



What kind of Bayes are you using?

JUNE 29, 2022 – MARK GARRATT, SANDEEP CONOOR

Synopsis: Bayesian methods have proven so dominant in the last 20 years that they have become the preferred tool for predictive models, including Marketing Mix Models (MMM). With a few exceptions, almost every vendor says: “Yes, we use Bayes.” Under the hood, however, there are a lot of different versions of Bayes, and they produce different results. These differences are big enough to change decisions made by media managers and trade specialists. This whitepaper clarifies various types of Bayes models and shows how in4mation insights has advanced marketing science by developing **Robust Bayes™** - the most accurate regression method for dealing with increasingly fragmented data.

Bayes has come to mean a lot of things:

Many marketers are familiar with the term “Bayes”, and some even know that it is named after the Rev. Thomas Bayes, an English vicar who lived from 1701-1761. The probability calculus that Thomas Bayes invented remained a curiosity piece for 200 years because problems of any realistic scale were too hard to estimate. But it was only in the 1990’s that estimation methods conceived in the early Cold War by physical chemists became practicable due to computing technologies. Now, Bayesian methods dominate every field of science. But what does “Bayes” mean in marketing?

First, let’s clear up some confusion. The kind of Bayes that gets applied to MMM in a regression setting is not a **Bayesian Network**. A Bayesian network uses data based on discrete events to derive the conditional probability of outcomes of interest. The variables are usually small sets of alternatives (events) cross-tabulated with other events. The method is not used for problems like MMM where the dimensions (e.g., stores, time periods, UPC level products, etc.) have much greater granularity than the typical network problem.

At the next level of possible confusion is the difference between **empirical Bayes** models (also called random effects or mixed-effects models) and what is often called fully Bayes. Both empirical Bayes and **fully Bayes** have the same basic mechanics:

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

The *prior* is a set of constraints or normative values that is not dependent on the data at hand. These can come from industry experience, a prior study, or a set of prior studies combined into a database of norms. The likelihood is based on the sample data (consumer research) or on the ePOS and media spend data (MMM). The posterior is generated by combining the prior and the likelihood. The posterior is what is reported as the outcome of the study; for instance, they are the coefficients or the betas that drive the KPIs for a marketing mix model (MMM).

Both empirical Bayes and fully Bayes share the property we call *shrinkage*. The idea is that you can get more precise results (lower variance betas) by introducing bias via the prior. This may sound odd that you want to introduce bias into something. But think of it this way, your data is not “truth;” it is time-bound, potentially confounded (multicollinear), conditional on many factors (who owns it, releasability, granularity, aggregation), and subject to its other hidden biases. **The amazing result of Bayesian analysis is that by introducing bias via the prior, the model is more predictive of unobserved outcomes (e.g., future time periods, observations that are “held out”) than by using just the sample consumer data or ePOS/spend data alone.** For an approachable mathematical overview of this, see Casella (1985).

Where empirical Bayes and fully Bayes part ways is that fully Bayes methods precisely model the uncertainty in the *shrinkage points* (the priors) that govern the posterior betas. This leads to more accurate estimates of central tendency (the posterior mean) but also ones where the *spread* in the betas across the units of observation (the individuals, stores, DMAs, UPCs, etc.) is more informative and more influenced by the data. This difference matters a lot. Empirical Bayes and fully Bayes models both use shrinkage to pull outliers towards a point of central tendency. This constraining effect helps immensely in producing reasonable results as we dive deeper into granularity – when we are trying to get estimates for scores of retail chains, hundreds of UPCs, or results down to individual people. However, as we will show later, the empirical Bayes models do this at the price of being very rigid; they force the units of observation into a tight normal distribution, not allowing the betas to vary very much and creating a false sense of precision. The fully Bayes methods, on the other hand, allow the patterns in the spread of the betas to emerge more faithfully from the data. Now that the reader is familiar with basic terms, we are ready to demonstrate important differences between methods and how they would change the decisions made by media managers and trade specialists.

Shrinkage in action:

The following chart (Fig. 1) shows the result when we take a promotional variable (in this case an ad feature) and compare the results of doing separate regressions on each UPC versus combining all those regressions into one fully Bayesian model. The data used here and in all the following examples consists of 2000 UPCs with data from 2019 to 2021 from the same, large CPG category:

In Fig. 1, the x-axis contains values of the coefficient for the ad feature using separate regressions and the y-axis shows the same data evaluated using a fully Bayesian model. A brief look at the values of the x-axis shows that they vary more widely than the Bayes model.

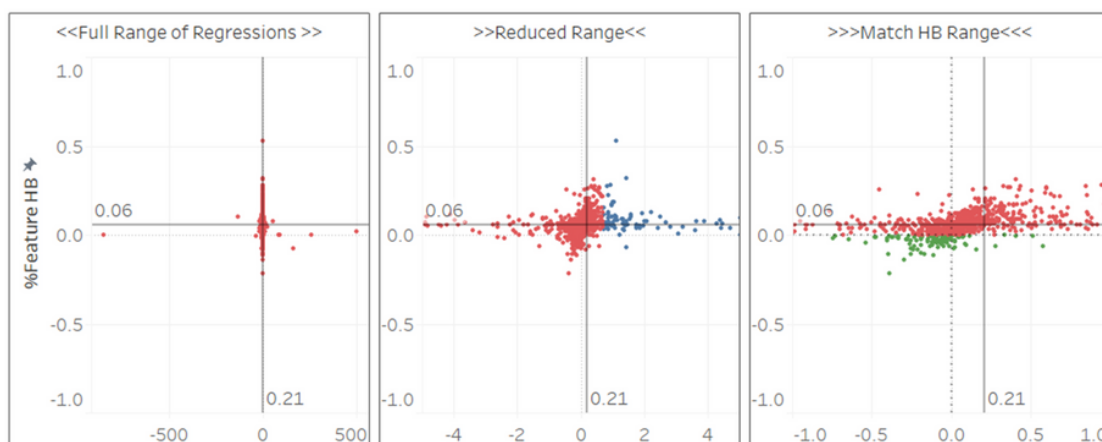


Fig. 1: Comparison of coefficients for ad feature: separate regressions (x) versus Bayes (y)

Let us note right away that the spread in the individual regressions is far too wide to produce a reasonable estimate of the ad feature effect. First, the effect should be positive; if you run an ad, sales should increase. Second, the lift from a feature ad in this category typically ranges from 5% to perhaps 100%. These would translate into values on the x-axis between 0.05 and 0.7. However, from the middle chart, we can see that there are many values above 0.7 (dots shaded in blue) among the separate regressions and none over that value for Bayes.

The leftmost chart shows the full range of the individual regressions. These values are likely the result of numerical problems with the models due to having UPCs with too few weeks of ad features. We would certainly call these *outliers*. However, the dots in blue in the middle chart are also outliers by any reasonable standard. The chart on the right attempts to match the scale of the separate regressions with the scale of the Bayes model. Here (and in the other charts), we see that the range of the Bayes estimates is much tighter, between -0.3 and 0.3, as a result of Bayesian shrinkage. Also, we can see from the chart on the right that the proportion of wrong signs (negative values for ad feature, colored green) is quite a bit lower for the Bayes model and the absolute values of those wrong signs are minor. **So, shrinkage is having the beneficial effect of both keeping the estimates positive and keeping the positive values in a reasonable range for ad feature lifts in this category (between about 5% lift and 40% lift).**

This example demonstrates the immense value of shrinkage to analysts looking for models that produce reasonable outputs. To compare a Bayes model to separate regressions is, admittedly, to use a strawman. Industrial MMM suppliers stopped using separate regressions about 20 years ago, as methods for random effects became widely available (say, through SAS PROC MIXED) and as Bayesian routines started to become technically feasible. However, many analysts, when pushed against deadlines, might run individual regressions, and many data mining packages might offer the option to run hundreds of separate regressions. To take this story further, we can actually compare empirical Bayes results to fully Bayes results. To do that, we introduce one more idea – **hierarchical Bayes**.

Hierarchical Bayes (HB)

The HB model we describe here follows Rossi et al. (2005). In an HB model, there are two levels: the *lower model* and the *upper model*. In the lower model, the betas are the familiar drivers of the dependent variable (usually sales). There is a beta for distribution, a beta for price, a beta for social media advertising, etc. These betas are the basis for calculating *elasticities* such as the impact of an increase of X% in social media spend on sales.

In the upper model, the betas are themselves predicted by variables that we have a *prior belief* could be the basis for differences amongst the betas. These upper model variables might be quite different based on the units of observation. If the units of observation are people, the upper model variables might be demographics. If the units are UPCs, the upper model variables might be product characteristics. If the units are DMAs (geographies), the upper model variables might be weather, house prices, unemployment statistics, neighborhood socio-economics or COVID infection rates, etc.

The value of the upper model is to permit **a relaxation of the shrinkage property**. Instead of all values of the MMM coefficients being shrunk towards one central tendency, different UPCs might be shrunk to different points corresponding to package types, sub-brands, etc. Different stores might be shrunk towards different shrinkage points depending on store format, location, or price zone.

Marketing mix problems often use data that maps naturally to the lower and upper models. In the typical marketing mix problem in CPG, there are usually two dimensions that drive the scope and complexity of the problem: the number of products (e.g., UPCs, SKUs, PPGs, etc.) and the number of geographies (e.g., stores, RMAs, CTAs, SMAs, DMAs, States, etc.). The observational units of analysis in the typical CPG MMM model are “product-geography-time.” For instance, in the example we are using in this white paper, the breakdown of the model is “UPC-chain-week.” A single upper model can contain shrinkage points corresponding to *multiple dimensions*: the UPCs may shrink to a single central tendency, but also be relaxed to have different shrinkage points for a brand or package type. The chains may shrink to a single central tendency but also to the type of outlet (grocery, convenience, dollar stores, etc.) or the physical geography. We do both types of shrinkage in the same model.

Robust Bayes™

Years of work with HB models also pointed out a potential weakness that comes about with any kind of shrinkage. Using in-store displays as our example, what values should we give to UPCs that had never executed a display in the time period? Should they get the coefficient value of the central tendency? Should they get the value associated with the relaxed shrinkage point for package type or brand? Should they be set to zero? Now, for sure, this problem of what to do for UPCs that don't display only comes up with Bayesian models – because Bayesian models have a prior that will populate a value whether or not there is data; *other methods would completely fail to provide a value*. This property can be quite useful; it is a *lookalike* analysis. If brand X has never put package Y on display then maybe we could look at brand X's other displays, or how package Y displays did with other brands? Only a hierarchical Bayes model can do that triangulation precisely.

However, along the way, we learned that **there has to be a minimum amount of data for the hierarchical Bayesian model to provide reasonable and accurate results**. For the prior to provide good shrinkage points, it has to receive reasonable values from the data. The ideal case, for instance (using in-store displays again) is for a UPC with only 5% of weeks with a display to get help from a shrinkage point based on other UPCs with say 10% to 30% of weeks with a display. The stronger data helps the weaker data and *leads to better predictions for the weaker data*. However, if UPCs with less than 5% of weeks with a display or no displays at all contribute to the upper model, they are likely to *over-bias* the shrinkage point and then populate the no display cases with these over-biased values. These then feed into the next cycle of the Bayesian algorithm and further perpetuate the bad data.

This is especially problematic when a promotion only occurs in a small fraction of chains/stores or in a small fraction of weeks. Furthermore, if the promotion or advertising condition never occurs in a particular retailer, or for a particular product, there may be *structural* reasons for that. For instance, convenience stores may never have features, dollar stores may never have discounts, Walmart may have different promotional formats than Target, etc. **Therefore, it's critical to mask the data and make it inoperative where structural gaps exist**. The statistical methodology that we use to implement this, including the rules for inclusion and exclusion, is called **Robust Bayes™**. This represents a major adaptation of HB to MMM models that must deal with increasingly granular data such as digital shopper and retailer media.

To see these differences clearly, we will now work through the four basic approaches we have touched on in this paper using in-store displays as our theme. They are:

1. Separate OLS regressions
2. Random Effects Models/Empirical Bayes
3. Hierarchical Bayesian Models
4. Robust Bayes

The chart below (Fig. 2) shows the values for up to 2000 UPCs for the coefficients for in-store displays using four different models:

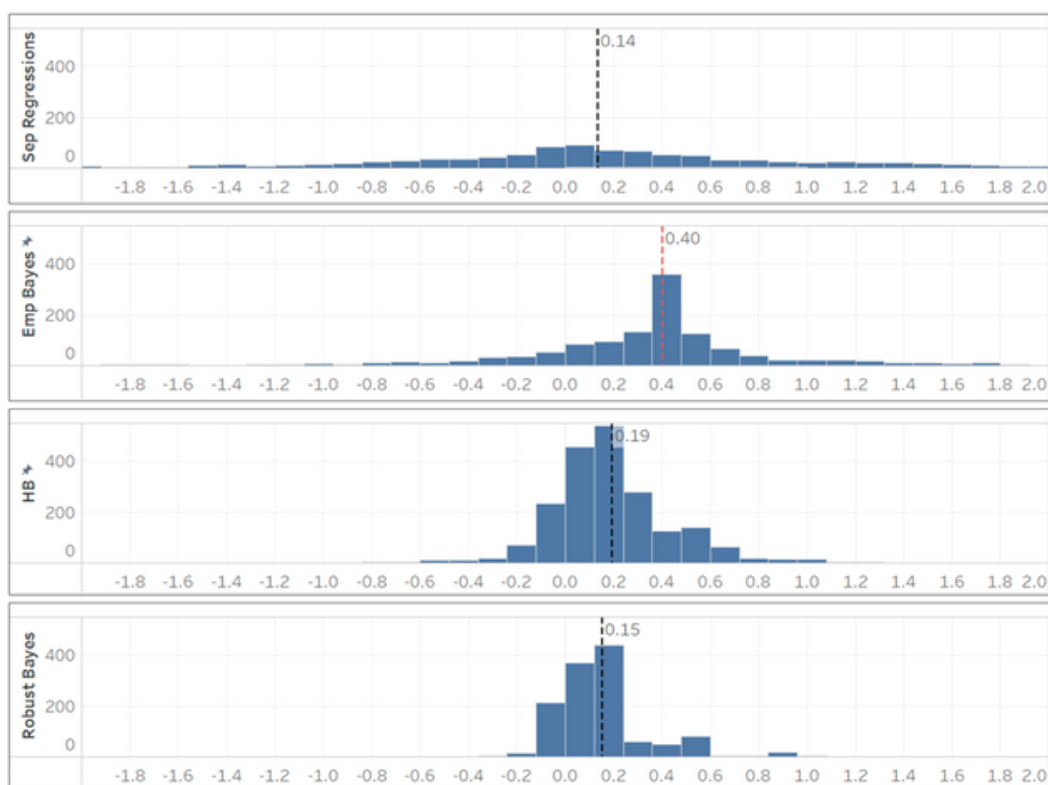


Fig. 2: Distribution of coefficients for displays using four estimation methods

At the top of Fig. 2 are the “strawman” separate OLS regressions, next down is empirical Bayes. The empirical Bayes results come from SAS PROC MIXED. Empirical Bayes methods are used widely by Nielsen, Neustar, and other suppliers. The third chart down is hierarchical Bayes and the last chart is Robust Bayes. Here are some key observations:

1. The results for the separate OLS regressions are very spread out, reflecting some real signal in the data but with volatility due to different numbers of weeks of causal and different degrees of freedom for error (df) in each. Note that there are many outliers outside of the boundary of -2 to 2 as well.
2. The Empirical Bayes estimate is higher than the other three at 0.40, which corresponds to a 50% lift in volume. The dotted vertical lines represent OLS (median), empirical Bayes (the fixed effect), and the posterior means for HB and Robust Bayes. Note that empirical Bayes does not model the variance of the prior, so the spread of the estimates is restricted to be normal (gaussian) with a large central peak and long tails.
3. The hierarchical Bayes (HB) method has a lower mean value of 0.19 which corresponds to a lift of 22% and begins to show some interesting dispersion. Fig. 3 shows there is a mass of UPCs with values near zero and fewer, larger values to the right. If you look carefully, you will notice that there are more UPCs captured in HB than in empirical Bayes (i.e., the blue bars have more mass). This is a potential problem. As the bar chart (below, left) shows, about 37% of the 2000 UPCs never had a display. So, the HB approach is using the prior to fill in values for UPCs that have never had a display.

In fact, it is important to mention that the empirical Bayes method would do this as well. The red dashed line in Fig. 2 is the *fixed effect* of the empirical Bayes model. This would be the default prediction for UPCs that had never had a display. But to be fair to empirical Bayes, we have removed these values in a manual step.

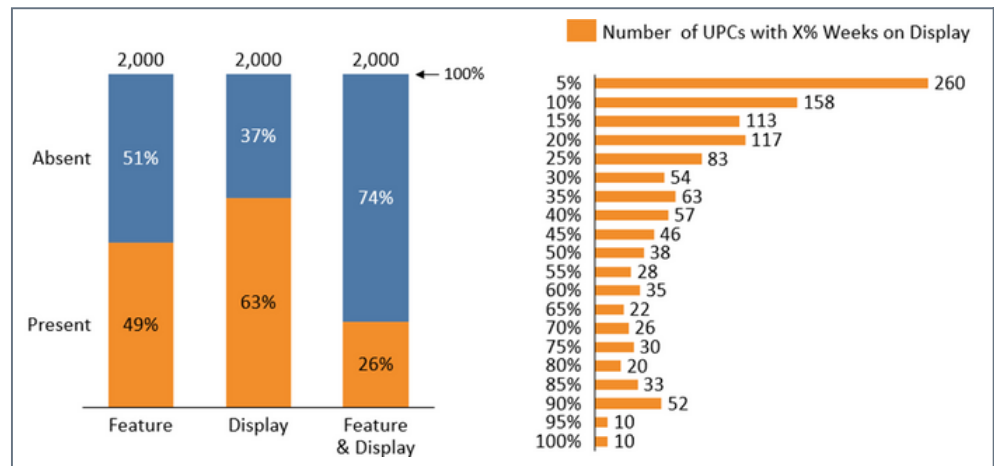
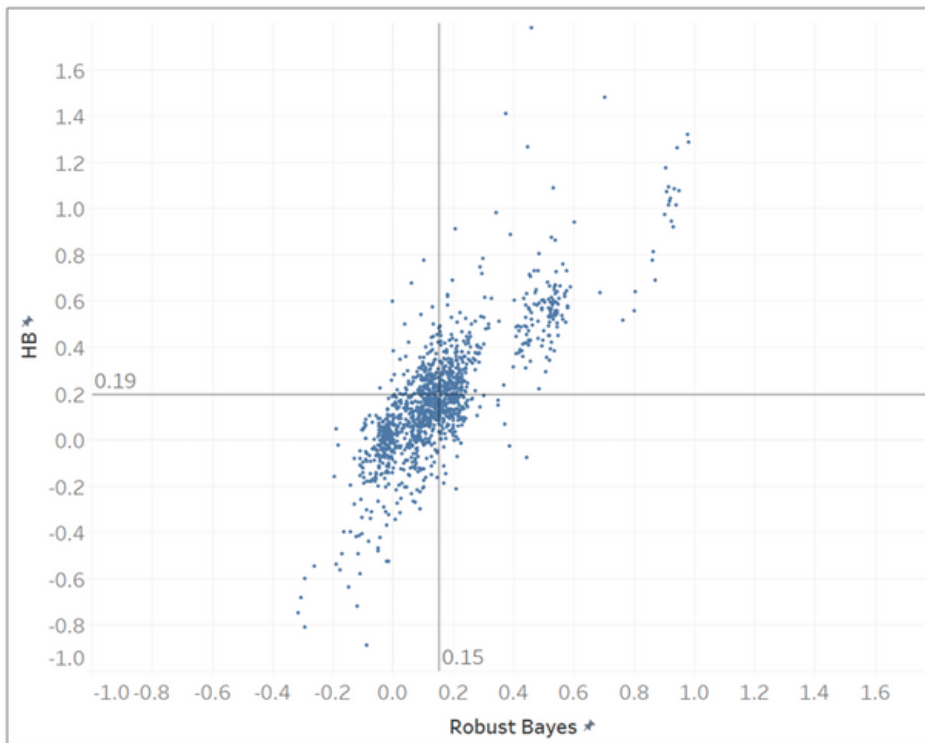


Fig. 3: Percent of UPCs out of N=2000 with each type of promotion (left); count of UPCs with various % weeks on display (right). For instance, 63% (1255) of the 2000 UPCs had a display in the 2-year time period. Of those, 260 had displays in less than or equal to 5% of weeks, 158 in 5% < x <=10% of weeks, etc. Note that in each week of display, for each UPC, the percent of stores in the chain penetrated is also a factor. Both elements (% of weeks, % of stores) contribute to the precision of the final estimate of display lift.

4. In Robust Bayes (Fig. 4), we have applied a rule that eliminates UPCs that have never had a display from both giving and receiving data from the prior. Note that we have not had to eliminate these UPCs from the overall model, just from the estimate of display lift. In Robust Bayes, the UPCs still benefit from the prior – but they have to have some “skin in the game.” As the horizontal bar chart (above, right) shows, among the UPCs most had a display in only 5% of weeks (260 UPCs) while there were 52 UPCs that were almost constantly on display with 95% of weeks promoting. For each UPC, in each week when it is displayed, there is also a % penetration of the chain (number of stores) that supported the display. Both factors drive the precision in the coefficient, and it is this precision that determines how much each UPC contributes to the prior.

When we look at the Robust Bayes chart from Fig. 2 – we start to get the most accurate picture of what is going on. We see a large number of displays that have a moderate effect (0-0.2 or about 0-20% lift) and a smaller number of UPCs in the 0.2 to 0.6 range (20% to 82% lift). One way to get even more insight is to compare hierarchical Bayes with the Robust Bayes method:



In Fig. 4, we compare the coefficients for display for the subset of UPCs that meet the minimum criterion of having any display in the 2019-2021 period. Clearing out these “duds” is the first benefit of Robust Bayes. But notice also how the precision or range of the estimates has improved for UPCs that did have a display: from -1 to 1.6 for regular HB to about -0.3 to 1 for Robust Bayes. Both methods have teased out 2-3 clusters of displays, but Robust Bayes has more precision. Note that both Bayes methods have the ability to pull out clusters of high and moderate performers – a feature lost in the cruder empirical Bayes approach.

Fig. 4: Comparison of hierarchical Bayes (HB) to Robust Bayes

As a final exercise, we will look at the distributions of coefficients for *features and displays* – when a store both advertises a UPC and has a display for it. About 63% of UPCs had a display in the period. But only 26% of UPCs had both a feature **and** a display. What happens when there’s even more missing data?

The Robust Bayes estimate is also much closer to the median of the individual regressions and this is true for all the advertising and promotion effects we have looked at. This makes sense because fully Bayes models work at the UPC level, using a sensitive prior to stabilize estimates that are far away from the central tendency. By comparing the different distributions, we see more divergence between the empirical Bayes estimate and the fully Bayes estimates.

Empirical Bayes gives a lift value of 73% versus Bayes at about 28%. Not only that, empirical Bayes gives a false sense of precision around that estimate.

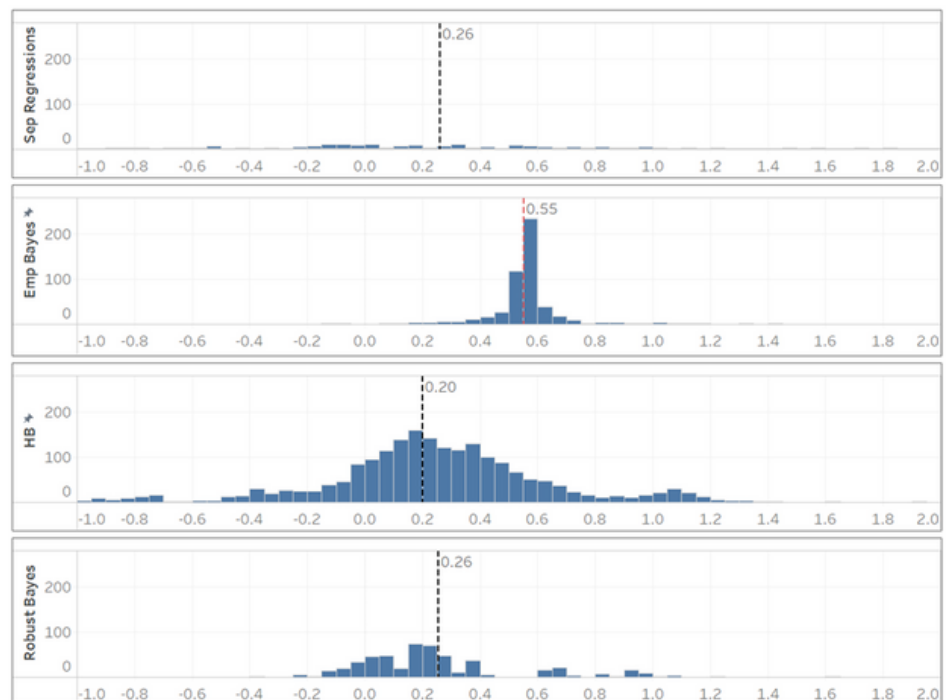


Fig. 5: Distribution of coefficients for feature and display using four estimation methods

Taking both the mean and the variance of these coefficients into account, there is a big difference between the statements that a modeler would give to their client:

Advice using Empirical Bayes (Random Effects or Mixed Models):

The lift from combining feature and display is 73% over base, and we're sure of that: about 80% of all the promotions fall between 57% lift and 82% lift. You should keep doing these.

Advice from using Robust Bayes™:

The lift from combining feature and display is 28% over base, but there's a lot of variation to that. About 36% of UPCs have very low lifts and you should consider not doing those. About 47% have performance near the average of 28%. There are, notably, 16% of promotions with lifts averaging 145%. We should look into those top performers to see what they are doing.

What is the in4mation insights advantage?

We hope we have demonstrated that fully Bayes models are superior to both empirical Bayes and to running separate regressions. We hope to also have shown that hierarchical Bayes (HB) models have a flaw when it comes to dealing with sparse variables and that the solution is **Robust Bayes™**. This is not just an "academic" level of difference; the completely different nature of key takeaway proves that. Is this something you should be concerned about? We think so.

You may ask "what makes you think that the Robust Bayes estimate is *right*, and the empirical Bayes estimate is *wrong*?" First, let's look at some basic fit statistics to determine how close the predicted values come to the actuals:

	Robust Bayes™	Empirical Bayes
R-squared (sample)	96.0%	96.2%
R-squared (holdout)	98.5%	94.5%
MAPE (sample)	22.7%	24.2%
MAPE (10% holdout)	28.7%	33.2%

Both models have a very high and similar R-squared. The fully Bayes model performs better on the MAPE¹. The fully Bayes model has an average of 22.7% error at the UPC/weekly level versus the empirical Bayes model at 24.2%. The fully Bayes model shows an even better improvement in the holdout, the key measure of predictive validity. These performance scores alone don't disqualify empirical Bayes. In fact, many MMM models can reach similar measures of fit with very different specifications. So, it is always important to ask *whether the size of the effects is reasonable*. It is here where we have our doubts about empirical Bayes. Even though the individual regression approach was crude and prone to outliers, there was still enough data to do them. Samples sizes for most UPCs were over 52 weeks. The *median* value of these effects for feature and display, which protects against outliers, is identical to fully Bayes (0.26) while the empirical Bayes estimate is far away from both (0.55). It is both the flexibility of the prior in fully Bayes models and the fact that fully Bayes models are *built up* from the granular level and regulated by priors versus having a distribution imposed on them from the *top down* like empirical Bayes. Because they are built up from the bottom, the uniqueness of exceptional performers is also retained in fully Bayes models, which means when we find an exceptional UPC or an exceptional chain or DMA, etc. it's likely to be true versus an artifact of the method and worth following up.

As the marketing world of digital media and digital shoppers becomes more complex and fragmented, there will be more incidences of sparse variables. For instance, with retailer media networks, executions are being mapped down to the UPC/retailer level. This generates variables that start to have the granularity of the feature and display variables that we have discussed in this paper. The behavior of regression models is very different with intermittent media and promotion variables than with continuous streams like adstocked linear TV. We can't expect MMM models built in the 1980's and 1990's around linear TV to continue to perform well when the nature of the media inputs is changing. As much as the MMM industry may have tried to "bury the math", it is perhaps time to open up the hood.

¹ MAPE is the Mean Absolute Percent Error defined as $MAPE = \frac{1}{n} \sum_{i=1}^n ABS(observed - predicted)/observed$ where i is a UPC/week.

References:

1. G. Casella (1985), [An introduction to empirical Bayes data analysis](#), *Amer. Statistician*, vol. 39, no. 2, 83–87.
2. P. Rossi, G.M. Allenby, and Robert McCullough (2005), [Bayesian Statistics and Marketing](#), *John Wiley & Sons Ltd.*, Chapter 3, Section 7.

About the Authors



Mark Garratt
Partner & Co-Founder
in4mation insights

Mark Garratt is a partner and co-founder of in4mation insights. He is an accomplished analytics professional with a distinguished career in both business and academia. Mark has been a trusted advisor to some of the world's biggest brands and an analytics leader at CPG companies including P&G, SABMiller Brewing, and The Gillette Company.



Sandeep Conoor
Senior Director,
Marketing Science, R&D
in4mation insights

Sandeep Conoor is a Senior Director, Marketing Science in R&D at in4mation insights. He plays a leading role in several challenging projects. These include prediction of customer activity at a large e-commerce website using Hidden Markov models, analysis of customer visit frequency and marketing mix for national restaurant chains, and building a new product sales forecasting system.

Contact Us

If you would like to discuss how Robust Bayes™ can make a difference in optimizing your unique media and marketing mix, reach out to us for a conversation at info@in4ins.com.