Sawtooth Software

RESEARCH PAPER SERIES

Hierarchical Bayes: Why All the Attention?

Bryan Orme,
Sawtooth Software, Inc.
2000

530 W. Fir St.
Sequim, WA 98382
(360) 681-2300
www.sawtoothsoftware.com

Hierarchical Bayes: Why All the Attention?

Bryan Orme, Sawtooth Software

This article was originally published in Quirk's Marketing Research Review, March 2000.
Bryan Orme is Vice President of Sawtooth Software, Inc. (www.sawtoothsoftware.com). *The author wishes to thank Richard M. Johnson for his helpful comments and technical papers on Hierarchical Bayes that made this article possible.*

If you've been to a technical market research conference lately, you've likely heard presentations advocating a technique called Hierarchical Bayes estimation (HB). The possible applications for HB are far-reaching. If there is heterogeneity among individuals, HB can significantly improve upon traditional aggregate models such as OLS regression or logit for conjoint/choice analysis, customer satisfaction, brand image studies or any other situation in which respondents provide multiple observations.

Until recently, the individuals advocating HB were academics and a few practitioners expert in statistics. HB is demanding both in terms of computational time and complexity. For realistic market research data sets, the run times were counted in days rather than minutes or hours. Given that no easy-to-use HB software existed and computers were not fast enough to deal with real world problems in a reasonable time frame, it is not surprising that some practitioners were skeptical of HB and the hype surrounding it.

Until recently, we too at Sawtooth Software were doubtful that HB would soon achieve very widespread use in the marketing research community. But recent advances in the processing speed of PCs have exceeded our expectations and knowledgeable academics such as Greg Allenby of Ohio State have taught tutorials, published algorithms on HB estimation, and have supported the efforts of individuals such our ourselves in creating off-the-shelf HB software.

What is Hierarchical Bayes?

The Hierarchical Bayes model is called "hierarchical" because it has two levels. At the higher level, we assume that individuals' parameters (betas or part worths) are described by a multivariate normal distribution. Such a distribution is characterized by a vector of means and a matrix of covariances. At the lower level we assume that, given an individual's betas, his/her probabilities of achieving some outcome (choosing products, or rating brands in a certain way) is governed by a particular model, such as multinomial logit or linear regression.

Initial crude estimates of betas are estimated for each respondent to use as a starting point. New estimates are updated using an iterative process called "Gibbs Sampling." The model estimates individual betas as well as the mean and covariances of the distribution of betas. In each iteration, an estimate is made for each parameter, conditional on current estimates of the others. This is done by making a random draw from each conditional distribution. Eventually, after many iterations, this process converges to correct estimates for each parameter. In other words, the HB algorithm produces betas that fit each individual's outcome reasonably well, but

"borrows" information from other respondents to stabilize the estimates.

After a certain number of "burn-in" iterations (often 10,000 or more), convergence is assumed and the estimates of respondent betas are saved after each or (preferably) every *n*th subsequent iteration. These saved results are called "draws" and they reflect the uncertainty around each respondent's estimated betas. Often hundreds or even thousands of draws are saved per respondent. Point estimates of betas are computed for each respondent by averaging the respondent's draws.

Why All of the Attention for HB?

- In application after application where respondents provide multiple-observation data, HB estimation seems at least to match and usually to beat traditional models. Conjoint analysis is a prime example of an application that benefits from HB estimation.
- HB estimation is robust.
- HB permits estimation of individual-level models, which lets marketers more accurately target/model individuals. More specifically, HB permits estimation of models too demanding for traditional methods: even when estimating more beta coefficients per individual than there are individual observations.
- Aggregate estimation models confound heterogeneity and noise. By modeling individuals rather than the "average," HB can separate signal (heterogeneity) from noise. This leads to more stable, accurate models whether viewed in terms of individual- or aggregate-level performance.
- The "draws" (replicates) for each respondent provide a rich source of information for more accurately conducting statistical tests and, for example, estimating nonlinear functions of parameters such as shares of preference.

We do not suggest that HB is a panacea. However, we have been impressed by the way HB handles numerous real-world and synthetic data sets that we have tested. It generally beats other analytical techniques with which we are familiar. We expect HB soon to become a mainstream analytical technique for market research.

The remainder of this article will deal with three common research situations that can benefit from HB estimation: regression analysis, choice-based conjoint (discrete choice) and Adaptive Conjoint Analysis (ACA).

Hierarchical Bayes Regression

Regression analysis is widely used in marketing research for quantifying the relationship between predictor variables and an outcome. The predictor variables are termed *independent variables* and the outcome the *dependent variable*. As an example, in customer satisfaction modeling, the independent variables can be respondents' evaluations of brands on different aspects such as quality, performance, and service after the sale. The dependent variable is usually a measure of overall satisfaction with the brand or likelihood of purchasing that brand again.

Multiple regression models take the general form:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots b_nX_n$$

where,

Y	= dependent variable
b_0	= constant
$b_1 \ldots b_n$	= regression weights (betas)
$X_1 \ldots X_n$	= independent variables

The goal of the model is to minimize the difference between the predicted and actual values of the dependent variable. The degree of fit is termed R^2. An R^2 of zero implies that the predictor variables provide no information to predict the dependent variable, and a value of 1.0 implies perfect fit.

Often in marketing research we tend to apply regression analysis to a group of observations that individually have different relationships between the independent and dependent variables. Consider people's opinions toward anchovies on pizza. Anchovies are generally either liked or despised. It is rare to find an individual who is lukewarm about anchovies. The distribution of individuals with respect to anchovy preference is not a normal bell-shaped curve, but perhaps has two "humps," reflecting the mixture of two very different populations.

Consider a hypothetical satisfaction study for pizza in which respondents tasted four different pizzas (some with anchovies and some without) and then rated each pizza on an overall desirability scale. To analyze the data, we apply a regression model to predict respondents' satisfaction for pizza based on whether it had anchovies or not. Let's assume that the independent variable (X_1) indicating whether a pizza had anchovies or not was dummy coded (0=no anchovies; 1=has anchovies). Further assume that half of the population loves anchovies and their true beta weight b_1 (the increase in satisfaction due to the presence of anchovies) is +10 (plus or minus some error). The other half of the population despises anchovies, b_1 = -10, again plus or minus some error.

When we pool the data and estimate b_1, we discover that b_1 is close to but not significantly different from 0, and the R^2 is also near zero. (Both values would be exactly zero if respondents answered without noise and all used the rating scale in the same way.) Without any additional information, we'd be tempted to report that anchovies do not affect people's satisfaction with pizza whatsoever. And we would be dead wrong. The aggregate regression model has ignored heterogeneity and simply tried to describe the "average" respondent. Moreover, because aggregate regression cannot distinguish between (confounds) heterogeneity and noise, the estimate of b_1 is not as precise as it could be.

Hierarchical Bayes (HB) can deal much better with this situation. HB "borrows" information from other respondents to compute relatively stable individual-level results when respondents provide multiple observations (in our example, respondents evaluated multiple pizzas). One can even estimate useful individual-level models for more independent variables than a respondent

has given observations-an impossible feat for traditional regression.

By estimating betas separately for each individual rather than just for the average of all people, HB separates heterogeneity (signal) from noise. The use of HB for this problem would reveal that anchovies significantly affect peoples' satisfaction for pizza. For HB, the average R^2 (the result of R^2 measured at the individual level and then averaged across respondents) is significantly greater than 0. If we submitted the individual-level betas to a cluster analysis, we'd learn that there were two distinct types of people with opposite opinions. We'd note that mean value for b_1 was near zero. But, because HB has been able to separate the heterogeneity from the noise, the average estimate of b_1 is more precise, and closer to zero than with aggregate regression.

Those readers attuned to the assumptions of HB will point out that this hypothetical situation is at odds with HB's assumption that respondent betas conform to a normal distribution. The beta weights are indeed tempered by this assumption, but the observations provided by each individual still strongly influence the individual-level betas. We've analyzed a synthetic data set conforming to the pizza example with HB and observed that it deals well with this problem. Respondents are separated into their respective populations, the individual estimates of beta closely fit the true betas, and estimates of aggregate betas are measured more precisely than under aggregate regression.

It is worth noting that Latent Class methods are also useful in dealing with heterogeneous populations. For this simple pizza example, a two-group Latent Class solution would be entirely appropriate. However, Latent Class solutions are subject to local minima, they typically do not achieve proper individual-level estimates and, like cluster analysis, the analyst must decide how many groups (classes) are appropriate.

While the pizza example above is a very simple illustration, the principles are important and relevant to more complicated regression problems in marketing research: for example, ratings-based conjoint analysis or price-elasticity measurement from scanner data, where the unit of analysis is stores rather than individuals. The major take-aways are as follows:

- If respondents (or another unit of analysis such as stores) provide multiple observations, HB can be used to estimate individual-level betas.

- HB can distinguish between the heterogeneity and noise that aggregate regression modeling confounds. This results in more precise estimates of average betas than under aggregate regression.

- The individual-level beta weights can be used to segment respondents, using methods such as cluster analysis, neural networks, CHAID or AID, or banner points (filters) such as in cross-tabs.

Another problematic issue that often derails multiple regression models is lack of independence (collinearity) among the independent variables. Consider a customer satisfaction study in which respondents evaluate multiple brands on various product-related features (independent variables) and then provide an overall evaluation of the brand (dependent variable). The goal of such a study might be to derive the weight (importance) each feature has in driving overall satisfaction. If some of the attributes have overlap in meaning for many of the respondents, such as "reliability" and "durability," regression modeling will have a difficult time distinguishing the relative weight of these two related items. As a result, collinearity leads to unstable estimates of beta weights. HB significantly reduces this problem by distinguishing heterogeneity from noise and by leveraging information from respondents whose ratings reflect better discrimination between "reliability" and "durability" to improve the measurement for respondents who did not make such distinctions. The result is more precise estimates of both individual and aggregate beta weights.

Marketers should be more concerned with profiling and targeting individuals and segments rather than the market average. HB methods support this strategy and are very valuable for problems that have traditionally been analyzed using aggregate regression. Whether the researcher's interest lies in achieving aggregate- or individual-level estimates of beta, for studies in which respondents provide multiple observations, HB usually beats aggregate regression.

HB for Choice Data

Choice-based conjoint (discrete choice) measurement has grown in popularity over the last five years. Many researchers assert that choice-based tasks are more realistic for respondents than ratings- or rankings-based conjoint questions. However, choice-based conjoint data don't contain as much information per unit of respondent effort as traditional conjoint analysis. Respondents evaluate multiple products in choice sets, but they typically only indicate which *one* within the set they would choose. We don't learn how much more desirable the chosen product is over those not chosen, nor do we ascertain the relative values of the non-chosen product concepts. As a result, stable individual-level estimation was previously not feasible. Researchers, rather, pooled respondent data using methods such as logit to model the "average" respondent.

Using the logit rule on aggregate data led to IIA (red-bus/blue-bus) problems in simulations. A new alternative in a choice simulation took share from existing products in proportion to their shares. Cross-elasticities and substitution rates among competing products were assumed to be equal, which certainly wasn't realistic. To alleviate these problems, some analysts turned to building more complex models with additional terms to account for respondent characteristics, cross-effects, availability effects and interactions. These models were complicated to build, and the specification could balloon into a very large number of terms. Estimating so many terms ran the risk of overfitting. Still, for the expert logit modeler, the results could be quite satisfactory and could largely overcome the IIA problems resulting from aggregation.

Other techniques such as Latent Class analysis were developed to deal with the problems of aggregation and IIA. The Latent Class approach segmented the market into relatively homogenous groups and fit an average model within each group. Latent Class analysis is an

important development and is very useful for market segmentation. Even though Latent Class helps reduce IIA problems, it fails to provide accurate individual-level estimates.

Then came Hierarchical Bayes. The HB algorithm can also be adapted for choice data, where the model is a logit specification and the fit is measured in terms of log-likelihood. Its ability to borrow information from other respondents to stabilize part worth estimation for each individual is particularly valuable for choice data. Rather than rely on the logit rule for market simulations, the researcher can apply a first choice (maximum utility) rule to the individual-level estimates (or the multiple draws). The first choice rule is immune to IIA difficulties.

Applying HB to choice data lets analysts largely solve IIA problems and capture complex cross-effects (through market simulations) using very simple model specifications (such as main effects only).

HB for ACA Part Worth Estimation

According to a 1997 industry survey conducted by Wittink, Vriens and Huber, ACA is the most widely used methodology in the world for conjoint analysis. Given its popularity, it is not surprising that ACA has been widely scrutinized and been the subject of a great deal of debate. Most notably, in a 1991 *Journal of Marketing Research* article by Green, Krieger and Agarwal, ACA was criticized because of potential scale incompatibilities between the self-explicated priors and conjoint pairs sections of the interview.

ACA version 4 was released shortly after the 1991 JMR article. It used a slightly different technique from earlier ACA versions for combining self-explicated priors and conjoint pairs information. Even after the upgrade, ACA remained an approach that, while having a loyal following and working well in practice, was still based on computational procedures that to some academics and statisticians were not theoretically pure.

In 1995, Allenby, Arora and Ginter published an article also in the *Journal of Marketing Research* reporting improvements for ACA through Hierarchical Bayes estimation. Allenby and a number of co-author's collective work on HB methods has been ground breaking and important.

HB provides two major benefits for ACA part worth estimation:

> 1) HB improves the quality of each individual's utility estimates by "borrowing" information from other individuals. This translates to more accurate predictions of both individual choices and share estimations. We have tested the results on dozens of real and synthetic data sets. HB at least matches and usually beats traditional ACA utility estimation.
>
> 2) HB provides a more theoretically sound way of combining data from the self-explicated and paired comparison sections of the interview. Because the priors information can be applied in a purely ordinal way as constraints, it entirely avoids the issue of combining two separate sets of metric dependent variables with potentially

different variances.

Not only is the technique more defensible, but the results are generally better. Notably:

- HB utilities are less biased toward equal utility increments spacing between levels as compared to ACA v4.

- HB importances reflect slightly more discrimination than under ACA v4.

- HB does a better job of estimating utilities for the levels not taken forward into pairs when using "Most Likelies" and "Unacceptables."

In addition to those benefits, ACA surveys can now be shorter. HB estimation does not require the calibration concept data (unless one wants to calibrate the data for purchase likelihood simulations). Therefore, this sometimes-confusing section can be cut from ACA surveys. Rather than reducing the length of the interview, the researcher might decide instead to add a few more pairs questions to further stabilize utilities.

Summary and Conclusion

Hierarchical Bayes estimation is coming of age for market researchers. Academics have published the algorithms and off-the-shelf software is available. PCs are now fast enough to handle small to medium-sized market research problems in a reasonable time (usually between 30 minutes to 4 hours). But large marketing research problems may still require many hours of processing time.

By using HB estimation, researchers can improve the reliability and predictive validity of their models. HB estimation helps with some common, vexing challenges, including trying to estimate stable individual-level models from sparse data, multicollinearity and the IIA (red-bus/blue-bus) problem in logit simulations. Moreover, the draws generated by HB are useful for statistical testing and estimating non-linear functions of the parameters.