdout observations: for the nonlinear model, we find a posterior mean RMSE of 0.52, with a 95% credible interval of [0.42 0.69]; for the time-varying model, we find a nearly identical posterior mean RMSE of 0.47, with a slightly wider 95% credible interval of [0.28 0.73]. Thus, by standard model selection metrics, a typical analyst would not be able to distinguish these two models. This is as we predicted: the conflation issue arises because of the model form, not the specific machinery used to estimate the nonlinear and time-varying effects. As long as the model is sufficiently flexible, conflation will exist.

References

Boughanmi, K. and Ansari, A. (2021). Dynamics of Musical Success: A Machine Learning Approach for Multimedia Data Fusion. *Journal of Marketing Research*, 58(6):1034–1057.

Brezger, A. and Steiner, W. J. (2008). Monotonic Regression Based on Bayesian P-splines: An Application to Estimating Price Response Functions from Store-Level Scanner Data. *Journal of Business & Economic Statistics*, 26(1):90–104.

Dew, R., Ascarza, E., Netzer, O., and Sicherman, N. (2024). Detecting Routines: Applications to Ridesharing Customer Relationship Management. *Journal of Marketing Research*, 61(2):368–392.

Haschka, R. E. (2023). Endogeneity-Robust Estimation of Nonlinear Regression Models Using Copulas: A Bayesian Approach with an Application to Demand Modelling. Available at SSRN: https://ssrn.com/abstract=4451591.

Kim, J. G., Menzefricke, U., and Feinberg, F. M. (2007). Capturing Flexible Heterogeneous Utility Curves: A Bayesian Spline Approach. *Management Science*, 53(2):340–354.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.

Stan Development Team (2024). Stan Modeling Language Users Guide and Reference Manual. https://mc-stan.org.

Van Heerde, H. J., Mela, C. F., and Manchanda, P. (2004). The Dynamic Effect of Innovation on Market Structure. *Journal of Marketing Research*, 41(2):166–183.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA.

Winer, R. S. (1979). An Analysis of the Time-Varying Effects of Advertising: The Case of Lydia Pinkham. *Journal of Business*, pages 563–576.