Challenges, Opportunities, and Innovations in Media Mix Modelling

Not for external distribution.

charles.shaw@msixandpartners.com
M/Six Data Union

15 Dec 2022

Section 1

Motivation / Preliminaries

Predictive Modelling

Data Science (. . . / Artificial Intelligence / Machine Learning / . . .) := a set of methodologies to automatize the process of

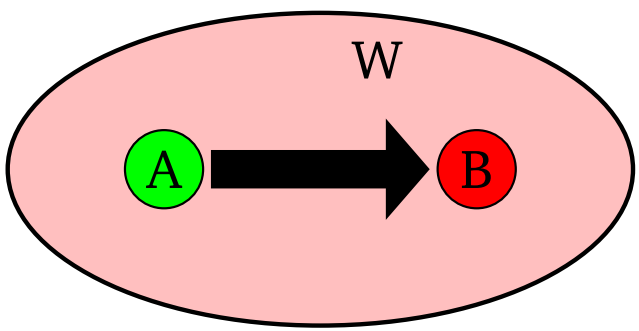
observation $\rightarrow$ modelling $\rightarrow$ predicting $\rightarrow$ testing

This means to discover laws automatically. This can be done instanteneously or temporaneously.

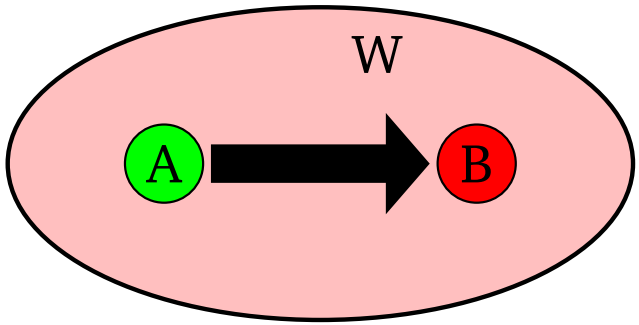
The relationship between set of variables A and set of variables B: a (probabilistic) law.

Goal:

To find automatically **maps** between sets of variables.



Predictive Modelling



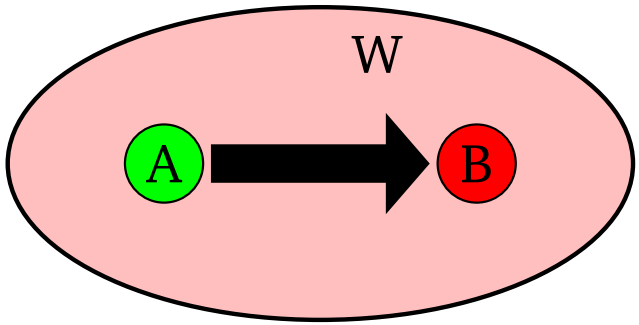
If the law is continuous := an equation (or a neural network).

If the law is discrete := a classification problem.

But it is always a mapping, which is unavoidably probabilistic.

In complex systems these **maps** are conditional probabilities. This is the probability of observing B, when you have A, within the context of whatever happens in the world.

Predictive Modelling



$$P(X_B \mid X_A, X_W)$$

Prediction is the estimation of the conditional probability of a (future) event, given the available information about other (past) events.

Predictive Modelling

The conditional probability

$$p(X_B \mid X_A, X_W)$$

is a tool for:

1. forecast ("what is the expectation value of X_B"?)
2. test hypothesis ("what happens if X_A changes"?)
3. quantify risk
4. analyse scenarios
5. stress test

Marketers are primarily interested in 1 and 2. Other business domains (eg finance) may also be interested in 3-5.

Predictive Modelling

What we commonly call a prediction is the **expectation value**

$$\hat{X}_B = E[X_B \mid X_A^-, X_W^-] = \sum_{X_B} X_B \, p(X_B \mid X_A^-, X_W^-)$$

this is the **regression** and for a linear model (multivariate Gaussian) this is the linear regression formula which we know how to solve analytically given a set of training data. We also want to know the uncertainty of $\hat{X}_B$ (the standard deviation). The uncertainty, in a general (nonlinear) framework is the entropy.

$$H(X_B \mid X_A^-, X_W^-) = -\sum_{X_A, X_B} p(X_B \mid X_A^-, X_W^-) \log p(X_B \mid X_A^-, X_W^-)$$

Entropy is the amount of uncertainty of variable B, which is left when we "fix" A. We may have a lot of uncertainty if we do not know much about B given the knowledge of the past. If we know more about what is happening with B than before then we get **uncertainty reduction**.

Causality

The **reduction of uncertainty** on variables X_B given the knowledge of the past (variables X_A^-, X_W^-), discounting for the past X_B^-.

$$H(X_B \mid X_B^-, X_W^-) - H(X_B \mid X_A^-, X_B^-, X_W^-) = TE(X_A \to X_B \mid X_W^-)$$

this is the **transfer entropy**[1] that for linear models (multivariate Gaussian) coincides with **Granger causality**[2].
TE is the uncertainty left on B:

if you know the past of B (and the past of the rest of the world); or

if you know the past of B and the past of A

Intuitive explanation:

$H(X_B \mid X_A^-, X_B^-, X_W^-)$ should be smaller than $H(X_B \mid X_B^-, X_W^-)$ because there is more information (you know also the past of A).

How much does the past of A provide about B? If past of A provides full information about B then $H(X_B \mid X_A^-, X_B^-, X_W^-)$ is zero and we have maximum TE.

If past of A provides no information about B then $H(X_B \mid X_A^-, X_B^-, X_W^-)$ would be equal to $H(X_B \mid X_B^-, X_W^-)$ and there would be no TE i.e. non-causal.

[1] Transfer entropy is a way to measure causality.

[2] aka Wiener-Granger causality. Wiener, N., 1956. The theory of prediction. In: Beckenbach, E. (Ed.), Modern Mathematics for Engineers. McGraw-Hill, New York

High-dimensional problem

To estimate the probability we must estimate a number of quantities between

$$\frac{N^2}{2} \leq \text{quantities to estimate} \leq N!$$

High-dimensional problem:

Information increases linearly with observations. But model parameters increase at least with the square.

Models with 20-30 independent variables are fairly standard.[3] This is already a large number. One approach is to use dimensionality reduction (shrinkage) methods eg Tikhonov regularization.
Data in marketing tends to be increasingly high-dimensional, which means that the models are hard to parametrise. Data is also inter-related, which poses additional challenges.

[3] To compare: 52! is the number of stars in the known universe, and 80! is the number of atoms in the known universe.

Collinearity

Recall the formula for the estimated coefficients in a multiple linear regression:

$$\hat{\beta} = (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{y}$$

This is obviously going to lead to problems if $\boldsymbol{x}^\top \boldsymbol{x}$ isn't invertible. Similarly, the variance of the estimates,

$$Var(\hat{\beta}) = \sigma^2 (\boldsymbol{x}^\top \boldsymbol{x})^{-1}$$

will blow up when $\boldsymbol{x}^\top \boldsymbol{x}$ is singular. If that matrix isn't exactly singular, but is close to being non-invertible, the variances will become huge.
There are several equivalent conditions for any square matrix, say $\boldsymbol{u}$, to be singular or non-invertible:(i) The determinant $\det \boldsymbol{u}$ or $|\boldsymbol{u}|$ is 0. (ii) At least one eigenvalue of u is 0. (This is because the determinant of a matrix is the product of its eigenvalues.) (iii) $\boldsymbol{u}$ is **rank deficient**, meaning that one or more of its columns (or rows) is equal to a linear combination of the other rows. Since we're not concerned with any old square matrix, but specifically with $\boldsymbol{x}^\top \boldsymbol{x}$, we have an additional equivalent condition:

- $\boldsymbol{x}$ is **column-rank** deficient, meaning one or more of its columns is equal to a linear combination of the others.

The last explains why we call this problem **collinearity**: it looks like we have p different predictor variables, but really some of them are linear combinations of the others, so they don't add any information. The real number of distinct variables is $q < p$, the column rank of $\boldsymbol{x}$. If the exact linear relationship holds among more than two variables, we talk about **multicollinearity**; **collinearity** can refer either to the general situation of a linear dependence among the predictors, or, by contrast to multicollinearity, a linear relationship among just two of the predictors.
Again, if there isn't an *exact* linear relationship among the predictors, but they're close to one, $\boldsymbol{x}^\top \boldsymbol{x}$ will be invertible, but $(\boldsymbol{x}^\top \boldsymbol{x})^{-1}$ will be huge, and the variances of the estimated coefficients will be enormous. This can make it very hard to say anything at all precise about the coefficients.

Multicollinearity

A multicollinear relationship involving three or more variables might be totally invisible on a pairs plot. For instance, suppose X_1 and X_2 are independent Gaussians, of equal variance σ^2, and X_3 is their average, $X_3 = (X_1 + X_2)/2$. The correlation between X_1 and X_3 is

$$Cor(X_1, X_3) = \frac{Cov(X_1, X_3)}{\sqrt{Var(X_1)Var(X_3)}} = \frac{Cov(X_1, (X_1 + X_2)/2)}{\sqrt{\sigma^2\sigma^2/2}} \quad (1)$$

$$= \frac{\sigma^2/2}{\sigma^2/\sqrt{2}} = \frac{1}{\sqrt{2}} \quad (2)$$

This is also the correlation between X_2 and X_3. A correlation of $1/\sqrt{2}$ (=71%) isn't trivial, but is hardly perfect, and doesn't really distinguish itself on a pairs plot![4]
If the predictors are correlated with each other, the standard errors of the coefficient estimates will be bigger than if the predictors were uncorrelated. If the predictors were uncorrelated, the variance of $\hat{\beta}_i$ would be

$$Var(\hat{\beta}_i) = \frac{\sigma^2}{ns_{X_i}^2} \quad (3)$$

just as it is in a simple linear regression. With correlated predictors, however, we have to use our general formula for the least squares:

$$Var(\hat{\beta}_i) = \sigma^2(\boldsymbol{x}^\top \boldsymbol{x})^{-1}_{i+1,i+1} \quad (4)$$

The ratio between Eqs. 4 and 3 is the **variance inflation factor** for the i^{th} coefficient, VIF_i. Folklore says that $VIF_i > 10$ indicates "serious" multicollinearity for the predictor.

[4] Consider also y=sin(x). In absence of confounders: no correlation, perfect causation.

Illustration: Problems with OLS

Consider a matrix $\mathbf{X}$ with $n = 20$ and whose elements consist of independent, normally distributed random numbers; the figure below plots the largest variance of the $\hat{\beta}_j$ estimates as we increase the number of columns in $\mathbf{X}$:

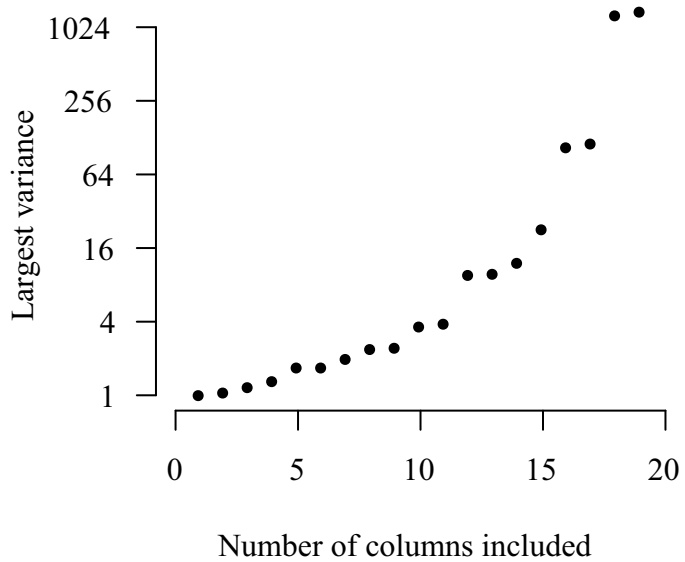


Figure: Problems with OLS

As $p \to n$, $\mathbb{V}(\hat{\boldsymbol{\beta}})$ increases without bound; the increase is substantial as p approaches n, and infinite when $p \geq n$. Maximum likelihood cannot handle high-dimensional data without causing major identifiability and efficiency issues.

Application to media / advertising

- Advertisers have a need to understand the effectiveness of their media spend in driving sales in order to optimize budget allocations. Media mix models (MMMs) are a common and widely used approach.
- MMMs are statistical models used by advertisers to measure the effectiveness of their advertising spend and have been around in various forms since the 1960s [B64, M78]. MMMs use aggregate historical time series data to model sales outcomes over time, as a function of advertising variables, other marketing variables, and control variables like weather, seasonality, and market competition. Metrics such as return on advertising spend (ROAS) and optimized advertising budget allocations are derived from these models, based on the assumption that these models provide valid causal results.
- Aim of this talk is to outline the various challenges such models encounter in consistently providing valid answers to the advertiser's questions on media effectiveness. I will also discuss opportunities for improvements in media mix models that can produce better inference. I will then present examples of Msix innovations that are turning problems into progress.

What is an MMM?

- MMMs seek to provide advertisers with answers to causal questions. For instance: 1) What was my ROAS on television in the past year? 2) What would my sales be if next year's expenditures were increased or decreased? 3) How should I manage my media expenditure to enhance sales?
- MMMs are often regression models based on a limited quantity of aggregated observational data, and such models produce correlational rather than causal conclusions. Only under very specific situations can these estimations be called causal.
- The work of the Data Science team aims to address the obstacles that MMMs have when attempting to provide reliable responses to these types of inquiries, as well as potential chances to enhance their capacity to deliver valid inference.

Things to think about

- variable selection i.e. the selection of independent variables in a regression model.
- overfitting
- out of sample performance

Possible approaches

1. "Kitchen sink" OLS: include all regressors
2. Stepwise OLS: begin with general model and drop if p-value > 0.05
3. Stochastic search heuristics / genetic algorithms: GAs where the optimality is determined with respect to some criteria (eg AIC) - which is the strataQED methodology
4. Regularized regression

What is regularized regression?

Regularized regression (eg LASSO):

- Lasso is a special kind of penalised regression. Not only does it penalise overfitting but it coerces some betas to have a value of exactly zero with sufficiently large tuning parameter.
- It is used in many areas as a one step model selection and model fitting method. Many other penalties (such as ridge) do not have this property.
- The most attractive feature is high-dimensional controls, even if the number of controls exceeds the number of observations.
- In short, a penalized regression where overfitting is penalized, basically a trade off between bias and variance.

Regularized regression: Prediction vs inference

- Statistical learning
  - Focus on prediction and classification.
  - Wide set of methods: support vector machines, random forests, neural networks, penalized regression, etc.
  - Typical problems: predict user-rating of films (Netflix), predicting success of telemarketing, etc
- Econometrics and allied fields
  - Focus on "causal" inference using OLS, IV/GMM, Maximum Likelihood (ML) / Restricted ML / EM
  - Typical question: Does x have a "causal" effect on y? But difficult to do causal analysis without a theoretic model or instruments.
- Statistical learning can augment more "traditional" methods. Benefit: Out-of-sample prediction, high-dimensional data, data-driven model selection.

Why can't we just do stepwise regression?

Stepwise methods will not necessarily produce the best model if there are redundant predictors (common problem). All-possible-subset methods produce the best model for each possible number of terms, but larger models need not necessarily be subsets of smaller ones, causing serious conceptual problems about the underlying logic of the investigation. Models identified by stepwise methods have an inflated risk of capitalising on chance features of the data. They frequently fail when applied to new datasets. In summary, here are some of the problems with stepwise variable selection.

- It yields R-squared values that are badly biased (high).
- The F and chi-squared tests quoted next to each variable on the printout do not have the claimed distribution.
- The method yields confidence intervals for effects and predicted values that are falsely narrow.
- It yields P-values that do not have the proper meaning and the proper correction for them is a very difficult problem.
- It gives biased regression coefficients that need shrinkage (the coefficients for remaining variables are too large; see Tibshirani [T96]).
- It has severe problems in the presence of collinearity.
- It is based on methods (e.g., F tests for nested models) that were intended to be used to test prespecified hypotheses.

Model selection

- The standard linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i$$

  Why would we use a fitting procedure other than OLS?

- Model selection.
  - We do not know the true model. Which regressors are important?
  - Including too many regressors leads to overfitting: good in-sample fit (high R^2), but bad *out-of-sample performance*.
  - Including too few regressors leads to *omitted variable bias*.
- Model selection becomes even more challenging when the data is high-dimensional.
  - If $p > n$, the model is not identified.
  - If $p = n$, perfect fit. Meaningless.
  - If $p < n$ but large, overfitting is likely: Some of the predictors are only significant by chance (false positives), but perform poorly on new (unseen) data.
- Large p is often not acknowledged in applied work:
  - The true model is unknown ex ante. Unless an analyst runs one and only one specification, the low-dimensional model paradigm is likely to fail.
  - The number of regressors increases if we account for non-linearity, interaction effects, parameter heterogeneity, spatial and temporal effects.
  - Especially if p is large, inference is problematic. Need for false discovery control (multiple testing procedures) - rarely done.

Prediction

- Bias-variance-tradeoff. OLS estimator has zero bias, but not necessarily the best out-of-sample predictive accuracy.
- High-dimensional data. The general model is:

$$y_i = x_i^\top \boldsymbol{\beta} + \epsilon_i$$

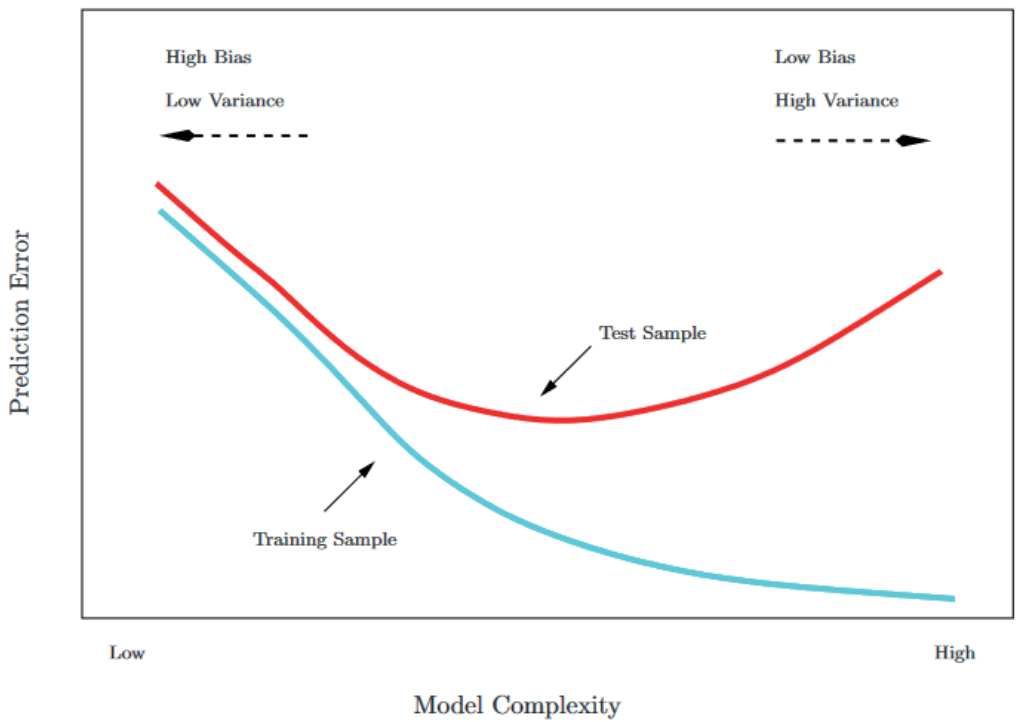
- We index observations by i and regressors by j. We have up to $p = dim(\beta)$ potential regressors. p can be very large,
- OLS leads to disaster: If p is large, we overfit badly and classical hypothesis testing leads to many false positives.
- This becomes manageable if we assume *(exact) sparsity*: of the p potential regressors, only s regressors belong in the model, where

$$s := \sum_{j=1}^{p} \mathbb{1}\beta_j \neq 0 \ll n$$

- In other words: most of the true coefficients β_j are actually zero. But we do not know which ones are zeros and which ones are not.
- We can also use the weaker assumption of approximate sparsity: some of the β_j coefficients are well-approximated by zero, and the approximation error is sufficiently small.

Overfitting

- A full model with all predictors (kitchen sink approach) will have the lowest bias (OLS is unbiased) and R^2 (in-sample fit) is maximised.
- However, the kitchen sink model likely suffers from overfitting.
- Removing some predictors from the model (i.e., forcing some coefficients to be zero) induces bias. On the other side, by removing predictors we also reduce model complexity and variance.
- The optimal prediction model rarely includes all predictors and typically has a non-zero bias.
- High R^2 does not translate into good out-of-sample performance.



Source: Tibshirani/Hastie

Illustration of regularization regression: LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator, Tibshirani, 1996 [T96]) has found its uses in MMM. Below is a quote from Deloitte and Facebook's joint MMM guide:

> *For the highly correlated data in a synchronised marketing campaign, even if there is no magic and definitive solutions around this topic, regularization techniques, such as Lasso or Ridge regression, have demonstrated their reliability in order to face multicollinearity among many regressors, preventing overfitting and helping on variable mis-selection. This approach tends to improve the predictive performance of MMMs, providing more flexible models that allow deeper levels of detail and more robust results, with a better balance between analytical and business requirements of the projects.*

Source: Deloitte, "The future is modeled A how-to guide for Advanced Marketing Mix Models" [D]

LASSO, "ℓ^1"- norm

Minimize:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p} \mid \beta_j \mid$$

There is a cost to including lots of regressors, and we can reduce the objective function by throwing out the ones that contribute little to the fit. The effect of the penalization is that LASSO sets the $\hat{\beta}_j$ for some variables to zero. In other words, it does the model selection for us. In contrast to ℓ^1-norm penalization (AIC, BIC) more computationally feasible.

LASSO: Choice of the penalty level

The penalization approach allows us to simplify the model selection problem to a one-dimensional problem. But how do we select λ? Three approaches:

- Data-driven: re-sample the data and find the λ that optimizes out-of-sample prediction. This approach is referred to as cross-validation.
- "Rigorous" penalization: Belloni et al. (2012, Econometrica [B+12]) develop theory and feasible algorithms for the optimal λ under heteroskedastic and non-Gaussian errors. Feasible algorithms are available for LASSO and square-root LASSO.
- Information criteria: select the value of λ that minimizes information criterion (AIC, AICc, BIC, EBIC_γ, etc)? Not always the right strategy.

LASSO: Theory-driven penalty

While cross-validation is a popular and powerful method for predictive purposes, it lacks theoretical justification [5].
The theory of the "rigorous" LASSO has two main ingredients:

- Restricted eigenvalue condition (REC): OLS requires full rank condition, which is too strong in the high-dimensional context. REC is much weaker.
- Penalization level: We need λ to be large enough to "control" the noise in the data. At the same time, we want the penalty to be as small as possible (due to shrinkage bias).

This allows to derive theoretical results for the LASSO: consistent prediction and parameter estimation. The theory of Belloni et al. (2012) allows for non-Gaussian and heteroskedastic errors.

[5] its theoretical validity is an open question in the settings p >> n, but we are not working in those scenarios in MMM

LASSO: Information criteria

Some available information criteria:

$$AIC(\lambda, \alpha) = N \log(\hat{\sigma}(\lambda, \alpha)) + 2df(\lambda, \alpha)$$

$$BIC(\lambda, \alpha) = N \log(\hat{\sigma}(\lambda, \alpha)) + df(\lambda, \alpha) \log(N)$$

$$AICc(\lambda, \alpha) = N \log(\hat{\sigma}(\lambda, \alpha)) + 2df(\lambda, \alpha) \frac{N}{N - df(\lambda, \alpha)}$$

$$EBIC_{\gamma}(\lambda, \alpha) = N \log(\hat{\sigma}(\lambda, \alpha)) + df(\lambda, \alpha) \log(N) + 2\gamma df(\lambda, \alpha) \log(p)$$

df is the degrees of freedom. For the LASSO, df is equal to the number of non-zero coefficients.

- Both AIC and BIC are less suitable in the large-p-small-N setting where they tend to select too many variables.
- AICc addresses the small sample bias of AIC and should be favoured over AIC if n is small.
- The BIC underlies the assumption that each model has the same probability. While this assumption seems reasonable if the researcher has no prior knowledge, it causes the BIC to over-select in the high-dimensional context.
- Extended BIC imposes an additional penalty on the number of parameters. The prior distribution is chosen such that dense models are less likely [CC08].

Alternative estimators that have been inspired by the LASSO

- Elastic net (Zou and Hastie, 2005 [ZH05]). The Elastic Net applies a mix of ℓ^1 (LASSO-type) and ℓ^2 (ridge-type) penalization:

$$\hat{\beta}_{elastic} = \arg\min \frac{1}{N}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \frac{\lambda}{N}\left[\alpha\sum_{j=1}^{p}\psi_j|\beta_j| + (1-\alpha)\alpha\sum_{j=1}^{p}\psi_j\beta_j^2\right]$$

  where $\alpha \in [0,1]$ controls the degree of ℓ^1 (LASSO-type) to ℓ^2 (ridge-type) penalization. $\alpha = 1$ corresponds to the LASSO, and $\alpha = 0$ to ridge regression [6]

- Model selection is a much more difficult problem than prediction. The LASSO is only model selection consistent under the rather strong irrepresentable condition (Meinshausen and Buhlmann, 2006 [MB06]). This shortcoming motivated the **Adaptive LASSO**). The Adaptive LASSO is variable-selection consistent for fixed p under weaker assumptions than the standard LASSO.

[6] (in GLMNet α is set $= 0$ but this can be changed if needed)

Assessing prediction performance - real data!

We randomly split our data into two samples (75%/25%)[7]. One we will fit models on, and the other we will use to test their predictions.

	OLS	LASSO	Elastic Net	Ridge	Adaptive LASSO
in-sample RMSE	3.94E+07	7.82E+07	8.51E+07	3.94E+07	8.51E+07
out-of-sample RMSE	4.08E+10	2.88E+10	3.44E+10	4.08E+10	3.44E+10

- OLS exhibits (joint) lowest in-sample RMSE, but worst out-of-sample prediction performance. **Classic example of overfitting.**
- LASSO exhibits best out-of-sample prediction performance.
- Takeaways:
  - Statistical learning provides wide set of flexible methods focused on prediction and classification problems.
  - Penalized regression outperforms OLS in terms of prediction due to bias-variance-tradeoff.
  - LASSO / Ridge is just one such method, but has some advantages: closely related to OLS, sparsity, well-developed theory, etc.

[7] Analogous results were achieved with 50/50 split

Structure for the rest of this talk

- Next, I motivate MMM analysis by explaining two popular methodologies that may be used to address causal questions and highlight why they are unrealistic or infeasible in the context of answering all queries an advertiser may have regarding the efficacy of their advertising channels. Hence the reason advertisers turn to MMMs.
- I then provide context by explaining some of the regression modelling techniques that are extensively employed in MMMs today. We show a typical model specification that seeks to account for carryover and diminishing returns, as well as the type of data commonly accessible to modellers for fitting such models.
- Following this, I outline the obstacles a modeller may encounter while attempting to obtain accurate estimates using a regression MMM. These obstacles can be roughly categorised into three basic categories: data constraints, selection bias, and modelling.
- I offer some thoughts regarding the necessity to accept uncertainty in the modelling process, the need for transparency between the modeller and the end user of the model outputs, and the need to educate end-users of MMMs of their capabilities and limits.
- I conclude by outlining highlighting some relevant work and products under development.

Ensuring causal, not casual, inference

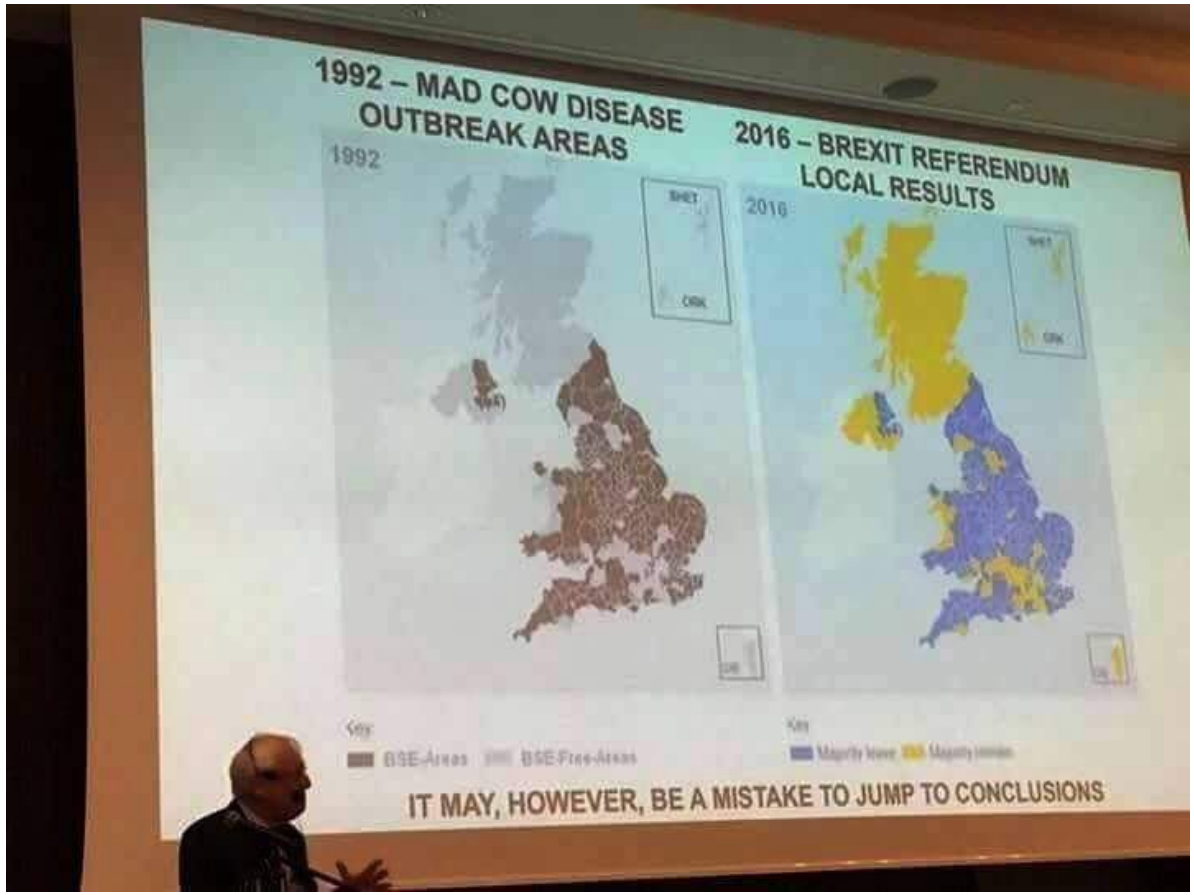


Figure: Correlation ! = causation.

As with other domains, causal inference in marketing requires subtlety. Next section discusses two ways that advertisers might use to answer queries regarding the effectiveness of their advertising. It explains why these approaches are infeasible or impractical for advertisers, which explains why they resort to MMMs instead.

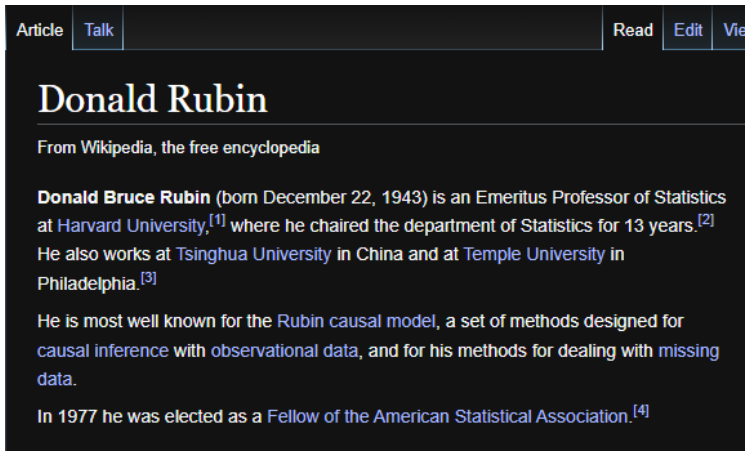
Standard approaches to causal inference

Randomised trials. Generally accepted gold standard for answering causal questions is to perform a randomized experiment.

- ✗ **Problem:** this is not possible for most advertisers, if not all of them. Other things that keep people from using randomised experiments are the technical difficulties of setting them up, the lost opportunities that come with having a control group, the costs of having a test group, and weak advertising effects that may need very large sample sizes

"**Potential Outcomes" approach to causal inference**. If advertisers can not do large-scale randomised experiments and have to rely on historical data instead, they could use the potential outcomes framework, also called the Rubin causal model for causal inference to try to figure out what caused what (see Imbens and Rubin, [IR15]).

- ✗ **Problem:** Requires assumptions which are nearly impossible to meet in MMM



Alternative methods for dealing with selection on unobservables:

- ✗ Instrumental Variables (IV)
- ✗ Regression Discontinuity Design (RDD)

See Appendix for discussion.

Section 2

Building models from data

Regression

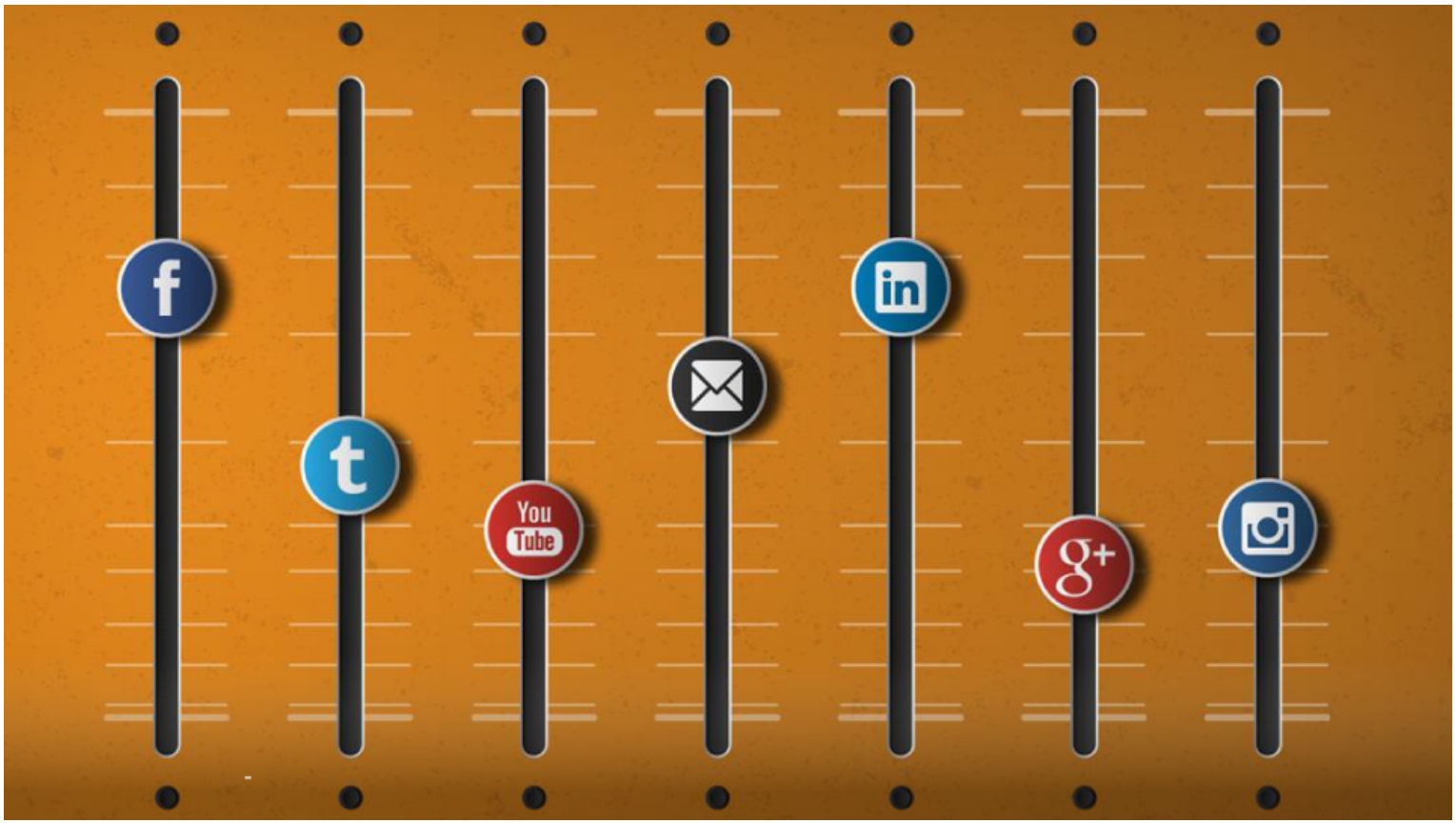
- We briefly discuss the types of data that are typically available in a typical MMM dataset before delving into the specifics of the regression model, noting some difficulties even in the definition and collection of the data.
- Data used to fit MMMs are historical weekly or monthly aggregated national data, although geolevel or even store level data can be used. The data includes:
  - **response data**, which are typically sales but can be other KPIs such as store visits,
  - **media metrics** in the different media channels, such as impressions, clicks, GRPs, with media spend being the most common,
  - **marketing metrics** such as price, promotion, product distribution, and
  - **control factors** such as seasonality, weather and market competition.
- Data collection and data quality are both very hard challenges with MMM. Aside from the logistics of gathering and putting together all the data needed for an MMM, the data itself could be of different quality and level of detail. First, for an MMM to work, the response data and the ad spend data need to be at the same level. So, an advertiser might have very accurate data on sales at the SKU or store level, but most advertising is done at the brand or product level, and usually for a whole country.
- There is not always a clear link between the product for which the MMM is sought and the amount of money spent on advertising. In the MMM, it is also hard to account for the halo effects of advertising for related brands. So, deciding which advertising campaigns are relevant to which SKUs is partly subjective and could lead to mistakes by under- or over-assigning advertising to certain SKUs.

MMM in the real world

- Data used to fit MMMs usually end up at the lowest common denominator of geographical granularity. In other words, the granularity of the entire model is determined by whichever ad channel is accessible at the lowest level. Recent Bayesian hierarchical model research [S+17] demonstrates that it is sometimes possible to combine several types of data into a single model.
- Advertisers normally have an effective data gathering method in place for the response data, which is typically sales. There are more third-party sources of sales, price, and promotion data in the majority of industries where MMMs are extensively utilised. Thus, advertisers may rely on these third parties to additionally supply rival data in addition to the internal source of sales, price, and promotion data for their own brands. However, competition variables for pricing, advertising, and distribution can be hard to get and are frequently left out of MMMs.

MMM data

- The collection of ad exposure data is more difficult because advertising campaigns are frequently planned and carried out by a number of intermediaries, including agencies. To obtain information about the advertising campaigns they have supported, advertisers may need to go via numerous different organisations. This procedure's intricacy makes it quite likely that some crucial data will be overlooked or misinterpreted. Even though the amount of advertising expenditure can be found, the information regarding advertising exposure might be trickier to find and is frequently calculated differently depending on the media and the vendor providing the information.
- In particular, this applies to offline media. For instance, while circulation figures for print publications can be given, they are not always a reliable indicator of the actual number of people who see the advertisement. Another potential cause of data inaccuracy is the labour-intensive collection procedure and the use of proxy variables.


You
Tube

A typical regression model

$$\text{Sales} = \text{pricing} + \text{media} + \text{discounts} + \text{seasonality} + \text{promotion} + \ldots$$

- More formally, a regression MMM specifies a parameterized sales function chosen by the modeler, e.g

$$y_t = F(\boldsymbol{x}_{t-L+1}, \ldots, \boldsymbol{x}_t, \boldsymbol{z}_{t-L+1}, \ldots, \boldsymbol{z}_t; \Phi)$$

- where y_t is the sales at time t, $F(\ldots)$ is the regression function, $\boldsymbol{x}_t = \{x_{t,m}, m = 1, \ldots, M\}$ is a vector of ad channel variables at time t, $\boldsymbol{z}_t = \{x_{t,c}, m = 1, \ldots, C\}$ is a vector of control variables at time t and Φ is the vector of parameters in the model.
- L indicates the longest lag effect that media or control variables has on sales.
- In order to enable optimization of media budgets and to capture diminishing returns, the response of sales to a change in one ad channel can be specified by a one dimensional curve which is called the response curve for that channel.

A typical regression model

- A more familiar linear functional form specification is:

$$y_t = \beta_0 + \beta_j \times \frac{x^{\alpha}_{decay_{t,j}}}{x^{\alpha}_{decay_{t,j}} + \gamma^{\alpha}} + \beta_{hol} \times \text{hol}_t + \beta_{sea} \times \text{sea}_t + \beta_{trend} \times \text{trend}_t + \ldots + \beta_{ETC} \times \text{ETC}_t + \epsilon$$

- The components of the above specification can be explained as follows:

$$y_t = \beta_0 + \overbrace{\beta_j \times \frac{x^{\alpha}_{decay_{t,j}}}{x^{\alpha}_{decay_{t,j}} + \gamma^{\alpha}}}^{\text{S-Curve component for each media } (j)} + \underbrace{\beta_{hol} \times \text{hol}_t + \beta_{sea} \times \text{sea}_t + \beta_{trend} \times \text{trend}_t}_{\text{holiday, seasonality, and trend effect}} + \ldots + \beta_{ETC} \times \text{ETC}_t + \epsilon$$

In the above specification, y_t is the dependent variable (e.g. either Direct Website Leads or Customer Sales). Main components of the function are:

1. Adstock transformation:

$$x_{decay_{t,j}} = x_{t,j} + \theta_j \times x_{decay_{t,j-1}}$$

2. S-curve transformation:

$$\text{SCurve}_{(x,j)} = \beta_j \times \frac{x^{\alpha}_{decay_{t,j}}}{x^{\alpha}_{decay_{t,j}} + \gamma^{\alpha}}$$

where

β_0 = Intercept
y_t = "revenue" or "conversion" at time t
t = time index (week)
j = media index (e.g. FB, TV, OOH)
β, α, γ, θ = regressor specific to each media j
γ implemented on the S-curve is a transformed γ where γ_{tran} = quantile $(X_{decay_j} \times \gamma)$
$\beta_{ETC} \times ETC_t$ = further independent variables to be added to the model (eg competitor, promotion)
ϵ is an error term, assumed to be well-behaved

We wish to model the following:

- **Context variables.** These are context variables that can help explain the dependent variable behavior in time and that are not paid media. Most common examples of these are: competitors, price and promotion, temperature, unemployment rate, etc.
- **Paid media variables.** These are the names of the media variables that will be used in the model. It is recommended to use metrics that better reflect media-exposure such as impressions, clicks, or gross ratings points (GRPs) instead of spend.
- **Organic variables.** Typically newsletter emails sent, push-notifications, social media posts, etc. Compared to paid media vars, organic vars are often marketing activities without a clear marketing spend.

Adstock modelling

In general, there are two things to consider

Diminishing Marginal Returns: The core premise of television advertising is that exposure to television advertisements creates awareness in the minds of customers to a certain extent. Beyond that, the influence of ad exposure begins to fade over time. Each additional unit of GRP would have a smaller impact on sales and awareness. As a result, the sales generated by additional GRP begin to decline and stabilise. This effect can be observed in the graph above, which shows a non-linear relationship between TV GRP and sales.

Carry over effect or "decay effect": Carry over effect refers to the impact of previous advertisements on current sales. The previous month's GRP value is multiplied by a minor component called lambda. Because the impact of prior months' advertisements fades with time, this component is also known as the Decay effect.

This technique is effective for representing the true carryover effect of marketing initiatives in a better and more precise way. It also aids in the better understanding of degradation effects and how they can be applied to campaign planning. It represents the premise that advertising's impacts might lag and fade after initial exposure. To put it another way, not all of the consequences of advertising are perceived right away—memory grows and people sometimes delay action—and awareness fades over time. In the current Msix model we can select between two adstock approaches in the code:

1. **Geometric:** Traditionally the exponential decay function is used and controlled by theta, the decay parameter. For example, an ad-stock of theta = 0.75 means that 75% of the impressions in Period 1 were brought to Period 2. The traditional exponential adstock decay effect is defined as:

$$\text{decay}_{t,j} = x_{t,j} + \theta \times \text{decay}_{t-1,j} \tag{5}$$

2. **Weibull:** Weibull survival function (Weibull distribution) provides much more flexibility in the shape and scale of the distribution. The formula is defined as:

$$\text{decay}_{t,j} = x_{t,j} + \left(1 - \exp\left(\frac{\text{decay}_{t-1,j}}{\alpha}\right)\right)$$

Diminishing marginal returns

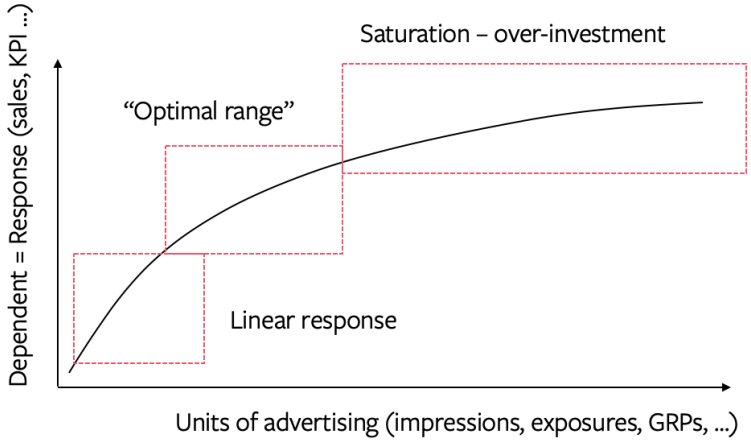


Figure: Diminishing returns.

Each extra unit of advertisement increases response at a decreasing rate (Figure 3). This important marketing principle is portrayed as a variable transformation in marketing mix models. A number of functions can be used to simulate the nonlinear response of a media variable on the dependent variable. For example, we can use a power transformation (x^{α}) or a basic logarithm transformation (taking the log of the units of advertising $\log(x)$). The modeller examines the multiple variables (different levels of parameter x) for the maximum significance of the variable in the model and the highest significance of the equation overall in the case of a power transformation.
However, the most common approach is to use the flexible S-curve transformation:

$$\text{SCurve(x)} = c \times \frac{x^{\alpha}}{x^{\alpha} + \gamma^{\alpha}}$$

where c is the coefficient obtained from regression (maximum saturation), x is the level of (ad-stocked) impressions (exposures), $\alpha > 0$ is the shape parameter, and $\gamma > 0$ is the inflection parameter.

Modelling growth, seasonality, and holidays

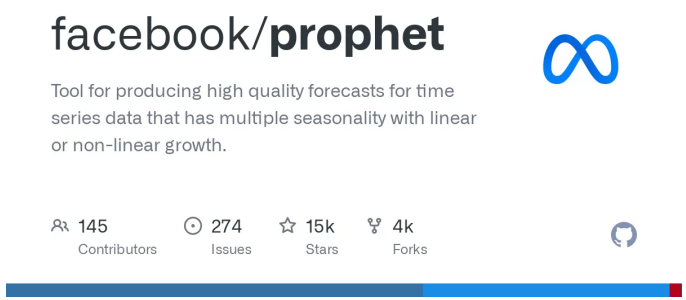
- Working with time series data can be challenging, and the many techniques used to create models may be quite tricky to fine-tune. If you are working with data that contains numerous seasonalities, this is especially true. Traditional time series models, such as SARIMAX, also include a number of strict data requirements, including stationarity and uniformly spaced values.
- In terms of neural network design, other time series models like Recurring Neural Networks with Long-Short Term Memory (RNN-LSTM) may be quite complicated and challenging to work with. Time series analysis thus has a high entrance hurdle for the typical data analyst. In order to address this, Facebook Research released a paper in 2017 titled "Forecasting at Scale" that launched the open-source project Facebook Prophet. Although it has some limitations, it suits our purposes well in the sense that we use to supplement the main model i.e. modelling growth, seasonality, and holidays.
- Historically, traditional time-series methods have dominated the forecasting area. Classic models like Auto-Regressive Integrated Moving Average (ARIMA) and Exponential Smoothing (e.g. Holt-Winters) have undergone extensive research and offer components that may be understood. However, their performance in practical applications is constrained by their limiting suppositions and parametric nature. A proficient forecasting specialist can combine algorithms and adapt data to fulfil particular requirements for better performance. This calls for in-depth expertise in both the application's area and traditional time series modelling.

Methods of statistical machine learning

- Early forecasting competitions consistently saw subpar performance from machine learning (ML) based models. In the past, forecasting experts have even criticised neural networks (NN) as being uncompetitive. They were also criticised for being "black boxes." However, NN-based data-driven algorithms have once again become popular in forecasting due to the boom in the availability of large scale time series. The amount of data required for the training of ML and Deep Learning (DL) approaches is no longer insufficient. But in the sphere of predicting, the explainability of these models is still largely an unsolved research issue.
- Additionally, they frequently necessitate intensive technical work to preprocess data and adjust hyperparameters. As a result, the majority of non-expert forecasters working in various industries do not employ the most precise state-of-the-art models for their particular task. Instead, they are more concerned with locating a model that is a reasonable amount of accuracy, subject to explainability, scalability, and little adjustment.
- These are frequently thought of as prerequisites for forecasting applications in business. Despite their low forecasting effectiveness, practitioners frequently choose classic statistical techniques because they are simple to use and have a clear functional structure. Therefore, it is necessary to create new techniques that can close the gap between traditional time series modelling and ML-based approaches.

Hybrid methods

- Facebook Prophet offers an interpretable model that scalable to numerous forecasting applications, serving as a early example of hybrid methods. The forecasting framework gives beginners complete automation and gives domain specialists the ability to fine-tune.
- Prophet is a well-known forecasting tool, and it is also one of the few that they frequently stick with as their abilities advance. It has opened up traditional time series forecasting to a broad audience and made it practical. Users have faced difficulties as a result of its shortcomings in crucial areas like extensibility and the absence of local context.
- Prophet's utility in industrial applications has been constrained by the absence of local context, which is necessary for forecasting the near future. It is challenging to expand the initial forecasting library for Prophet because it was created on top of Stan, a probabilistic programming language.



Prophet

- A core concept of the Prophet model is its modular composability. The model is composed of modules which each contribute an additive component to the forecast. Most components can also be configured to be scaled by the trend for a multiplicative effect. Each module has its individual inputs and modelling processes.
- However, all modules must produce h outputs, where h defines the number of steps to be forecasted into the future at once. These are added up as the predicted values $\hat{y}_t, \ldots, \hat{y}_{t+h-1}$ for the time series future values $y_t, \ldots, y_{t+h-1}$. If the model is only time-dependent, an arbitrary number of forecasts can be produced. In the following descriptions, that special case will be treated mathematically equivalent to a one-step ahead forecast with $h = 1$.

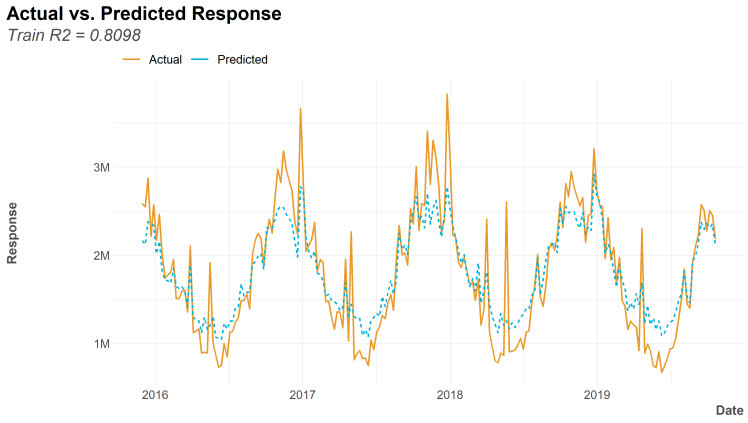
$$\hat{y}_t = g(t) + s(t) + h(t) + e_t \tag{6}$$

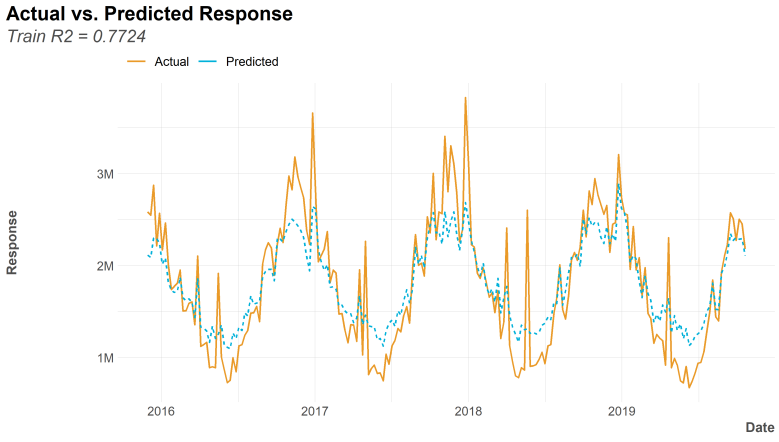
- where,

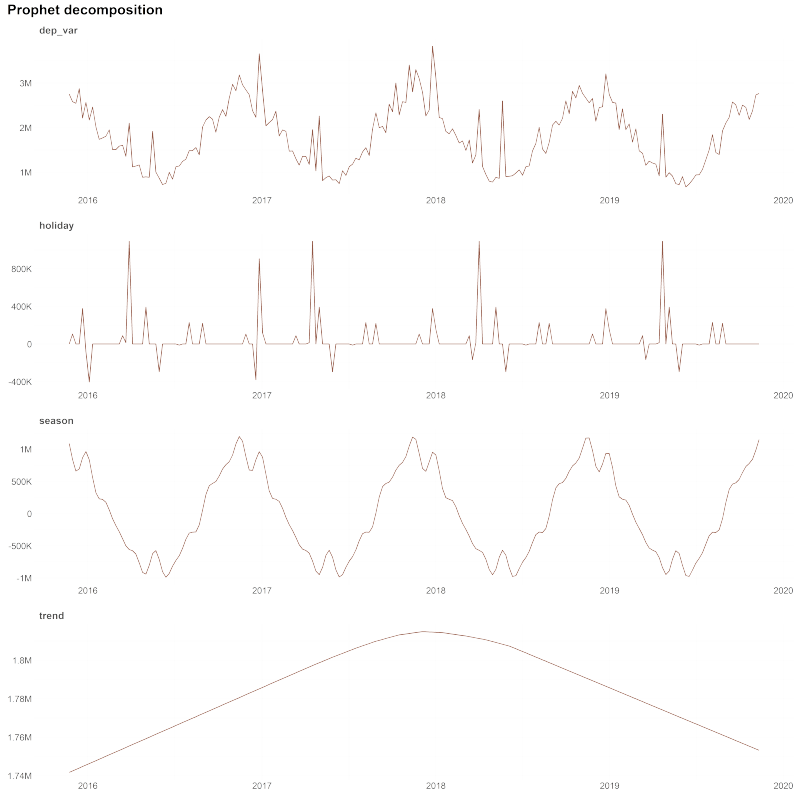
$$\begin{aligned} g(t) &= \text{Trend i.e. growth at time } t \\ s(t) &= \text{Seasonal effects at time } t \\ h(t) &= \text{Event and holiday effects at time } t \end{aligned}$$

  and error e_t.
- All model component modules can be individually configured and combined to compose the model. If all modules are switched off, only a static offset parameter is fitted as the trend component. By default, only the trend and seasonality modules are activated. The full model is summarized in equation 6. In the Appendix we discuss each of the components in more detail.

Visual example







Section 3

Challenges

Challenges

Next section goes over a variety of problems that cast doubt on the accuracy of the findings from MMMs fitted to observational data. These come up frequently for the modeller, but they are frequently not acknowledged or discussed with the MMM's end user. Limitations on the amount of data, bias in the selection of data, and modelling are the three main areas of difficulty.

Limited amount of data

- The modeller frequently has a small amount of data at their disposal. Only 156 data points make up a typical MMM dataset, which is made up of three years' worth of weekly national data. The modeller is required to create an MMM from this, frequently with 20 or more ad channels. For each ad channel, which is typically the case, flexible functional response forms must be required.
- In this case, the number of parameters in the model may be greater than the number of available data points. A lagged effect with a declining return may require 3 to 4 parameters for each channel to be properly modelled. Typical MMMs fall short of the minimum need of 7–10 data points per parameter for a stable linear regression, regardless of how well causal effects are assessed.

Correlated input variables

- Advertisers frequently distribute their budget across advertising channels in a correlated manner, which may make sense from the perspective of maximizing ad effectiveness. These linked advertising decisions can also interact with other marketing decisions to produce large sets of associated input data. For instance, only during a specific season or when another marketing variable is at a high level can a particular ad channel be detected at a high level.
- Highly correlated input variables might produce coefficient estimates with substantial variance when fitting a linear regression model. This can then result in inaccurate sales attribution to the advertising channel. Figure 4 depicts a simplified depiction of a typical scenario a modeller would see as an example.

Correlated input variables

- Two fitted response surfaces that each have a distinct slope relative to each ad channel and each match the data well are shown in the image. The modeller may discover that many surfaces in this dataset offer strong out-of-sample sales prediction accuracy but perform badly when the advertiser deviates from historical spending habits.
- This is due to the fact that when one of the ad channels moves independently of the other, the data reveals little about the results of sales. Another effect is that the estimated association may drastically alter as a result of minute adjustments to the data or the inclusion or exclusion of factors that appear unconnected in the model.

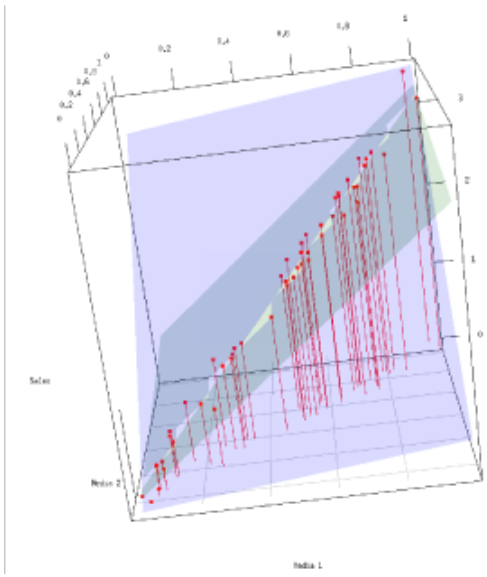


Figure: Two estimated response surface in the presence of correlated variables. Sales is on the z-axis, and there are two ad channels, with the spend of each on the other two axes. The ad channel spend levels are strongly correlated with each other and each plane fits the observed data well despite having different slopes. From [CK17]

Limited range of data

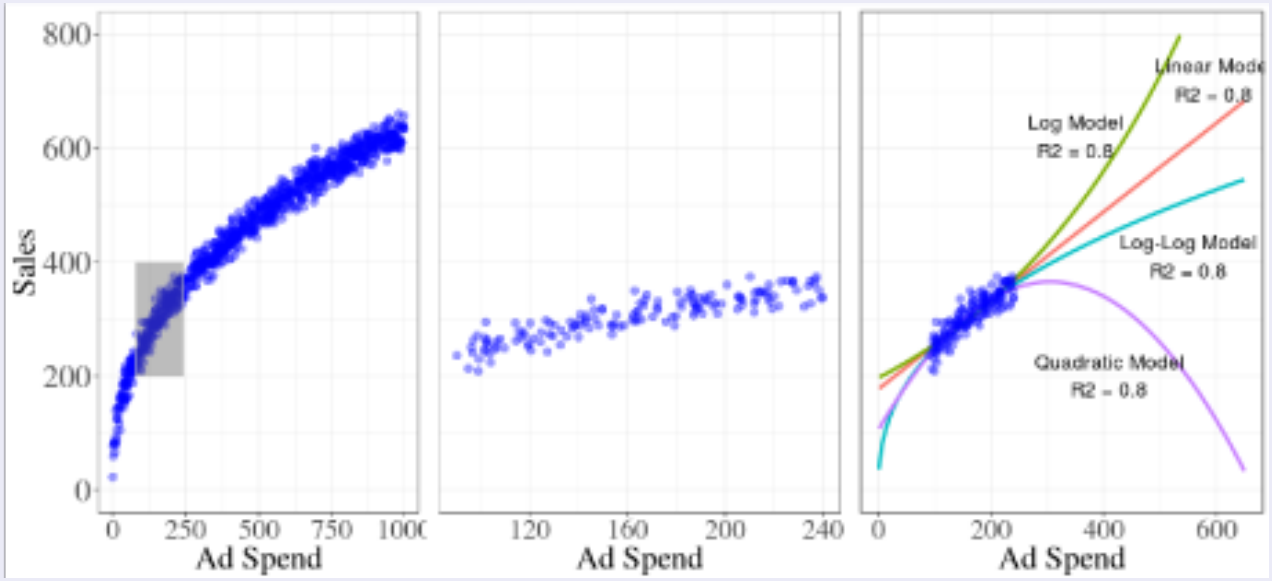


Figure: Extrapolation uncertainty is illustrated when models are fitted to a small set of data. If the advertiser had spent in those ranges, the response curve shown in the figure's left panel might have looked like that. Due to the advertiser's spending only falling within a certain range, the modeller would only have access to the centre panel. We display four fitted response curves on the right panel. From [CK17]

Limited range of data

- Advertisers frequently choose an advertising budget that fits their company's requirements. When models are fitted to this constrained historical data and the model is expected to provide insights beyond the scope of this data, this becomes a problem. For instance, the advertiser might want to know the answer to the question, "What if I double my ad spend next year?"
- As seen in Figure 5, fitting a model to a small set of data subjects the advertiser to high levels of extrapolation uncertainty. The right panel displays four fitted response curves that, while all fitting the observed range of data equally well, result in very different sales results as ad spend increases.
- What would happen if I stopped spending money on ads? is another common extrapolation technique. This extrapolation happens because the advertiser needs to extrapolate back to zero spend in order to determine the average ROAS for an advertising channel. Due to the data available regarding current spend levels in these circumstances, the fitted model may provide reasonable estimates of the marginal ROAS but subpar estimates of the average ROAS.

Model selection and uncertainty

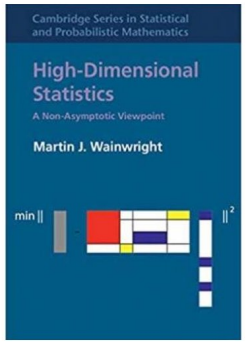
- The modeller faces the challenge that the functional form of $F(\boldsymbol{x}_t, \boldsymbol{z}_t; \Phi)$ and the members of $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ to include in the model are ambiguous. Uncertainty of the functional form for $F(\boldsymbol{x}_t, \boldsymbol{z}_t; \Phi)$ has been long-acknowledged (see [Q64]) and is due to the complexity of the sales response process.
- MMMs fall into the broader category of demand modeling which has a long history in the field of economics (e.g. Deaton and Muellbauer [DM80]; Berry, Levinsohn and Pakes [B+95]). The demand modelling literature provides a starting point for considering what variables to include. However, it does not provide strong guidance on the functional form of the model or the proper control variables.
- The modeller may utilise a model selection procedure based on accuracy measures like R^2 or prediction error to choose a model form and to regard the model with the highest accuracy as a legitimate causal model. However, compared to the size of the input space $(\boldsymbol{x}_t, \boldsymbol{z}_t)$, the number of data points is typically small.

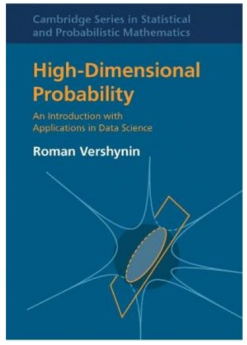
Model selection and uncertainty

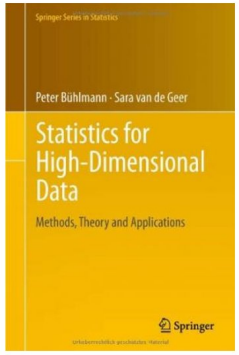
- Because MMM datasets often have poor signal-to-noise ratios, where the ad variables are the signals of interest, the issue of model selection based on prediction accuracy is made even worse. In the absence of ad spend variables, the explanatory power of a few input variables, such as seasonal proxies, price, and distribution, is frequently sufficient to attain high predictive power. By testing various requirements, it is simple to find models with good predicted accuracy. Choose whether to use sales volume or log sales volume as the outcome variable, how to control for price, and how to account for lag effects and diminishing returns are just a few particular instances of specification decisions. Each MMM is supported by dozens of such options.
- The volatility in sales is typically significantly higher than the variance in media spend in real datasets that we have observed.[8] What part do the ad variables play in model selection, given that ad expenditure is a weak signal? A modeller will typically not be able to exclude models that give poor ROAS estimates if their main concern is predictive accuracy.
- Another issue is that the scope for meaningful cross-validation of an MMM is typically limited due to small datasets and the time-series nature of the data which does not allow for simple random sampling to produce a validation set.

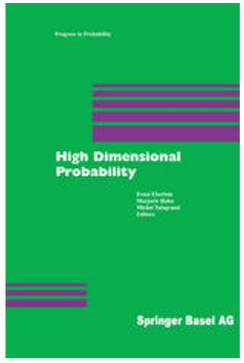
[8] [LR15] explore these challenges in the context of digital advertising

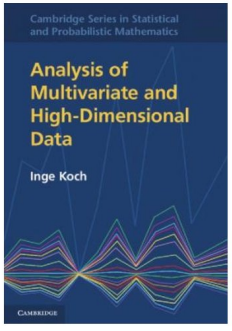
Opportunities

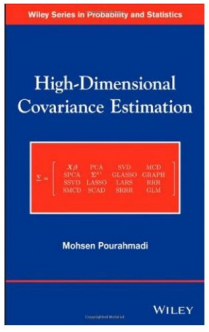


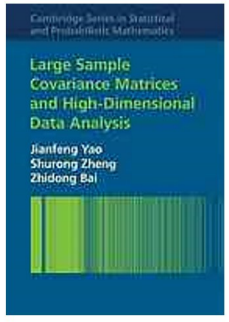


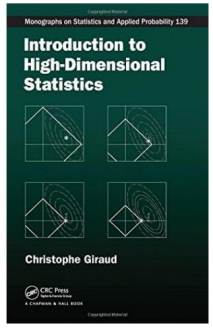












- From its onset, modern statistics engages in the problem of inferring causality from data. A common mindset is that causal inference is only possible using randomised experiments, but developments in statistics and related fields have shown that this view is oversimplified and restrictive.
- However, we now have a much better understanding of the assumptions and methodologies that enable causal inference from nonexperimental data. These texts cover some of the most fundamental ideas in high-dimensional inference, a vibrant research area where statistical theory meets scientific practice.

High-Dimension, Variable Selection and Post-Selection Inference

- L'Hour, J. (2020). L'econometrie en grande dimension. Documents de Travail de l'Insee - INSEE Working Papers, (M2020-01). Book forthcoming in 2023 "Machine learning for econometrics".
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. Journal of Economic Perspectives, 28(2):29-50
- Belloni, A., Chernozhukov, V., Chetverikov, D., and Hansen, C. (2018). High dimensional econometrics and regularized gmm. arXiv:1806.01888, Contributed chapter for Handbook of Econometrics

Using Machine Learning Tools in Econometrics

- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68

Higher-Dimensions and Endogeneity

- Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012a). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6):2369–2429
- Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. Annu. Rev. Econ., 7(1):649–688

Section 5

Innovations at Msix

Internal tool: dryad MMM

- dryad MMM aims to be the "state of the art" model to address the above issues in the MMM context. It provides new functionalities to empower practical applications by using a cross validation framework that assesses the predictive performance and statistical significance of a family of regularized models and of the corresponding features that contribute to prediction. The user can select which quality metrics to use to quantify the concordance between predicted and observed values, with defaults provided for each model.
- Statistical significance for each model is determined based on comparison to a set of null models generated by random permutations of the response; the same permutation-based approach is used to evaluate the significance of individual features. In the analysis of large and complex marketing datasets, such as dryad MMM provides summary statistics, output tables, and visualizations to help assess which subset(s) of features have predictive value for a set of response measurements, and to what extent those subset(s) of features can be expanded or reduced via regularization.

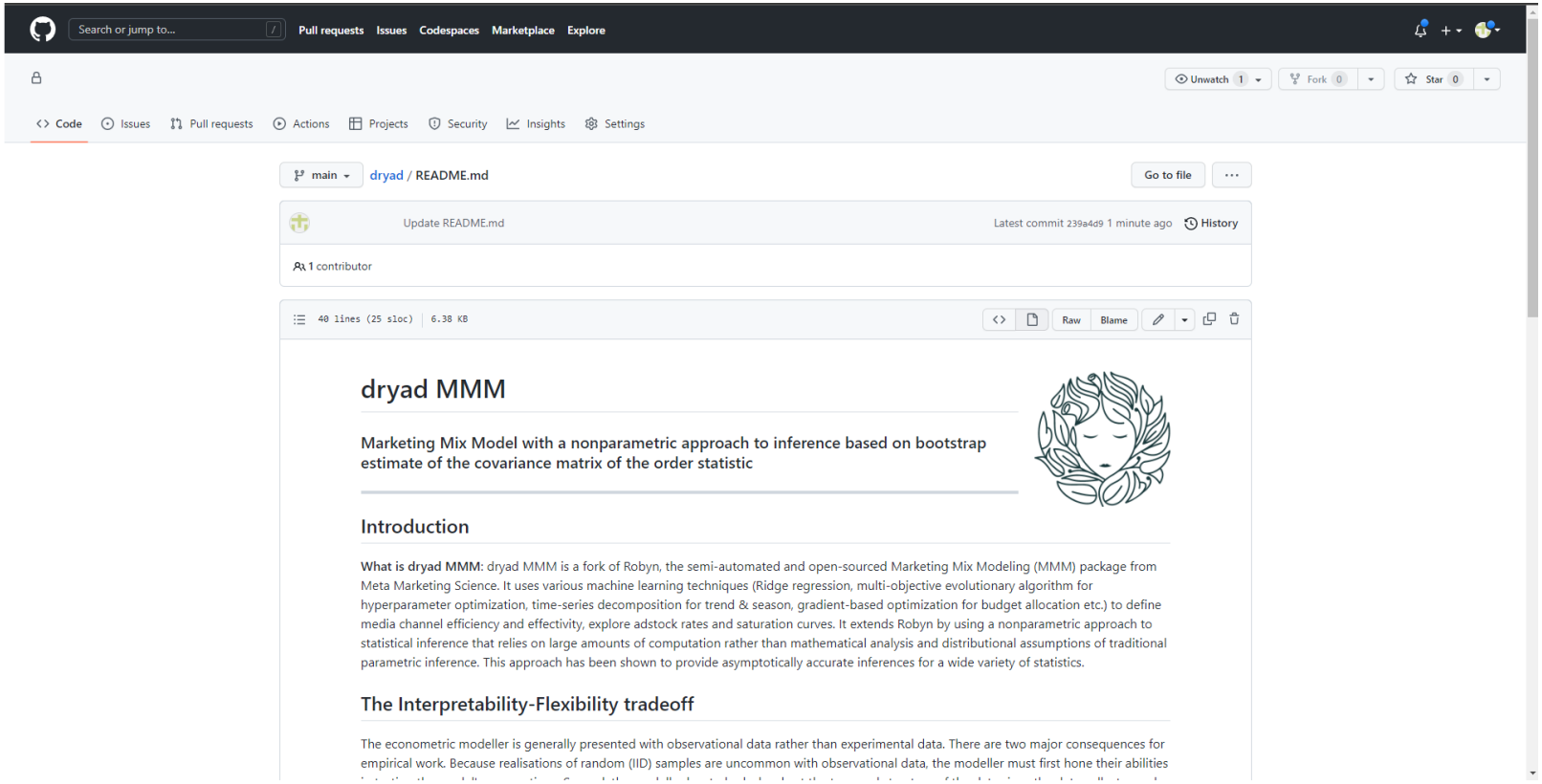


Figure: dryad MMM – custom library for high dimensional MMM inference (R). Under development.

Internal tool: Pegasus

Pegasus is an ongoing project, a collection of data analysis toolboxes extending the MMM frontier. It empower researchers to gain better insight to their data through interactive data visualization, powerful machine learning methods and combining different datasets in easy to understand visual workflows.

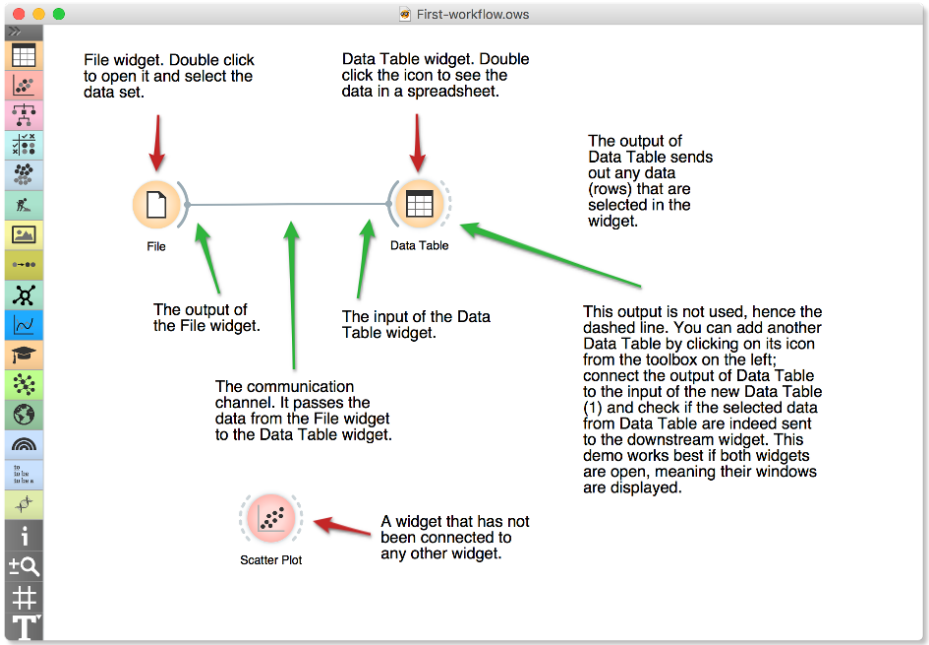
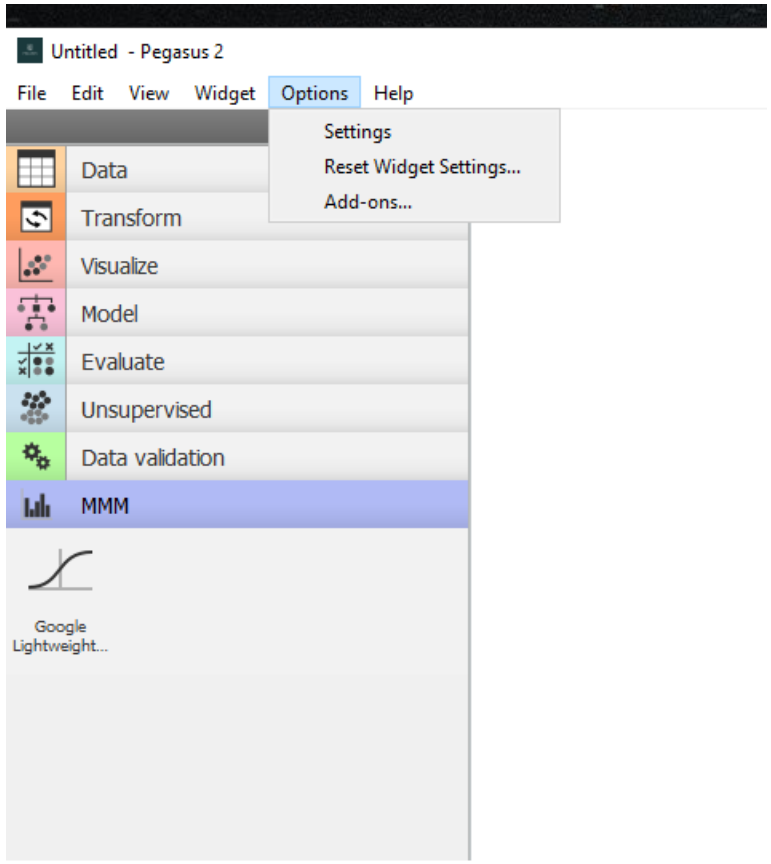


Figure: Pegasus includes state of the art MMM models, processing tools, and visualizations for multiple measurement techniques. This will allow the discovery of new scientific insights through multimodal data analysis.


Untitled - Pegasus 2
File
Edit
View
Widget
Options
Help
Settings
Reset Widget Settings...
Add-ons...
Data
Transform
Visualize
Model
Evaluate
Unsupervised
Data validation
MMM
Google Lightweight...

- We take a variety of approaches while processing data. Data analysis scientists frequently choose to code, whilst non-programmers employ a variety of open-source and commercial applications. Each strategy includes drawbacks including the challenge of utilising and customising task-specific scripts, adjusting to command-line tools, or the expense and rigidity of closed-source environments. Different output data formats are produced when numerous measuring techniques are used, which is frequent in business research. The combined data may be difficult to read if each format needs specialised software.
- Real understanding of complex, multidimensional data volumes generated during company activities necessitates a statistical methodology. For data exploration, analysis, and creating final representations, interactivity is essential.
- New methods for data analysis have emerged as a result of the recent machine learning and deep learning growth in the tech sector, which has also impacted the field of marketing analytics. User-friendly software for such investigations was lacking until relatively recently, and tools that allow the deployment of contemporary machine learning algorithms typically demand substantial programming abilities.
- We believe that expandable, well-designed, user-friendly technologies have the greatest promise in addressing these problems.

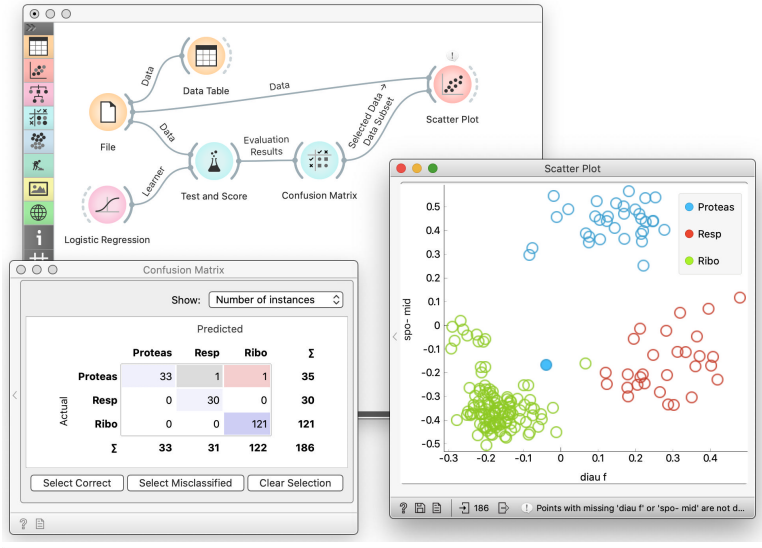
Pegasus provides:

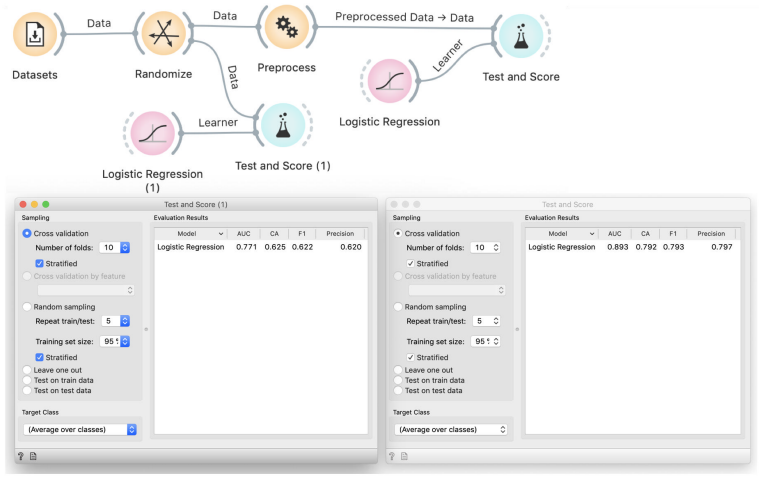
- Immediate feedback: Pegasus abides by the basic tenet that efficiency is improved by the ability to see the results of activities and change right away. Users can build trust in the results and knowledge with the analysis techniques by having the ability to examine the results at every stage of the study;
- Visual programming: Pegasus does not impose a predetermined order of actions while using visual programming. Instead, it provides "widgets," which are components, that either process, visualise, or model inputs. As long as they share connection types, users are free to connect them however they see suitable. This strategy enables the development of adaptable workflows;
- Interactive visualisations: Pegasus allows users to interact with the items that are being shown, which may affect additional analyses. The associated data is sent to the output, for instance, when a point on a scatter plot is selected. This idea gives consumers the power to further examine intriguing components found in the visualisation;
- Statistical Machine learning: Pegasus is a machine learning tool that has assessment components for both supervised and unsupervised models. It mostly transforms well-known machine learning tools, like scikit-learn or XGBoost, into approachable GUI elements. Numerous clustering techniques, t-SNE, random forests, and more specialist MMM-related tools are included;
- Extendability and modularity: Pegasus is mostly written in Python and contains computationally demanding C++ code for extensibility and modularity. It is based on the NumPy, SciPy, Pandas, and scikit-learn Python data science libraries. The Python Script widget enables adding unique Python code to the workflow in cases where the existing components are insufficient. Additionally, Pegasus offers a programming interface that is well-documented for adding additional modules and components.

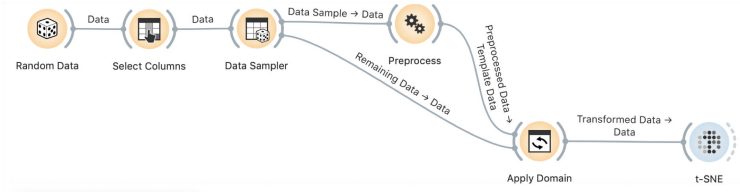
Pegasus summary:

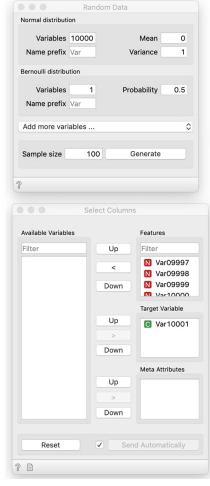
- We provide Pegasus as a single installer, which is a bundled distribution of Orange that has been expanded with particular, preselected add-ons.
- The new Pegasus MMM components were created with maximum flexibility in mind. As a result, the MMM analyst has access to a number of machine learning features, including the ability to immediately use clustering and classification on business data.
- We develop Pegasus as an internal tool available for the sole benefit to the Msix Data Science team. An installation package is developed for Windows only.

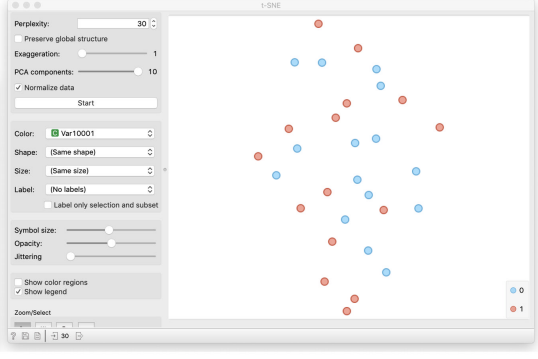
Selected Screenshots

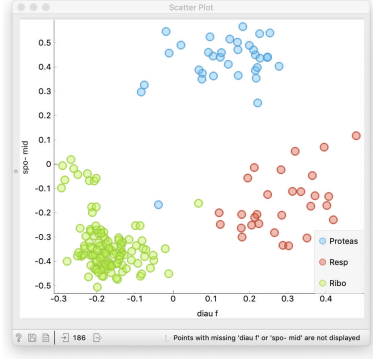


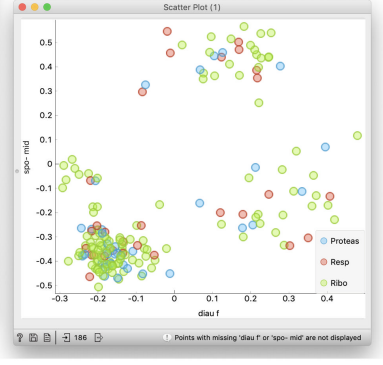












Technologies that we rely on in the Data Science team

Recap

- Challenges, such as Data
  1. a lot of them - large dimensions, hard to parametrise
  2. interrelated - multivariate models
  3. complex - hard to make predictions and validate
  4. non-stationary - small reliable observation set
- Implication:
  - 1 and 2 make the number of parameters large
  - 3 and 4 make the training set small
- Opportunities: make use of recent advancements and research frontier
- Innovations: develop tools to tackle hard problems and obtain a commercial edge

Recent client quote

> "*Your work achieved what we previously considered impossible.*"

Questions?

Section 6

References

References I

[AK99] Angrist, J. & Krueger, A. B. (1999). Empirical strategies in labor economics. In O. C. Ashenfelter & D. Card (Eds.), Handbook of labor economics (Vol. 3, pp. 1277–1366).

[B+95] Berry, S., Levinsohn, J. & Pakes, A. (1995). Automobile prices in market equilibrium. Econometrica, 63 (4), 841–890.

[B+12] Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6), pp.2369-2429.

[B64] Borden, N. H. (1964). The concept of the marketing mix. Journal of advertising research, 4 (2), 2–7.

[C+10] Chan, D., Ge, R., Gershony, O., Hesterberg, T. & Lambert, D. (2010). Evaluating online ad campaigns in a pipeline: causal models at scale. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 7–16). KDD '10. Washington, DC, USA: ACM. doi:10.1145/1835804.1835809

[C+17] Chen, A., Chan, D., Perry, M., Jin, Y., Sun, Y., Wang, Y. and Koehler, J., 2018. Bias correction for paid search in media mix modeling. arXiv preprint arXiv:1807.03292.

[CK17] Chan, D. & Koehler, J. (2017). Bayesian methods for media mix modeling with carryover and shape effects. research.google.com

[CP17] Chan, D. and Perry, M., 2017. Challenges and opportunities in media mix modeling. Google Research working paper.

[CC08] Chen, J. and Chen, Z., 2008. Extended Bayesian information criteria for model selection with large model spaces. Biometrika, 95(3), pp.759-771.

[D] Deloitte, The future is modeled: A how-to guide for Advanced Marketing Mix Models deloitte.com

[DM80] Deaton, A. & Muellbauer, J. (1980). An almost ideal demand system. American Economic Review, 70 (3), 312–326.

[HK08] Hinkelmann, K. & Kempthorne, O. (2008). Design and analysis of experiments, vol 1, (2nd ed.). Wiley.

[MP02] Minerva, T. and Paterlini, S., 2002, May. Evolutionary approaches for statistical modelling. In Proceedings of the 2002 Congress on Evolutionary Computation. (Vol. 2, pp. 2023-2028). IEEE.

[IR15] Imbens, G. W. & Rubin, D. M. (2015). Causal inference for statistics, social, and biomedical sciences. Cambridge University Press.

[NW06] Jorge, N. and Stephen, J.W., 2006. Numerical optimization.

[LP07] Lambert, D. & Pregibon, D. (2007). More bang for their bucks: assessing new features for online advertisers. In Proceedings of the 1st International Workshop on Data Mining and Audience Intelligence for Advertising (pp. 7–15). ADKDD '07. San Jose, CA: ACM.

[LR15] Lewis, R. A. & Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. Quarterly Journal of Economics, 130 (4), 1941–1973.

[MB06] Meinshausen, N. and Buhlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. The annals of statistics, 34(3), pp.1436-1462.

[M78] McCarthy, J. E. (1978). Basic marketing: a managerial approach (6th ed.). Homewood, Il: R.D. Irwin.

References II

[P09] Pearl, J. (2009). Causality: models, reasoning and inference (2nd ed.). Cambridge University Press.

[Q64] Quandt, R. E. (1964). Estimating the effectiveness of advertising: some pitfalls in econometric methods. Journal of Marketing Research, 1 (2), 51–60.

[S15] Stan Development Team. (2015). Stan modeling language user's guide and reference manual. Retrieved from `http://mc-stan.org`

[S+17] Sun, Y., Wang, Y., Jin, Y., Chan, D. & Koehler, J. (2017). Geo-level bayesian hierarchical media mix modeling. research.google.com.

[T96] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp.267-288.

[VK11] Vaver, J. & Koehler, J. (2011). Measuring ad effectiveness using geo experiments. research.google.com

[W+17] Wang, Y., Jin, Y., Sun, Y., Chan, D. & Koehler, J. (2017). A hierarchical bayesian approach to improve media mix models using category data. research.google.com.

[ZV17] Zhang, S. S. & Vaver, J. (2017). Introduction to the Aggregate Marketing System Simulator. research.google.com

[ZH05] Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), pp.301-320.

Section 7

Appendix

Randomized experiments

- The generally accepted gold standard for answering causal questions is to perform a randomized experiment.
  - A randomised experiment answers the question of what would happen if an advertiser did action X by randomly splitting a population into a test group, where action X is done, and a control group, where nothing is done. Randomization controls for all other sources of variation, so the only difference between the test and control groups, from a statistical point of view, is the action X.
  - Then we can say that action X caused any change in the results between the two groups that we can see.[9]
- Randomized experiments can be done at different levels, such as by the user, by the store, or by the area.
  - The level of the experiment is limited by how well action X can be targeted and how well results can be tracked. Even though randomised experiments are the gold standard, they are not used as much as they could be to find answers that an MMM might give.
  - To answer the same kinds of questions that MMMs do, the advertiser may need to run a lot of tests under a lot of different circumstances over time.
  - For example, let us say an advertiser wants to know how effective an ad channel is at different levels of ad spend, not just at one level. The advertiser also uses an MMM to get the most out of their media budgets, which will take a lot of experiments to figure out.
  - **Problem:** this is not possible for most advertisers, if not all of them. Other things that keep people from using randomised experiments are the technical difficulties of setting them up, the lost opportunities that come with having a control group, the costs of having a test group, and weak advertising effects that may need very large sample sizes [10].

[9] For a general introduction to experiments, see [HK08].

[10] See the excellent paper in Quarterly Journal of Economics by [LR15]

Potential outcomes

- If advertisers can not do large-scale randomised experiments and have to rely on historical data instead, they could use the potential outcomes framework, also called the Rubin causal model for causal inference (see Imbens and Rubin, [IR15]) to try to figure out what caused what.
- Some interesting papers from the Google Research team [C+10, C+17, LP07] explore this framework in the context of digital advertising.
- Suppose we want to estimate the ROAS of a single ad channel for an advertiser and the ad channel is either on or off during each time period t. Let y_t be the sales for time period t and x_t be an indicator of whether the ad channel was on or off. For any time period t, there are two potential sales outcomes:

$$y_t^1 : \text{ potential outcome if ad channel turned on}$$
$$y_t^0 : \text{ potential outcome if ad channel turned of}$$

- The causal ROAS over the time periods is given by

$$\mathbb{E}\left[y_t^1 - y_t^0\right]$$

Potential outcomes

- The challenge is that only one potential outcome is observed for each time period. A naive estimate would be to compare sales in the periods when the ad channel was on, to sales in the periods when the ad channel was off. This would give:

$$\begin{aligned}\mathbb{E}\left[y_t \mid x_t = 1\right] - \mathbb{E}\left[y_t \mid x_t = 0\right] &= \left[\mathbb{E}\left[y_t^1 \mid x_t = 1\right] - \mathbb{E}\left[y_t^0 \mid x_t = 1\right]\right] + \\ &= \left[\mathbb{E}\left[y_t^0 \mid x_t = 1\right] - \mathbb{E}\left[y_t^0 \mid x_t = 0\right]\right]\end{aligned}$$

- The $\mathbb{E}\left[\cdot\right]$ operator here indicates the sample average in the data. The term in the first set of square brackets represents the causal effect of interest during the treated time periods. It is the difference between the average sales that occurred when $x_t = 1$ and the average sales that would have occurred during the same time periods if $x_t = 0$.
- Selection bias is represented by the term in the second set of square brackets. This term would by design have an expected value of zero in a randomised experiment. In this case, the word "selection bias" refers to any biases in the treatment selection mechanism that are also connected with the outcome (sales), which could refer to anything that affects ad expenditure.

Potential outcomes

- Selection bias can be caused by the advertiser's activities, those of potential customers, or those of rival advertisers. For instance, if a marketer bases the timing of their commercials on the seasonality of consumer demand, then

$$\mathbb{E}\left[y_t^0 \mid x_t = 1\right] \neq \mathbb{E}\left[y_t^0 \mid x_t = 0\right]$$

- It is this selection bias that makes trying to answer causal questions with observational data one of the most demanding problems in applied statistics. This selection bias needs to be accounted for in order to produce a valid causal result. One way to control for selection bias is through the use of the matching estimator described below.
- Let $\mathbf{z}_t$ be the vector of control variables which could potentially affect sales at time t. These control variables would include other ad channels. The matching estimator can be defined as

$$\mathbb{E}\left[y_t^1 - y_t^0 \mid x_t = 1\right] = \sum_{\mathbf{z}|=1} \mathbb{E}\left[y_t \mid x_t = 1, \mathbf{z})\right] - \mathbb{E}\left[y_t \mid x_t = 0, \mathbf{z})\right] \times P(\mathbf{z}_t = \mathbf{z} \mid x_t = 1)$$

  $P(\cdot)$ indicates the empirical probability. The matching estimator assumes that there are observations where $x_t = 0$ for all combinations of control variable values that occur in the case when $x_t = 1$.

Potential outcomes

- In the case of correlated media variables, for instance, this assumption could not always be valid. The number of possible combinations of various control values grows exponentially as the number of control variables rises. This assumption is nearly impossible to meet in a typical MMM case, when the number of data points is tiny, in order to apply the matching estimator to produce the complete range of ROAS estimations.
- With both randomized experiments and use of matching estimators being infeasible or impractical, advertisers often turn to regression models to answer their questions around advertising effectiveness.

Growth function $(g)t$ and Change Points

- The growth function represents how the data have been trending. The useful concept added to Prophet is that the growth trend may be constant throughout the data or may vary at certain changepoints.
  - Changepoints are points in time where the direction of the data changes. It could be because new cases of COVID-19 are starting to decline after peaking once a vaccination is become available. Or there can be an unexpected increase in instances when a new strain is spread across the community, etc. Change points can be manually established or automatically detected by Prophet. Both the quantity of data used in automated changepoint identification and the influence change points have on the growth function are adjustable.
- Within Prophet there are three primary alternatives for the growth function:
  - Linear Growth: Prophet's default option is linear growth. It employs a series of linear piecewise equations with variable slopes between change locations. The growth term for linear growth will resemble the familiar equation, y = mx + b, with the exception that the slope (m) and offset (b) are variables and will change value at each changepoint.
  - Logistic: When your time series contains a cap or a floor where the values you are modelling get saturated and can not rise over a maximum or minimum value, you should utilise the logistic growth setting (think carrying capacity). When logistic growth is used, the growth term will resemble a conventional logistic curve equation (see below), with the exception that the carrying capacity (C), growth rate (k), and offset (m) are all variable and will change value at each change point.

$$g(t) = \frac{C(t)}{1 + x^{ik(t-m)}}$$

  - Flat: In the event that there is no long-term growth, you might select a flat trend (but there still may be seasonality). The growth function will have a constant value if flat is selected.

Modelling trend

- The usual approach to modelling trend is to model it as the combination of an offset m and a growth rate k. The trend effect at a time t_1 is given by multiplying the growth rate by the difference in time since the starting point t_0 on top of the offset m.

$$T(t_1) = T(t_0) + k \cdot \Delta_t = m + k \cdot (t_1 - t_0) \tag{7}$$

- We allow the growth rate to change at a number of locations. Thus, the trend is modelled as a continuous piece-wise linear series. This results in an interpretable, yet non-linear form of trend modelling. It is simple to interpret, as in a segment between two points, the trend effect is given by the steady growth rate multiplied by the difference in time. We can generalize the trend by defining a time-dependent growth rate $\delta(t)$ and a time-dependent offset $\rho(t)$.

$$T(t) = \delta(t) \cdot t + \rho(t)$$

- The piece-wise linear trend only varies the growth rate at a finite number of changepoints. A set C of n_c changepoints are defined at different times as $C = (c_1, c_2, \ldots, c_{n_c})$. Between changepoints, the trend growth rate is kept constant. The first segment's growth rate and offset are given as δ_0 and ρ_0 respectively. Rate adjustments at each changepoint can be defined as a vector $\delta \in \mathbb{R}^{n_C}$, where δ_j is the rate change at the j^{th} change-point. The growth rate at time t is determined by adding the initial growth rate δ_0 with the summation of the rate adjustments at all the change-points up to time step t. Each growth rate change δ_j is a parameter to be fitted on the data.

Modelling trend

- Prophet provides a practical, semi-automatic mechanism for the selection of relevant change-points. Given the number n_C of desired change-points, n_C equidistant points along the series are selected as initial changepoints. Optionally, their growth change rate parameters can be regularized during model training.
- This is is similar to a fully automatic changepoint selection, as only the most relevant changepoints will be selected, if any. The user can also opt to manually define the specific times of a custom number of changepoints. To avoid overfitting on a small number of final points, the final trend segment (after the last changepoint) is set to a larger set of observations (eg 15 % of training data). When making predictions into the unobserved future, the final growth rate is used to linearly extrapolate the trend.

The Seasonality function $s(t)$

- The seasonality function makes use of a Fourier series. A simple approach to conceptualise the Fourier Series if you are new with it is as the sum of several successive sines and cosines. A certain coefficient is multiplied by each sine and cosine term. This allows us to approximate the seasonality (cycle trend) in our data.
- In this technique, a number of Fourier terms are defined for each seasonality as in Equation 82, where k refers to the number of Fourier terms defined for the seasonality with periodicity p. Fourier terms are defined as sine, cosine pairs and allow to model multiple seasonalities as well as seasonalities having non-integer periodicities such as yearly seasonality with daily data ($p = 365.25$) or with weekly data ($p = 52.18$). In a multiple seasonality scenario, different values for n can be defined for each periodicity.

$$S_p(t) = \sum_{j=1}^{k} \left(a_j \cdot cos\left(\frac{2\pi jt}{p}\right) + b_j \cdot sin\left(\frac{2\pi jt}{p}\right) \right)$$

- Fourier terms are a great tool for modelling seasonality as they produce smooth functions which are simple to interpret and stable to fit to data. However, Fourier terms only model deterministic seasonal shapes which are assumed to be fixed through time. A higher number of Fourier terms allows the model to fit a more complex seasonal pattern. Too much flexibility may lead to overfitting or to random patterns between observations.

Modelling seasonality $s(t)$

- Thus, each Fourier term corresponds to a frequency proportional to $\frac{j}{p}$, modelled by by a weighted combination of a sine and cosine transform. Every seasonality is associated with $2k$ number of coefficients. For time step t, the effect from all the seasonalities considered in the model can be indicated by $S(t)$ in below Equation 83, where $\mathbb{P}$ refers to the set of all the periodicities.

$$S(t) = \sum_{p \in \mathbb{P}} S_p^{\star}(t)$$

- Both additive and multiplicative seasonal patterns are supported. Each seasonal periodicity $S_p^{\star}$ can individually be configured to be multiplicative, in which case the component is multiplied by the trend.
- The framework activates daily, weekly and or yearly seasonality depending on data frequency and length. Each of these three types of seasonal periodicities is activated if the data frequency is higher resolution than the respective periodicity, and if at least two full periods of data are available.
- As an example, if the data is of daily frequency, the model will enable yearly seasonality if the data spans two years or more. Weekly seasonality will also be added if two or more weeks of data are available. Daily seasonality will not be activated, as the daily frequency is not of higher resolution to allow for intra-day seasonality. The default number of Fourier terms per seasonality are: $k = 6$ for $p = 365.25$ yearly, $k = 3$ for $p = 7$ weekly, and $k = 6$ for $p = 1$ daily seasonality.

Modelling events and holidays $h(t)$

- Prophet allows the modelling of two types of events; 1) user defined 2) country specific holidays. Given a country name, its national holidays are automatically retrieved and added to the set of events $\mathbb{E}$. Similar to seasonal effects, events can also be specified as either additive or multiplicative.
- Additionally, for a given event at time t_e, a window $[t_e - i, t_e + j]$ of $i + j$ days can be configured to be considered as special events of their own. For example, by setting a window of $[-1, 0]$ for Christmas day, will allow the day before Christmas to have its own effect on the forecast. Hereby, a new variable is created for each day within the window around event and added to the set of events $\mathbb{E}$.

Missing data

- Missing data is less of an issue when working with non-lagged input variables, as corresponding timesteps can simply be dropped. In doing so, one data sample will be lost per missing entry. With Auto-regression or lagged regressor modules in use, a missing data point would lead to $h + p$ data samples being dropped due to h missing forecast targets and p missing lag values. For example, a single missing point leads to the loss of 13 samples for a model forecasting $h = 3$ steps ahead with AR($p = 10$).
- We automatically implement a data imputation mechanism to avoid excessive data loss when working with incomplete data. The imputation mechanism follows the following heuristics:
  - **Data Imputation** If not specified or missing, events are assumed to not be happening. Missing Events are filled in with zeros, indicating their absence. All other real-valued regressor variables, including the time series itself, if autoregression is enabled, are imputed in a three step procedure.
  - **First** The missing values are approximated by a bi-directional linear interpolation. Hereby, the last and the first known value before and after the missing values are used as anchor points for the interpolation. This is done for up to 5 missing values in each direction. If there are more than 10 missing values, they will remain *NaN* after this step. The amount of missing values to interpolate is user-settable.
  - **Second** The remaining missing values are imputed with a centred rolling average. The rolling average is computed over a window of 30, and fills at most 20 consecutive missing values. The amount of missing values to fill with a rolling average is user-settable.
  - **Third** If there are more than 30 consecutive missing values, the imputation algorithm aborts and instead drops all the missing datapoints.
- Prophet utilizes L-BFGS [NW06], implemented in Stan for fitting model parameters to the data.