1.Model Specification:

$$Y_t = g_t + S_t + H_t + F(X_t) + \varepsilon$$

In which Y_t is sales.

1.1 Growth trend g_t:

Mainly two types of growth trend specification:

1. saturating trend function with change points
2. Linear trend function with change points

Trend function is a function of time, growth rate, change points and capacity. Change points are those time points where growth rate changes. See [1] for detailed information.

How to choose these change dates? Of course you can ask some experts about the dates where there is an obvious change of growth rate. You can also use feature selection. For example, for each month we specify a growth rate change (delta) and add regularization terms of these deltas when doing MLE(as will be explained in 2.1). Of course, the regularization term makes some delta close to zero or equal to zero. We consider there is no change of growth rate for those months with delta = 0. Through this process we automatically select the growth rate changing points.

1.2 Seasonality S_t:

Use Fourier Decomposition. See [1] for detailed information.

1.3 Holiday effect H_t: See [1] for detailed information.

1.4 Media Spend Information $F(X_t)$:

1.4.1. Adstock: lag effects of media spends.

Media spend at time t, x_t, may have influence on $Y_t, Y_{t+1}, Y_{t+2} \ldots$. We use add stock function to capture such lag effects. See [2] for detailed information.

1.4.2. Concavity

hills function or sigmoid function are often used to capture the decreasing marginal effect of media spend. See [2] for detailed information.

1.4.3. How to combine lab and concavity ?

If the media spend at each period is small compared to the cumulative spend, then put the adstock function of media spend into the expression of hills function. If you media spend is largely concentrated over several short time intervals with an on-and-off pattern, then put the hills function of media spend into the expression of adstock. See [2] for detailed information.

2. Estimation method

2.1 Maximization Likelihood

The estimation of the parameters in this model is simply maximum likelihood. Remember in the classical regression model, we have

$$Y = X\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2),$$

we assume the error term follows normal distribution. Then

$$Y \sim N(X\beta, \sigma^2),$$

Then we can choose the beta to maximize

$$\sum_i logf(X_i; \beta)$$

In the additive model, we have

$$Y_t = G(A_t,\ \theta) + X_t\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$$

Which implies that $Y \sim N(G(A_t,\ \theta) + X_t\beta, \sigma^2)$, In which theta and beta are parameters, while A_t, X_t and Y_t are data. Then we can choose the parameters to maximize

$$\sum_t logf(X_t, A_t; \beta, \theta)$$

In order to prevent overfitting, we can also add regularization term. In general, like ridge or lasso, we do not want the abstract value of coefficients to be too large. We can therefore specify a function R() such that R() reaches maximum when input is 0 but starts to decline whenever the input deviates from 0. Therefore, we want to choose beta and theta to maximize the following.

$$\sum_t logf(X_t, A_t; \beta, \theta) + R(\beta) + R(\theta)$$

In practice, we can assume that both beta and theta follows some sort of normal distribution with mean 0. Therefore, their density function must be a function with 0 as the center axis. Deviation from 0 lead to lower value of density function, which plays the same role as R. This is exactly what the STAN does.

In general, the ‘distribution parameters’ of parameters are actually hyperparameters, which you can fine tune to improve the model performance.

2.2 Bayesian Estimation

$$p(\theta,\ \beta|\ X,y) = \frac{p(y,x\ |\theta,\beta)p(\theta)p(\beta)}{p(X,y)} \propto p(y,x|\theta,\beta)p(\theta)p(\beta)$$

Which equals

$$p(y,x|\theta,\beta) = \prod_j^n p(y_j,x_j|\theta,\beta) = \prod_j^n p(y_j|\theta,\beta,x_j)p(x_j)$$

Therefore

$$p(\theta,\ \beta|\ X,y) \propto p(\theta)p(\beta)\prod_j^n p(y_j|\theta,\beta,x_j) \quad (*)$$

Since we know that $Y \sim N(G(A_t,\ \theta) + X_t\beta, \sigma^2)$, we can easily calculate $p(y_j|\theta,\beta,x_j)$ .In the maximum likelihood method, we set the prior $p(\theta)p(\beta)$ and maximize (*). We finally get a point estimation of θ,β . ($p(\theta)p(\beta)$ actually acts like regularization term)

But in the bayesian estimation, we do not maximize (*). Instead, we directly set $p(\theta)\ and\ p(\beta)$ and get the posterior distribution of parameters $p(\theta,\ \beta|\ X,y)$

Often times, $p(\theta,\ \beta|\ X,y)$ has an analytical form, but is super hard to get its statistical property without sampling, and sampling is very difficult. We use mcmc sampling method to sample $\theta,\ \beta$, and using these sample, we can draw the marginal distribution of $\theta,\ \beta$

Note : If the independent variables are highly correlated , then posterior coefficients may be highly correlated, and this may lead to a much longer time to converge. Try to regress them and get residual to get orthogonal components of these variables. !

3. Implication of Model

3.1.Forecast and Accuracy evaluation

How to evaluate the forecast accuracy ?The following link introduces the basic idea. Very intuitive and simple:
https://cran.r-project.org/web/packages/greybox/vignettes/ro.html#:~:text=Rolling%20origin%20is%20an%20evaluation,of%20how%20the%20models%20perform.

Rolling origin method works as follows:

1. With constant hold-out samples
2. With non-constant hold-out samples
3. With constant training samples.

We need to train the model (keeping hyperparameter fixed) at each iteration. To be specific, do the following procedure.

Choose some time points t1,t2,t3,.....T
For t in {t1,t2,,,,,T}:
    Estimate model using data from time 0 to time t. Get m(t)
End for
For h in {h1,h2,....H}:
    For t in {t1,t2….T}

Forecast, get $\hat{y}(t+h \mid m(t))$

End for

Take average over $\hat{y}(t_1+h \mid m(t_1)), \hat{y}(t_2+h \mid m(t_2)) \ldots$ *denote as* $\xi(h)$

End for

Through this process we can get $\xi(h_1), \xi(h_2) \ldots$ (of course, for different h, you can also specify different set of time points t1,t2….T)

But here is a problem. If the number of time points is too large. (extreme case is that every calendar day is a time point), then for given h, those predictions, $\hat{y}(t_1+h \mid m(t_1)), \hat{y}(t_2+h \mid m(t_2)) \ldots$ may be highly correlated. In some sense, this cannot provide good information on the model's accuracy. A rule of thumb is that, if you want to estimate $\xi(h)$, you'd better choose $h/2$ time points

3.2. Check ROA of each media

After the estimation, you want to check the marginal or average effect (ROA) of a media over a given range of time.Doing this is not difficult. But since our estimation of coefficients are just posterior distribution.(recall that we sample a lot of parameters Φ using Gibbs sampling) Therefore in general we can:For each parameter sample taken out from the posterior distribution, Φ_j, We calculate the ROA_j that is calculated using Φ_j. We can then plot the distribution of ROA.

3.3. Optimal Allocation of Media Spend

Sometimes we also want to know for a given interval, how to maximize the sum of y series. How to do this? The objective function is apparently related to the parameter, but what we have is a posterior distribution of parameters. Intuitively there are two methods.

Method 1: you can do the following

$$\max_x \frac{1}{N} \sum_j f(x, \Phi_j) \; s.t \; G(x) <= 0$$

From this you get ONE allocation result. x^*

Method 2: you can also do: For each Φ_j, do

$$\max_x f(x, \Phi_j) \; s.t \; G(x) <= 0$$

And get x^*_j

So you get many optimal allocation decision x^*_j, so you can plot its distribution. It can gives many useful information: the variance of the optimal allocation, and whether it is trustworthy

4.Model Selection

4.1 Using BIC

$$BIC = -2\log(pr(y|x,\theta,\beta)) + k\log(n)$$

First term contains the maximum likelihood term, while the k is the number of free parameters and n is the sample size. In maximum likelihood method you can directly put the estimation of the parameter into the expression.In Bayesian method, since you are faced with a posterior distribution of parameters, the log likelihood function above is the average over different parameter samples.

4.2 Variance of posterior distribution of parameters.

Do some of them have large variance? Large variance is undesirable.

4.3 Check the data information in altering the prior belief

We know that in Bayesian estimation, we get the posterior distribution by multiplying the likelihood function given the parameters and the prior distribution of the parameters. Sometimes we want to check for certain parameters, its posterior distribution and prior distribution, and how they are different. If the posterior distribution is very similar to the prior distribution for this parameter, then it indicates that the data does not have much information regarding the parameter. So, we should switch to other models where we do not have to estimate this parameter.(? Or we should increase the data sample size)

4.4 Check the auto-correlation of residuals.

For each mcmc sampling:

you get a parameter sample.

Put the parameter into the model and you get the predicted y, and hence residual, which represents the part that cannot be explained by the model. For this residual series, you can get the auto-regression coefficient (ACF).

For each mcmc sampling , you can get an ACF series (up to some order).

End for

Calculating the average of ACF across these mcmc samples. If residual correlation seems to be strong for many orders, then the model may be misspecified.

Reference:

[1] https://research.fb.com/wp-content/uploads/2017/11/forecastingatscale.pdf
[2] https://research.google/pubs/pub46001/