Taming Noisy Data Through Bayesian Analysis - A Market Share Tracking Example

Brent Johnson, Ph.D.

IPR's data science group is seeing an increase in companies trying to gain insight from inexpensive, streaming data sources. Due in part to growth in the Internet of Things, greater numbers of connected devices are generating larger quantities of data, which clients are trying to mine for quick, inexpensive insights. This is a challenge because such data are often noisy or biased. These data are, in a sense, data of **convenience**. They're analyzed not because they're the most accurate or most representative, but because they're cheap and available. Unlike an original, intentional data collection endeavor, there's often little a company can do to entirely eliminate the bias or representativeness of such data, its quality, the sample sizes or coverage, or the data cleanliness. The data simply are what they are. Take market share tracking, for example. Whereas companies once engaged (and many still do engage) in expensive, high quality, regularly recurring customer surveys to measure their own (and competitor) market share, the growth of data generated by internet-connected devices and other passive vehicles for data collection provides an alternative. If one can find both the right data stream that captures indicators of the sales or use of one's product, then one can generate powerful, real-time market share insights. But one needs the right tools to extract the signal from the noisy data. I've found great success in using Bayesian analysis to make greater sense of such dirty, or even downright polluted, data.

Why Bayesian estimates?

One of the first examples that convinced me of the power of Bayesian analysis for making sense of noisy data was its use in predicting baseball statistics[1]. I know that sounds like a trivial use case. What do sports statistics have to do with workplace needs such as measuring customer propensity to buy? However, if you think about it, it's not an entirely irrelevant parallel. In baseball, fans are often interested in batting averages[2]. A player's batting average is the number of times the player gets a hit divided by the number of times that player was at bat. It's a share metric-albeit it's the share of times a player obtains a hit as opposed the share of times one's product was purchased. Each time at bat, a baseball player either gets a hit or strikes out; a customer either buys one's product or doesn't buy.

Fans of both baseball and Bayesian analytics have shown how noisy, small-sample, early-season Bayesian estimates of batting averages actually turn out to be better predictors of end-of-season batting performance than the original early season batting averages themselves. Such Bayesian estimates[3] lessen the amount of noise or error in the estimates. Those estimates that were unstable due to their small samples (e.g., few batting occasions) are strengthened by learning from the larger, collective sample composed of other players.

A motivating marketing case study

In this case study, I demonstrate how Bayesian estimates can turn a noisy data stream, some of it coming from very small sample sizes, into an insightful marketing story. I demonstrate how to turn data dross into data gold, and I do so using a real-world data example. The data here are perturbed and disguised so as to preserve client confidentiality, but the value of Bayesian estimation is clearly demonstrated.

[1] https://www.jstor.org/stable/2287098
[2] https://en.wikipedia.org/wiki/Batting_average
[3] http://sas-and-r.blogspot.com/2012/04/example-927-baseball-and-shrinkage.html

The data for this case study come from a tracking service that passively collects data on company server configurations around the world. Embedded in these data are metrics revealing the presence of our client's server software product (and that of its competitors) on servers inside hundreds of thousands of business establishments. (Think Internet of Things in the commercial market space). When these servers are connected to the Internet, they automatically communicate with a central location (hosted by the tracking service) and report the server's health. Embedded in that report are details on the server's software configuration.

This tracking service receives status reports from millions of servers around the world. My interest is in a sample of these data–a snapshot of the configuration of the world's server population at quarterly time intervals. From these snapshots one can derive the share of servers (or business establishments) containing a client's particular software product vs. that of its competitors. Each quarterly data snapshot contains roughly 4 million records describing 5 million servers. And over all time periods, the data set (due to its width and additional variables) is nearly 100GB in size and warrants the label, "Big Data." The tracking service is relatively inexpensive and has the potential to provide a near real-time measure of market share and product performance.

The challenge in this effort, however, is the noisiness of the data. The client wanted separate market share estimates for 28 geographies, 4 company-size market segments, and for hundreds of product configurations. However, not all geographies, segments or product configurations were covered equally in the sample. There was no sound research design behind the data collection. The data isn't balanced and data fields are plagued by missing values. To give the reader a feel for the data, Table 1 below shows the most heavily sampled geographies and segments. For example, the most well-covered group is medium-sized businesses in the United States where I have records from over 60,000 establishments and over 100,000 servers.[4]

Segment	Study.Geography	Number of Establishments	Number of Servers
Medium establishments	US	60,308	116,795
Small establishments	US	43,067	56,374
Public sector	US	29,581	66,267
Medium establishments	UK	11,723	15,073

Table 1: The geographies and segments with the LARGEST sample sizes, 2017Q2.

In contrast to the above large-sample geographies, some geographies and segments had far smaller sample sizes. Below in Table 2 I show some of the geographies and segments with the least available data. At the top of this table is Japan large business where the data contained a mere 5 establishments in the sample and none of them possessed the type of servers capable of using the client's software. These are sparse data indeed! But the sparseness provides an ideal demonstration of the power of Bayesian modeling.

Segment	Study.Geography	Number of Establishments	Number of Servers
Large establishments	Japan	5	0
Public sector	Korea	5	1,343
Large establishments	Denmark	12	515
Public sector	Japan	17	16,347
Large establishments	Poland	18	477
Large establishments	Switzerland	19	290
Large establishments	Turkey	20	479
Large establishments	China	25	11,836

[4]These establishments owned far more servers in total than those shown here. The server count in Table 1 includes only the type eligible for the client's software).

Table 2: The geographies and segments with the SMALLEST sample sizes, 2017Q2.

I note one important limitation in these data: the sample sizes are small and subject to biases. The observations in this research do not reflect a random sample. Companies install this particular software and report back only on an opt-in basis. It's possible that companies left out of the sample are different from those that are in it. And in other research (not described here) the client discovered a bias in the market shares for one certain product category. In that research, the client and I took steps to correct for the bias, including the incorporation of additional data sources. However, for the present product category and analysis we did not find such a bias. For this product category, the missingness didn't appear systematic or correlated with any critical variable, nor were the shares at odds with other research. In fact, it appears that the the small sample sizes are due to the very recent launch of the tracking service in certain (mostly Asian) geographies and its use isn't correlated with the choice of this particular product category.

I share this background on sample sizes and because it highlights the power of Bayesian methods. Small sample sizes make for unstable market share estimates, estimates that can fluctuate wildly across segments, geographies, and time. One can see this in the following graph. Figure 1 shows the raw market shares prior to any Bayesian modeling. These are the maximum likelihood estimates (MLE). Presented this way to the client, each quarterly release of survey data caused a considerable reaction. Intense discussion ensued when, in one quarter, market share declined for a given product in a given geography. The next quarter, however, such angst often proved unwarranted when the subsequent results erased the decline and suggested the temporary drop was a mere small-sample artifact. This instability and over-reaction was particularly problematic because bonus compensation inside the client organization was based, in part, on reaching pre-set market share goals. Compensation of geography managers was at risk of being erroneously calculated (both positively and negatively) by such unstable results.

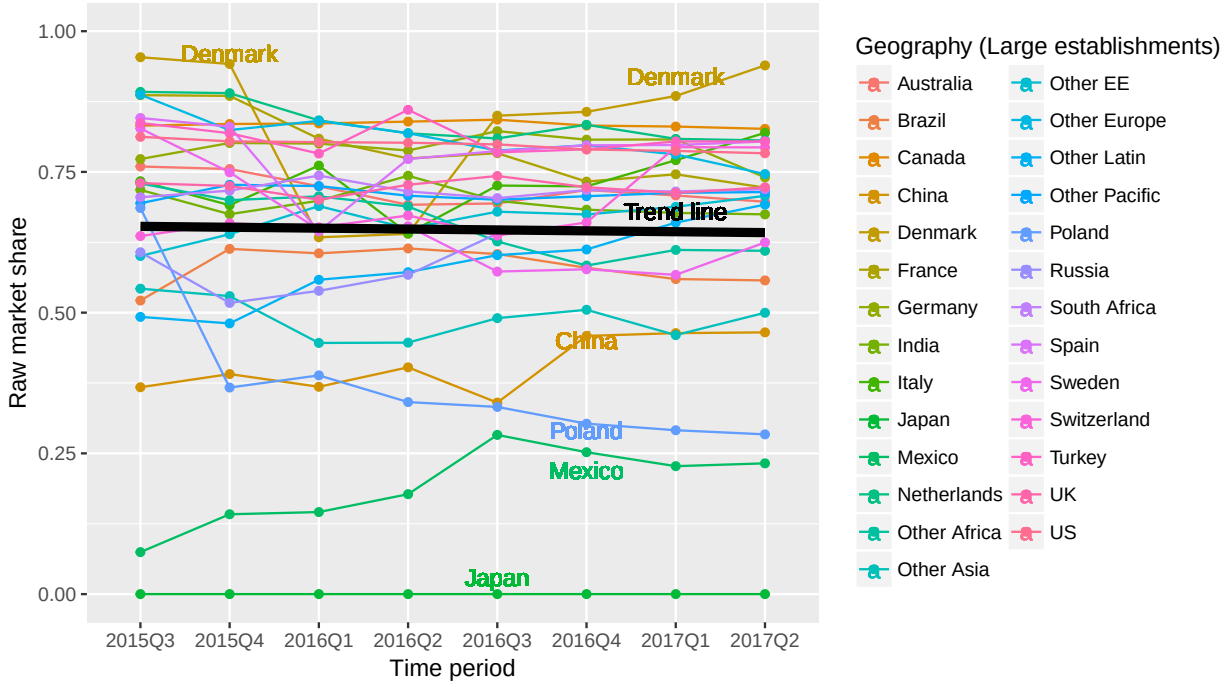


Figure 1: Raw market shares over time for large business establishments. The trend line is the unweighted average share across all geographies.

The raw market shares as shown in Figure 1 generated the following client questions:

- On average, is our market share truly declining over time? How significantly? In what geographies is the decline most severe?
- Is it true that our market share took a dip in Denmark? Then what caused Denmark to recover and become the highest share geography in the world?

- What's our market share in Japan? Surely our market share isn't truly zero.
- Why did our market share drop in China in Q3'2016?
- Is Mexico really bucking the trend with a rising market share? What are we doing correctly in that particular region?

The solution formula

I proposed a Bayesian solution to this client's problem. The solution reduces wave-on-wave volatility. And based on a cross-validation exercise, I will demonstrate that the Bayesian estimates are also significantly more accurate estimators than the original, raw market shares estimates.

I begin the Bayesian model description with the following equations where Y_{tij} is the count of servers containing the client's software product of interest in the the t^{th} time period, i^{th} geography, and j^{th}segment. It's modeled as a binomial outcome with probability p_{tij} based on n_{tij} total observations. The binomial distribution[5] is the appropriate probability distribution for modeling the number of independent yes/no-type events–such as whether a batter achieves a hit (in baseball) or a server contains the client's software (in marketing).

$$
\begin{aligned}
Y_{tij} &\sim binomial(n_{tij}, p_{tij}) \\
p_{tij} &= logistic(\mu_{tij}) \\
\mu_{tij} &= \alpha_0 + \alpha_1 * time_t + \beta_i * geography_i + \gamma_j * segment_j + \delta_{tij}
\end{aligned}
$$

I furthermore model the probability (p_{tij}) that a server contains the client's software as a logistically transformed function. In this function I first consider the time element. I have 8 time periods which I model using a simple linear trend. After all, clients often want to know, "Is my market share growing over time?" I can easily determine this growth by examining the α_1 coefficient in the trend equation above. I consider the intercept (α_0) and trend (α_1) parts of the model as "fixed effects." I next model geography and segment effects which are captured by β_i and γ_j, respectively. Recall that there are 28 geographies and 4 segments, so, this model will generate 28 different β parameters and 4 different γ parameters. The last element in the μ_{tij} equation above is an error term δ_{tij}. I model a unique δ_{tij} parameter for each time period, geography, and segment combination in the data.

In contrast to the trend parts of the model, I call the β_i, γ_j, and δ_{tij} parameters "random effects" and I model them with the following priors:

$$
\begin{aligned}
\beta_i &\sim normal(mean = 0,\ variance = .2) \\
\gamma_j &\sim normal(mean = 0,\ variance = .2) \\
\delta_{tij} &\sim normal(mean = 0,\ variance = .2)
\end{aligned}
$$

This means that I'm treating these particular geographies and segments probabilistically, or as just one out of many possible manifestations or ways of dividing the world and software market place. Notice too that I'm modeling these parameters as deviations (centered at zero and with a prior variance of .2) from the average market share at $time_t$. This centering helps with model identification. It is this random effects part of my model that's going to reduce noise and increase the accuracy of my market share estimates.

Readers knowledgeable of Bayesian statistics might recognize the equations above as similar to those of Clayton and Breslow (1993)[6], who first proposed such a random effects error term. The present model, however, is different from this in that I put a "regularization prior" on δ_{tij} and all the other above random effects. That is, I assume an informative prior variance (.2) rather than a random variance and hyperprior. I do this to further tame the noisy data and increase the accuracy of my share estimates.

There's now just one more important feature of this model–one that makes it even more unique. As described above, my use of the binomial distribution assumes that I'm modeling Y_{tij} independent server software installations. But that's not actually the case in practice. In practice, independent software purchase decisions

[5] http://stattrek.com/probability-distributions/binomial.aspx
[6] https://www.jstor.org/stable/2290687

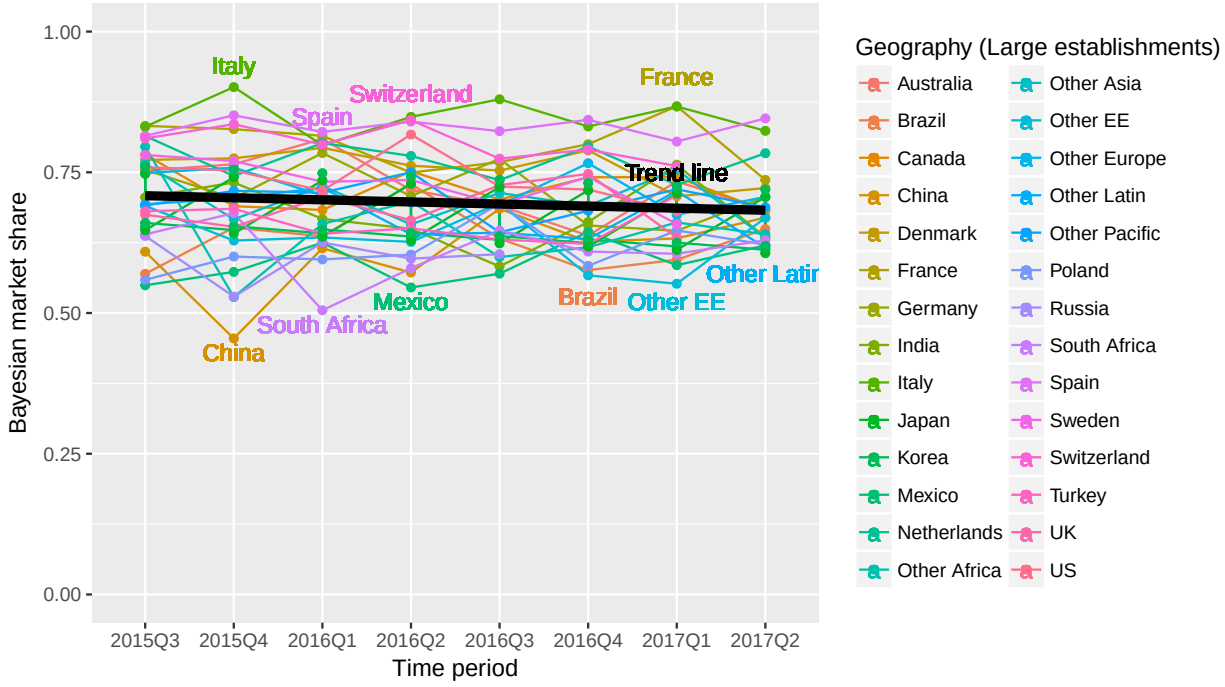
are not made for each server in the sample. Rather, IT managers make purchase decisions and often roll-out new server software *en masse*. IT managers tend to push new server software onto all (or nearly all) of the servers within their company or department collectively. Put statistically, it would be overly optimistic for me to believe that the large number of n_{tij} servers directly reflects my confidence in the data. To account for this, I need to redefine n_{tij} and Y_{tij} to be proportional to the number of decision-making units or the count of companies or establishments in the sample. Using China large business as an example (the last geography listed in Table 2), I employ a base of 25 observations for n_{tij} rather than 11,836 and I adjust Y_{tij} proportionately. This reflects the idea that the observed number of servers found in the sample is governed by a much smaller number of establishments. I'm now ready to fit a Bayesian model of server software market share.

Implementing the solution

For convenience, I fit the Bayesian model to summary market share data. For each time period, these summary data contain one record for each of 28 geographies and 4 market segments (i.e., up to 112 total records per time period). For the client's research goals and equations above, analyzing the summary version of the data does not result in any loss of precision, but it does make the model a lot easier to fit!

I fit this Bayesian model using Stan[7]–or more specifically the Stan package available for the R statistical language called RStan[8]. Stan uses a Monte Carlo sampler to estimate the model parameters. That is, it draws thousands of samples from the above distributions and equations in order to generate parameter estimates. And with an appropriately structured model, I can use those thousands of samples to compute and report-out Bayesian estimates of server software market share.

I won't bore the reader with all the model statistics and diagnostics available from RStan. I'll simply show the results and market share estimates of greatest interest to the client. Here below in Figure 2 is a plot of the Bayesian market share estimates over time in contrast to the earlier plot above (Figure 1). Again, these are the market shares over time, for the server product of interest, for each geography in the large business segment.



[7]http://mc-stan.org/
[8]http://mc-stan.org/users/interfaces/rstan.html

Figure 2: Market shares over time for large business establishments AFTER fitting with a Bayesian model. The trend line is the unweighted average share across all geographies.

Wow! Compared to the original chart in Figure 1 my Bayesian market share estimates in Figure 2 are now much more compressed. The shares have a smaller variance; the lines (one for each geography) are contained within a tighter range. Also, the lines are somewhat smoother over time. This is all the result of the Bayesian model. The market share estimates that emerge from a random effects model are different compared to the raw market shares. In the data science world, we say that the Bayesian or random effects model is a "shrinkage" estimator," as in each geography's Bayesian market share estimate gets "shrunken" towards the world-wide mean market share.

This shrinkage does not mean that all the market shares from all the geographies are forced to compress uniformly. The tool applies greater shrinkage to those share estimates in which I'm the least confident (as revealed by their variance) and that are farthest away from the mean. In the logistic model the variance (and confidence) in the observed data is governed by the sample size or n_{tij}. Market shares from geographies or segments with a small sample size are less trustworthy and, so, get shrunken more. By contrast, geographies such as the US with sample sizes numbering of tens of thousands of establishments are hardly changed at all when modeled using random effects. In this way, the weaker geographies borrow strength from the estimates coming from stronger geographies and the stronger geographies stand little changed.

With these Bayesian estimates I can now provide more definitive answers to the client's questions raised above after looking at the original data:

- Our client's WW market share decline is confirmed. After reducing uncertainty in the data, the decline is a little clearer and statistically significant. The slope coefficient (α_1) in the model is -.023 (std err = .004). Over the 8-time periods covered by the data, the client lost over 3 market share points.
- I also have rough estimates for Japan's share. It's among the many indiscernible geography lines near the trend line in Figure 2. To be clear, the Bayesian method hasn't miraculously discovered new data for Japan. Rather, we treat Japan–which didn't have a sample size big enough to register any servers containing our client's product–as full of missing values. Estimates for Japan's missing values are then derived based upon an assumption about Japan's missing values and what we know about Japan vis-a-vis the other geographies and what we know about the large business segment vis-a-vis the other segments. In this manner, share values for Japan are obtained through imputation.
- Poland and China no longer stand-out as consistently low-share geographies. It turns out that their sample sizes are relatively small and their values that set off alarm bells earlier are better thought of as outliers. The corrected share values for Poland and China fall more in-line with the trend (other than China's value in 2015Q4 which warrants attention).
- For the same reason, Denmark no longer stands out as the highest-share geography. And its share didn't truly dip over the middle duration of the study. Its share is now in the middle of the pack.
- I note that the market share for the US (not called out in Figure 2) changed very little as a result of the Bayesian estimation due to its extremely large sample size (and confidence). In fact, there's less than half a point difference between its raw and Bayesian estimate.
- There's a fairly consistent pattern in that the top-share geographies tend to be in Western Europe which includes Italy, Spain, France, Switzerland, and other Western European geographies not specifically called out in Figure 2. This is a noteworthy pattern.
- By contrast, the Latin American geographies–Mexico, Brazil, and Other Latin–dominate the bottom portion of the chart. This too is noteworthy as the consistency can be explained by the fact that marketing for this region is centralized by the client. This discovery in Latin America deserves more client attention.

When data are noisy like this, I am much more confident in Bayesian model estimates than the original, raw share estimates and so was the client. The client was pleased with the analysis and the display in Figure 2. The client could now evaluate its market share performance using an easy-to-acquire, quick-to-update data feed from a server tracking service–and a data stream purged of noise and containing a clearer signal. The client didn't need to conduct an expensive or long, drawn-out primary research survey to measure its market share. The client saved both money and time.

Gaining client confidence in Bayesian model results

In this case, it wasn't easy to convince everyone within the client organization that the Bayesian method was truly superior to simply reporting the raw market shares. A valid question is whether the estimates above are just window dressing. Sure, the estimates in Figure 2 are now much smoother than in Figure 1, but does that mean they're really better? Are they truly more accurate? When shrinking the estimates did I eliminate the signal as well as the noise? Could I have possibly over-fit the data and moved the market share estimates further away from their true values?

It can indeed be a challenge to explain why the simple raw share measure (which is the maximum likelihood estimator or MLE) isn't always the best predictor of a central tendency. It's hard to mathematically prove this to those without a strong statistical or mathematical background. I include the model equations in this white paper simply to satisfy the more technical reader, but there are few clients with a stomach for such detail. Clients, however, need a clear, tangible demonstration that Bayesian estimates are superior.

I've found that client confidence in Bayesian model results is more easily achieved not through equations or a lot of statistical jargon, but through the right example demonstration. Clients can be convinced of the superiority of Bayesian methods by testing their increased accuracy in practice. I do this through cross validation.

Evaluating the Bayesian method

Cross validation is a tool for assessing whether the results of a statistical analysis will accurately generalize to an independent data set. To conduct a cross validation in its simplest form, one first randomly splits one's data into training and validation samples. Then, one computes a model or the statistics of interest using the training data and observes how well it predicts data from the validation sample which is treated as a hold-out. My statistics of interest are the Bayesian vs. the raw (or MLE) market share estimates. I want to see which estimates better predict the actual market shares in the validation sample.

To conduct the cross validation, I take the individual data records from one example time period, (2016Q3 which is the midpoint of the study period). The data collected in this quarter includes records on over 4 million servers. I then draw a stratified random sample (stratified by geography and segment) such that 50% of the companies within each geography and segment are assigned to the training sample and the other 50% are assigned to the validation sample. The stratification ensures that small-sample geographies and segments, such as those shown in Table 2, are evenly assigned to the training and validation samples.

For the next step I start with the training data and compute the raw (maximum likelihood) product market shares for each geography and segment combination. I summarize the number of servers containing the client's software (Y_{ij}) and also the total number of servers (n_{ij}) for each geography-segment and express that as a share, Y_{ij}/n_{ij}, which is the maximum likelihood estimate. This results in 105 summary share observations.[9]

Still using the training data, I next compute the Bayesian share estimates. To do this I adopt nearly the same Bayesian model equation described on page 4. The only exception is that I'm now working with data from just one single time period. And since there's only one time period I don't need a slope coefficient (α_1) in my model. Instead, I can model the Bayesian estimates (μ_{ij}) as follows:

$$\mu_{ij} = \alpha_0 + \beta_i * geography_i + \gamma_j * segment_j + \delta_{ij}$$

The model runs faster than before, mostly because I'm now fitting it to just one time period. The fit is good and the performance of my Monte Carlo sampler is sound.

Figure 3 below compares the simple, maximum likelihood share estimates (left axis) to the Bayesian share estimates (right axis) for the training sample. There are 105 lines in the Figure–one for each segment-geography

[9]There are a total of 112 (28 x 4) possible geography-segment combination in the data. But in spite of the stratified random sampling I had to omit 7 geography-segment combinations which had missing values either before or after splitting the sample, such as Japan large business. For these combinations I can compute Bayesian share estimates but not the maximum likelihood market shares thus preventing a comparison. The resulting training and validation samples shared 105 geography-segment combinations in common.

for which I have complete data. In general, flatter lines indicate little difference between the Bayesian and maximum likelihood share estimates in the training sample. The relatively flat lines come from geographies or segments in which I have greatest confidence (higher sample sizes) or that are closer to the grand mean. By contrast, the non-flat lines reflect geographies and segments that have smaller sample sizes and for which the maximum likelihood estimates are likely to be less stable. The Bayesian estimates shrink the less stable shares towards an overall mean. This Bayesian model is doing exactly what I would expect.

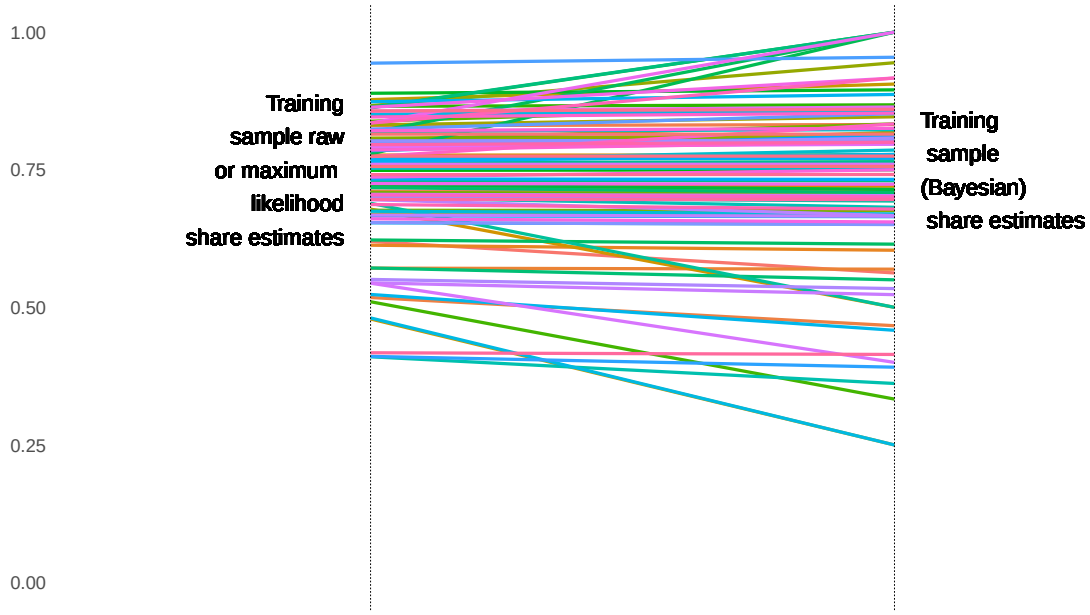


Figure 3: A comparison between the original, maximum likelihood share estimates computed from the training sample (left axis) to the Bayesian market share estimates from the training sample (right axis). There is one colored line in the figure for each unique geography and segment combination in the data (or 105 lines total). The Bayesian share estimates (left axis) show greater shrinkage and less variance.

With raw or maximum likelihood market shares and Bayesian market shares both computed from the training sample I am now ready for the moment of truth: a comparison to determine which share estimator is a better predictor of the hold-out market share estimates from the validation sample. I compute these market share estimates for the validation sample, $(Y_{ij})/(n_{ij})$, and I compare how well the two training sample estimates predict the validation shares. I do this for a number of performance metrics displayed in the first two rows of Table 3.

This comparison is perhaps sufficient for judging the performance of the Bayesian estimates. However, to confirm there's nothing unique about this particular training (or validation) sample, I can go one step further in my comparison and also do the reverse. I can switch direction and compute the raw, maximum likelihood share estimates and the Bayesian share estimates from the validation sample and then use those to predict the shares from the training sample, treating the training sample as my new hold-out. This provides a second test and comparison for how well the two share methods predict. To create these estimates from the validation sample I follow the same process that I just did for the training sample. That is, I apply the same Bayesian model (without the trend component) but now fit to the data from the validation sample. Rows 3 and 4 of Table 3 display how well the Bayesian and maximum likelihood estimates from the validation sample predict the hold-out market shares from the training sample.

Bayesian model evaluation results

Again, the purpose of this cross-validation is to evaluate whether the Bayesian share estimates do a better job of predicting the hold-out market shares than do the raw (maximum likelihood) share estimates. I can do this for the training sample's estimates by comparing rows 1 and 2 of Table 3. And I can evaluate the validation sample's estimates by comparing rows 3 and 4. I compare the two market share estimation methods using a variety of performance metrics.

Statistic	Holdout comparison	Mean absolute pct error (MAPE)	Mean absolute error (MAE)	Maximum share difference	# of share differences $> 50\%$	# of estimates closer to the holdout shares	Correlation with bench-mark shares
Training sample Bayesian estimates	**Validation sample** market shares	0.215	0.112	0.456	0	66	0.458
Training sample raw estimates	**Validation sample** market shares	0.22	0.123	0.667	2	39	0.408
Validation sample Bayesian estimates	**Training sample** market shares	0.171	0.11	0.448	0	71	0.443
Validation sample raw estimates	**Training sample** market shares	0.186	0.123	0.667	2	34	0.408

Table 3: Performance comparison of the Bayesian vs. raw (or maximum likelihood) share estimates. Rows 1 and 2 compare how well the two estimates predict market shares from the validation sample. Rows 3 and 4 compare how well the two estimates predict market shares from the training sample. On all metrics and for both hold-out samples the Bayesian estimates (rows 1 and 3) are better predictors than their corresponding raw (maximum likelihood) share estimates (in rows 2 and 4, respectively).

By all quantitative measures in Table 3, the Bayesian estimates do a better job of predicting the market shares in the two respective holdout samples. The Bayesian estimates computed from the training sample (row #1) do a better job than the raw, maximum likelihood estimates (row #2) at predicting validation sample's hold-out market shares. Likewise, the Bayesian estimates computed from the validation sample (row #3) do a better job than the raw, maximum likelihood estimates row #4) at predicting when the training sample is the hold-out market shares. I now look at some of these metrics in more detail.

- I first look at the "Mean absolute percentage error" or MAPE. Smaller MAPE values indicate less error and better prediction. This value for the training sample is 21.5% and 22% for the Bayesian and raw share estimates, respectively. The two corresponding figures for the validation sample are 17.1% and 18.6%. For both samples, the percent error in the Bayesian estimates is lower–albeit to a lesser degree for the training sample.

- Next, the "Mean absolute error" or MAE shows the average difference in share points for the two predictions vs. the holdouts. Compared to the MAPE, the mean absolute error measure is less sensitive to small share denominations. And on this measure, the two Bayesian estimates are lower than their raw share counterparts by roughly one full share point, on average (.112 vs. .123 and .11 vs. .123). Again, the Bayesian estimates are closer to the desired predictions in the hold-out samples.
- The "Maximum share difference" column highlights the biggest share outlier in each group. The worst performing Bayesian prediction from the training sample was off by 45.6 share points and the worst performing Bayesian prediction from the validation sample was off by 44.8 share points. This compares to the worst performing raw share estimates of 66.7 share points (for both the training and validation samples). This indicates two things. First, it tells me that accurate share prediction in every case is difficult for either method. Second, as high as these errors are, the Bayesian method is at least the least wrong.
- The fourth numeric column in Table 3 shows one of the most interesting quantitative metrics–the "Number of share differences greater than 50%". It's a count of extreme outliers. It's the number of share estimates that differ from the holdout estimates they're attempting to predict by more than 50 share points. The Bayesian estimates have zero such egregious predictions; the raw share estimates have two such cases in both the training and validation samples. This statistic reveals how well the Bayesian method performs in terms of reducing grossly wrong market share conclusions. To summarize, the Bayesian estimates, although imperfect, do indeed reduce the frequency of horribly wrong market share estimates. This in turn reduces the risk of over-reaction or bad marketing decision-making in response.
- In addition to the above metric for counting outliers, the overall "Number of estimates closer to the holdout shares" (column 7) is higher for the Bayesian estimates. For example, out of the 105 total geography-segment combinations that I'm trying to predict, 66 of Bayesian estimates from the training sample were closer to the holdout shares while only 39 of the raw share estimates were closer. (The corresponding numbers for the validation sample are 71 and 34, respectively). The Bayesian share estimates are more often closer to the hold-out shares than the raw, maximum likelihood share estimates.
- And finally, the correlation of the estimates with the hold-out share values is higher for the Bayesian estimates (.458 and .443 for the training and validation samples, respectively) than for their competing raw maximum likelihood counterparts (.408 and .408). The Bayesian estimates provide a better correlation measure.[10]

Examined on a number of metrics, the Bayesian method is indeed the more accurate method for predicting market share estimates from these sometimes noisy data. Cross-validation comparisons like this provide tangible evidence and are quite convincing for clients. Clients can more easily make the switch to alternative estimation methods when such methods clearly result in greater accuracy and fewer egregious prediction errors. In fact, it was this type of predictive performance, as demonstrated using sports statistics, that turned me on to such Bayesian estimates in the first place.

I should point out that the Bayesian estimates were not the most accurate predictors for every geography and segment in this cross-validation experiment. Column 7 of Table 3 tells us that there are 39 cases (geography-segment combinations) where the traditional market share estimates from the training sample were actually closer than the Bayesian share estimates. After hold-out tests such as this, I've heard clients criticize Bayesian estimates over this fact and ask questions such as, "Why did the Bayesian estimate not accurately predict all 105 market shares?" Or clients have asked me to mix-and-match the estimation methods. That is, clients have asked me to use the Bayesian method only for those geographies or segments where one can demonstrate they out-predict the raw shares and they've asked me to use the raw share calculations otherwise.

I resolve such concerns as follows. First, it's almost always the case that the list of geographies better predicted using the raw, maximum likelihood share estimates will change wave-to-wave. If I repeated the above cross-validation using data from another time period, I would find a different short-list of geographies better predicted by the raw shares. Hence, any decision rule regarding which geographies to report using

[10]The observant reader might notice in Table 3 that the values in rows 3 and 4 for the correlations (column 8), the maximum share differences (column 5), and the number of share differences >50% (column 6) are identical. This is not a coincidence. The raw estimated market shares for a sample are the same whether they're calculated for use as predictors or for use as holdout comparison points.

non-Bayesian estimates won't be reliable. Any such rules are likely to be ad hoc. Similarly, even a different randomization of the data in this particular cross-validation time period will result in a slightly different set of rules for which geography-segments are better predicted by the Bayesian method. However, on average and over all time periods, the Bayesian methods will more consistently out-perform.

Second, keep in mind that all the data over all time periods in this project was susceptible to error. There are no "true," known market shares available to provide us with an unequivocal comparison point. Even in this particular cross-validation time period, the hold-out sample shares were often computed from very small samples. (The sample size for Korean medium-sizes businesses in the validation sample, for example, contained just 3 establishments!) Whatever shares one uses for a hold-out comparison point and performance criterion are at best only an approximation of the truth. Such shares are sufficient for an overall performance comparison, but are not reliable enough to judge exactly which geography-segment combinations might be best predicted using only raw market shares. The safer rule is to simply adopt the Bayesian estimates altogether and stick with them.

One potential criticism of this Bayesian method is the existence of possible structural differences among the analytic groups. The present project and Bayesian model included three factors: geography, segment, and time. If there are known structural differences in the ways different geographies or segments use or purchase this particular server software product, then the researcher should attempt to account for this in the model. In our case, for example, it's possible that large enterprise businesses adopt new server software more slowly than the other size groups because such large businesses have more servers to migrate. The risks, effort, and the expense may be greater for large businesses than it is for smaller ones. If it were truly the case that this influences market shares, then one could improve the estimates by expanding one's Bayesian model to condition on such differences. One might not want to shrink the shares from large establishments to the same grand mean or to the same degree as the other segments. However, such group differences are easily checked and, if they do exist, are easily accommodated in one's model. In the present cross-validation there were no such stand-out differences among business size groups. For instance, the number of cases where the Bayesian model's estimates were less accurate (Column 7 of Table 3) did not contain a noticeable excess of any one business size and the model equations used above are entirely sufficient.[11]

Extending the application

Bayesian estimation is a powerful weapon in the data scientist's arsenal. Here I've demonstrated how Bayesian estimation tames noisy data, reduces outliers, and improves marketing decision making. Bayesian estimates are very often the better, more reliable estimates when one needs to make predictions for different groups or entities measured using imperfect data sources.

One can employ Bayesian estimates like this in a variety of contexts. This particular example used one single source of noisy market share data. In many cases, the researcher has multiple data sources available that report on a product's market share. And these sources may be biased or noisy, each in different ways. For example, the market share for a company's more popular products might be tracked and reported by secondary data sources or perhaps by multiple primary market research efforts. The Bayesian framework is ideal for combining these multiple data sources. In fact, even data sources that report just summary statistics–such as one's market share for the entire geography (without any segment detail) or of one's market share for the entire world (with neither segment or geography detail) are easily allowed to inform the more detailed estimates once one adopts a Bayesian framework. Such summary data sources help inform and tighten the more granular ones.

One can also employ Bayesian methods like this to other (potentially noisy) business metrics entirely. In the marketing world this could include revenue estimation, customer purchase incidence, purchase volumes or shopping basket sizes, new product adoption rates, trial and repeat behavior, customer loyalty indicators,

[11] Out of the 73 cases (39 from the training sample + 34 from the validation sample) where the Bayesian estimates didn't out-predict the raw maximum likelihood estimates, the mix of businesses sizes was 23%, 25%, 22% and 30% for large, medium, public sector, and small establishments, respectively. This was not enough to warrant concern. The mix is similar to each business size's presence in the total sample (which was 23%, 27%, 25% and 26%, respectively). The method is not penalizing any one size segment in particular.

brand switching, or recommender systems. In production or finance operations, candidate measures include quality control measurements, production yield indicators, cost estimates, or rates of return. The tools shown here are not just confined to market share (or batting average) examples.

Bayesian methods like those shown here almost always outperform in cases where there's a need for some sort of multiple group analysis and the groups are measured imperfectly. In this example the groups were summary geographies and segments. However, the "groups" could be singular observations or individuals. Bayesian methods like those estimated here are most helpful when one has at least three such groups of interest. In fact, when fit to groups of just one or two entities, the Bayesian estimates in this example will reduce to approximating those from simple maximum likelihood. [12]

Bayesian models are a little more involved to setup and execute. But as shown here, they're worth the effort and produce more accurate data predictions.

[12] This advice only pertains to the Hierarchical Bayes analysis demonstrated here. However, the Bayesian toolkit is a large one and yet other Bayesian methods are appropriate for improving estimates for fewer groups.