Sawtooth Software

RESEARCH PAPER SERIES

Understanding HB: An Intuitive Approach

Richard M. Johnson,
Sawtooth Software
2000

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

Understanding HB: An Intuitive Approach

Richard M. Johnson
Sawtooth Software, Inc.
March, 2000

Introduction:

Hierarchical Bayes Analysis (HB) is sweeping the marketing science field. It offers valuable benefits, but it's so different from conventional methods that it takes some getting used to. The purpose of this talk is to convey an intuitive understanding of what HB is, how it works, and why it's valuable.

We'll start with a simple example to show how Bayesian analysis is fundamentally different from what we usually do. Then I'll describe hierarchical models useful for estimating conjoint part worths and the algorithm used for computation. I'll finish with evidence that HB does a better job than conventional analysis, and describe some of the questions that remain to be answered.

The Basic Idea of Bayesian Analysis:

In a small air charter service there was a new pilot named Jones. Jones was an excellent pilot in every way but one – he had trouble making smooth landings. The company did customer satisfaction research, and one of the questions asked of passengers was whether the landing had been smooth. Fully 20% of Jones's landings were judged to be rough. All the other pilots had better scores, and the average percentage of rough landings for the rest of them was only 10%.

We could summarize that information in a table like this, which gives probabilities of rough or smooth landings for Jones and all other pilots combined:

Probabilities of Rough or Smooth Landings,
Conditioned on Pilots

	Rough	Smooth	Total
Jones	.20	.80	1.00
Other	.10	.90	1.00

One day a passenger called the president of the company and complained about a rough landing she had experienced. The president's first reaction was to think "Probably Jones again." But then he wondered if that was reasonable. What do you think?

These data do give us a probability: If Jones is the pilot, the probability of a rough landing is .2. This is the conditional probability of a rough landing, given that the pilot is Jones. But the president was considering a different probability: If it was a rough landing, what was the probability that the pilot was Jones?, which is the conditional probability of Jones, given it was a rough landing.

In conventional statistics we are accustomed to assuming a hypothesis to be true, and then asking how the data would be expected to look, given that hypothesis. An example of a conventional research question might be, "Given that Jones is the pilot, what's the probability of a rough landing?" And by a conventional analysis we can learn the answer, which is "Twenty Percent."

With Bayesian analysis we can turn things around and ask the other question: "Given that the landing was rough, what's the probability that Jones was the pilot?"

We can't answer that question from these data, because we don't know anything about how often Jones flies. For example, if there were a million pilots, all flying equally often, the answer to this question would obviously be different than if there were only two.

To answer the president's question we must provide one further kind of information, which Bayesians call "Prior Probabilities." In this case, that means, "Irrespective of this particular flight, what is the probability that Jones will be the pilot on a typical flight." Suppose this company has 5 pilots, and they are all equally likely to have drawn this flight on this particular day. Then a reasonable estimate of the prior probability that Jones would have been the pilot would be one fifth, or .2.

Suppose we multiply each row of the previous table by its prior probability, multiplying Jones's row by .2, and everyone else's row by .8, to get the following table:

Joint Probabilities of Rough or Smooth Landings,

	Rough	Smooth	Total
Jones	.04	.16	.20
Other	.08	.72	.80
Total	.12	.88	1.00

This table gives the probabilities of all combinations of pilots and outcomes, saying that out of 100 flights with this airline, we should expect 4 rough landings by Jones. Probabilities like these are called "Joint Probabilities" because they describe combinations of both variables of interest. Notice that they sum to unity. The row sums give the overall probability of each type of pilot flying, and the column sums give the overall probability of each kind of landing.

Once we have joint probabilities, we are able to answer the president's question. Given that the flight was rough (.12), the probability that Jones was the pilot was only .04 / .12, or one third. So the president was not justified in assuming that Jones had been the pilot.

Three kinds of probabilities are illustrated by this example.

> **Prior Probabilities** are the probabilities we would assign *before we see the data*. We assign a prior probability of .2 for Jones being the pilot, because we know

there are 5 pilots and any of them is equally likely to be at the controls for any particular flight.

Likelihood is the usual name for the *probability of the data, given a particular hypothesis or model*. This is the kind of probability we're used to: the likelihood of a rough landing, given that the pilot is Jones, is .2.

Posterior Probabilities are the probabilities we would assign *after we have seen data*. Posterior probabilities are based on the priors as well as information in the data. Combining the priors and the data, our posterior probability that the pilot of a rough landing was Jones is one third. After learning that flight had a rough landing, the probability that Jones was its pilot was updated from .2 to .333.

Bayes' Rule:

To discuss probabilities, we need some notation. For any two events, X and Y, we define:

P(X) = the marginal probability of X (e.g., without respect to Y)

P(X,Y) = the joint probability of X and Y (e.g. probability of both X and Y)

P(X | Y) = the conditional probability of X given Y

The definition of conditional probability is

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)} \qquad (1)$$

Starting from equation (1) we can derive Bayes' Rule by simple algebra. Multiply both sides of equation (1) by P(Y) to get

$$P(X \mid Y) * P(Y) = P(X, Y) \qquad (2)$$

Since X and Y could be any events, we can write equation (3) just by exchanging the roles of X and Y in equation (2):

$$P(Y \mid X) * P(X) = P(Y, X) \qquad (3)$$

Noting P(X ,Y) is the same thing as P(Y ,X), we can equate the left hand sides of equations (2) and (3), getting

$$P(X \mid Y) * P(Y) = P(Y \mid X) * P(X) \qquad (4)$$

Finally, dividing both sides by P(Y), we get what is known as "Bayes' Rule, "

$$P(X \mid Y) = \frac{P(Y \mid X) * P(X)}{P(Y)} \qquad (5)$$

Bayes' Rule gives us a way of computing the conditional probability of X given Y, if we know the conditional probability of Y given X and the two marginal probabilities. Let's apply Bayes rule to our example. Let X be the event "Jones was the pilot" and Y be the event "Rough Landing." Then Bayes' Rule says:

$$P(\text{ Jones} \mid \text{Rough }) = \frac{P(\text{ Rough} \mid \text{Jones }) * P(\text{ Jones })}{P(\text{ Rough })} = \frac{.2 * .2}{.12} = 1/3 \qquad (6)$$

This is the same computation that we performed before. Bayes' rule gives us a way to answer the question posed by the president. This same relationship between probabilities underlies the entire field of Bayesian analysis, including HB.

In real-life Bayesian applications, the probability in the denominator in equation (5) is often hard to compute. Also, since it often depends on arbitrary factors like the way measurements are made and data are coded, it is seldom of much interest. For example, the president wasn't wondering about the proportion of flights that had rough landings, which can be determined easily by other means. Our question was: *Given* a rough landing, what was the probability that Jones was the pilot?

Therefore, in practical applications the denominator is often regarded as a constant, and Bayes rule is expressed as:

$$P(X \mid Y) \propto P(Y \mid X) * P(X) \qquad (7)$$

where the symbol $\propto$ means "is proportional to." The main thing to remember is the relationship indicated in equation (7), which may be stated: ***Posterior probabilities are proportional to likelihoods times priors.*** If you can remember this slogan, you are well on the way to understanding Bayesian analysis.

Some Characteristics of Bayesian Analysis

We've used this simple example to introduce Bayes' rule, and to show how to produce posterior probabilities by updating prior probabilities with likelihoods obtained from data. Bayesian analysis differs from conventional analysis in several ways:

Bayesian analysis is sometimes said to involve "subjective probabilities" because it requires specification of priors. The priors in the example were obtained by assuming

each pilot was equally likely to have any flight, but such sensible priors are not always available. Fortunately, there is enough data in large-scale applications using HB that the priors have very little effect on the posterior estimates.

Bayesian analysis is sometimes said to involve "inverse probabilities." In conventional analysis, we regard parameters as fixed and the data as variable. In Bayesian analysis things are reversed. After the data are in hand we regard them as fixed, and we regard the parameters as random variables, with distributions that we try to estimate.

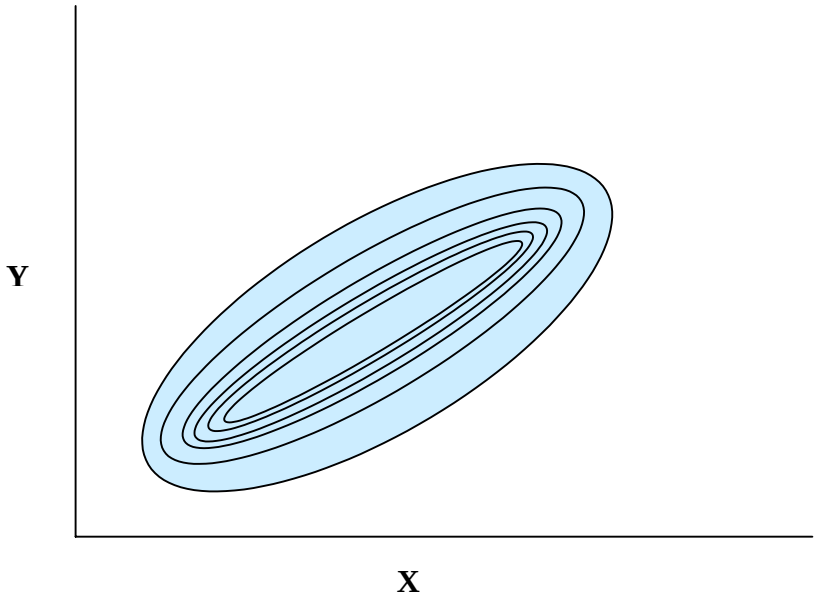
Although many Bayesian models are simple in concept, their actual estimation is often difficult, and requires computer simulations. Those simulations can take a long time – perhaps many hours rather than just a few seconds or minutes.

In exchange for these complexities, Bayesian methods offer important benefits. HB can produce better estimates of individual values. For conjoint analysis, we can get equivalent accuracy using shorter questionnaires. We can also get useful individual estimates where before we might have had to settle for aggregate estimates. And this is true not only for conjoint analysis, but also for customer satisfaction research and scanner data.

Markov Chain Monte Carlo Analysis:

Bayesian analysis has benefited enormously from recent increases in computer speed. A group of methods which have been especially useful for HB estimation are known as MCMC or "Monte Carlo Markov Chain" methods. One MCMC method particularly useful for HB is called the "Gibbs Sampler." The basic idea behind the Gibbs Sampler can be demonstrated by a small example.

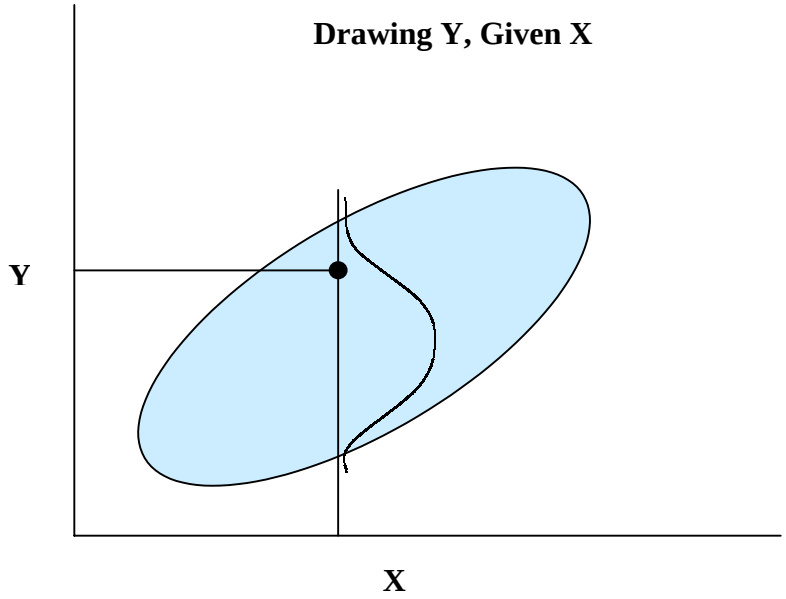
Suppose X and Y are two normally distributed variables with similar variability. The joint distribution of two variables in a sample is often shown with a contour map or a scatter-plot. If the variables are uncorrelated, the scatter-plot is approximately circular in shape. If the variables are correlated, the swarm of points is elongated.



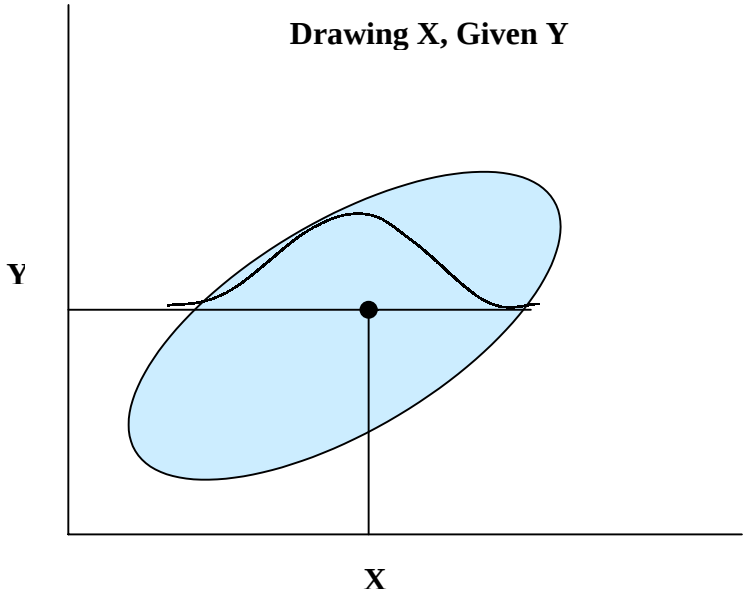
Suppose we don't know the joint distribution of X and Y, but wish we did. Suppose, however, that we know both conditional distributions: given a value for X, we know the distribution of Y conditional on that X; and given a value for Y, we know the distribution of X conditional on that value for Y. This information permits us to simulate the joint distribution of X and Y using the following algorithm:

Use any random value of X to start.

Step (1) Given X, draw a value of Y from the distribution of Y conditional on X.

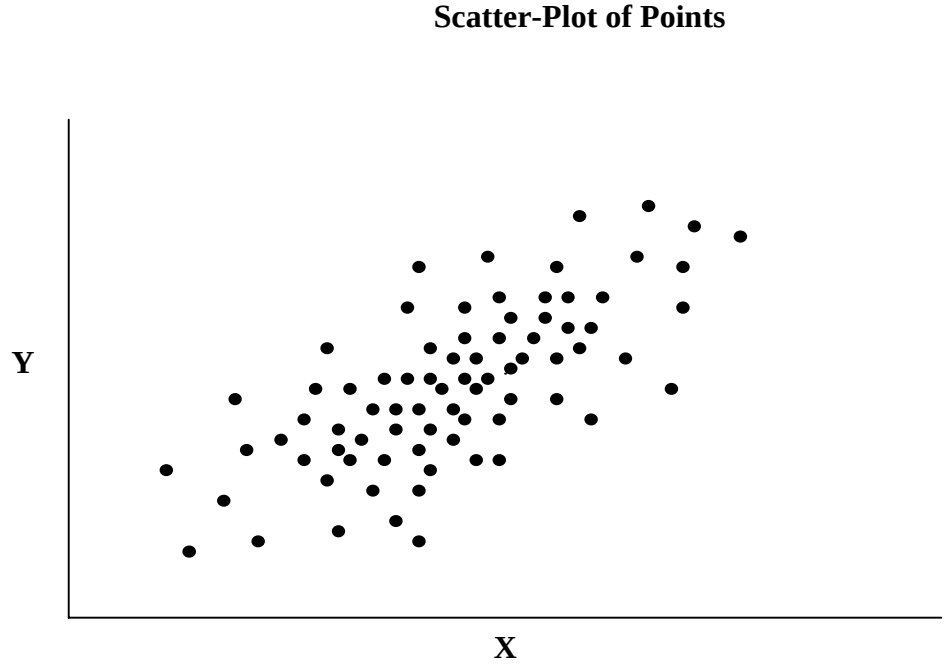


Step (2) Given Y, draw a value of X from the distribution of X conditional on Y.



Repeat steps 1 and 2 many times, labeling successive draws X1, X2, X3,…
And Y1, Y2, Y3…. Plot each pair of points Xi,Yi.

As the number of steps (iterations) increases, the scatter-plot of the successive pairs of X,Y approximates the joint distribution of X and Y more and more closely.



This principle is valuable for HB, because the *joint* distribution of many variables can horrendously complicated and impossible to evaluate explicitly. But the statistical properties of normal distributions permit us to estimate *conditional* distributions much more easily. We use a similar iterative process where we repetitively select each variable and estimate all the others conditionally upon it. This process is a Markov chain because the results for each iteration depend only on the previous iteration, and are governed by a constant set of transition probabilities.

Hierarchical Models for Conjoint Analysis:

In this presentation we skip technical details in an attempt to give an intuitive idea of how HB works. Explanatory papers are on the sawtoothsoftware.com web site providing further details.

Suppose many survey respondents have each answered several choice questions.
We want to estimate part-worths for each individual contained in a vector **b**, the mean for the population of individuals contained in the vector **a**, and the variances and covariances for the population contained in the matrix **C**.

The model is called "hierarchical" because it has two levels. At the *upper level*, we assume that individuals' vectors of part-worths are drawn from a multivariate normal distribution:

$$\textbf{Upper Level Model:} \qquad \mathbf{b \sim N(a , C)}$$

At the *lower level*, we assume a logit model for each individual, where the utility of each alternative is the sum of the part-worths of its attribute levels, and the respondent's probability of choosing each alternative is equal to its utility divided by the sum of utilities for the alternatives in that choice set.

$$\textbf{Lower Level Model:} \qquad \mathbf{u} = \Sigma\, \mathbf{b_i}$$

$$\mathbf{p} = \mathbf{exp(u)} \,/\, \Sigma\, \mathbf{exp(u_j)}$$

One starts with initial estimates for of **a**, the **b**'s, and **C**. There is great latitude in choosing these estimates. Our estimates of **b** for each individual are the numbers of times each attribute level is in the chosen alternatives, divided by the number of times each attribute level is present in all alternatives. Our initial estimate of **a** has all elements equal to zero, and for **C** we set initial variances at unity and covariances at zero.

The algorithm repeats these steps in each iteration.

> Step(1) Given current estimates of the **b**'s and **C**, estimate the vector **a** of means of the distribution.

Step(2) Given current estimates of the **b**'s, and **a**, estimate the matrix **C** of variances and covariances.

Step(3) Given current estimates of **a**, and **C**, estimate a new **b** vector for each respondent.

The process is continued for many thousands of iterations. The iterations are divided into two groups.

The first several thousand iterations are used to achieve convergence, with successive iterations fitting the data better and better. These are called "preliminary," "burn-in," or "transitory" iterations.

The last several thousand iterations are saved for later analysis, to produce estimates of the **b**'s, **a**, and **C**. Unlike conventional statistical analysis, successive iterations do not converge to a single "point-estimate" for each parameter. Even after convergence, estimates from successive iterations bounce around randomly, reflecting the amount of uncertainty that exists in those estimates. Usually we want a point-estimate of part-worths for each respondent, and this is obtained simply by averaging estimates of that individual's **b**'s for the last several thousand iterations.

During iterations, successive values of **a** and **C** are estimated by straight-forward statistical procedures that we shall not consider. However, successive values of the **b**'s are estimated by a "Metropolis-Hastings" algorithm which illustrates the Bayesian nature of the process, and which we shall now describe.

The following is done for each individual at each iteration:

Define the individual's previous estimate of **b** as $\mathbf{b}_{\mathbf{old}}$. Construct a candidate for a new estimate for that individual by adding a small random perturbation to each element of $\mathbf{b}_{\mathbf{old}}$, calling the resulting vector $\mathbf{b}_{\mathbf{new}}$. Using the data and the logit model, we compute the likelihood of seeing that respondent's set of choices, given each of those **b** vectors. Each likelihood is just the product of the predicted probabilities of all choices made by that respondent, given that estimate of **b**. We compute the ratio of those likelihoods, $\mathbf{l}_{\mathbf{new}} / \mathbf{l}_{\mathbf{old}}$.

Recall that the hierarchical model regards the individuals' **b** vectors to have been drawn from a multivariate normal distribution with mean vector **a** and covariance matrix **C**. We can use standard statistical formulas to compute the relative probabilities that $\mathbf{b}_{\mathbf{new}}$ and $\mathbf{b}_{\mathbf{old}}$ would have been drawn from that distribution, indicated by the height of the distribution's graph at each point. We compute the ratio of those probabilities, $\mathbf{p}_{\mathbf{new}} / \mathbf{p}_{\mathbf{old}}$.

Finally, we compute the product of ratios,

$$\mathbf{r} = (\mathbf{l}_{new} / \mathbf{l}_{old})*(\mathbf{p}_{new} / \mathbf{p}_{old}) = \frac{\mathbf{l}_{new} * \mathbf{p}_{new}}{\mathbf{l}_{old} * \mathbf{p}_{old}} \quad (8)$$

Recall that ***posterior probabilities are proportional to likelihoods times priors***. The **p**'s may be regarded as priors, since they represent the probabilities of drawing each vector from the population distribution. Therefore, **r** is the ratio of posterior probabilities of $\mathbf{b}_{new}$ and $\mathbf{b}_{old}$.

If **r** is greater than unity, the new estimate has a higher posterior probability than the previous one, and we accept $\mathbf{b}_{new}$. If **r** is less than unity we accept $\mathbf{b}_{new}$ with probability equal to **r**.

Over the first several thousands of iterations, the **b**'s gradually converge to a set of estimates that fit the data while also conforming to a multinormal distribution.

If a respondent's choices are fitted well, his estimated **b** depends mostly on his own data and is influenced less by the population distribution. But if his choices are poorly fitted, then his estimated **b** depends more on the population distribution, and is influenced less by his data. In this way, HB makes use of every respondent's data in producing estimates for each individual. This "borrowing" of information is what gives HB the ability to produce reasonable estimates for each respondent even when the amount of data available for each may be inadequate for individual analysis.

Typical HB Results for Choice Data:

Huber, Arora & Johnson (1998) described a data set in which 352 respondents answered choice questions about TV preferences. There were 6 conjoint attributes with a total of 17 levels. Each respondent answered 18 customized choice questions with 5 alternatives, plus 9 further holdout choice questions, also with 5 alternatives.

We examine HB's ability to predict holdout choices using part-worths estimated from small numbers of choice tasks. Part-worths were estimated based on 18, 9, 6, and 4 choices per respondent. Point estimates of each individual's part-worths were obtained by averaging 100 random draws, and those estimates were used to measure hit rates for holdout choices. The random draws were also used in 100 first-choice simulations for each respondent, with predicted choices aggregated over respondents, to measure Mean Absolute Error (MAE) in predicting choice shares. Here are the resulting Hit Rate and MAE statistics for part-worths based on different numbers of choice tasks:

TV Choice Data
Holdout Prediction With Subsets Of Tasks

# Tasks	Hit Rate	MAE
18	.660	3.22
9	.602	3.62
6	.556	3.51
4	.518	4.23

Hit rates are decreased by about 10% when dropping from 18 choice tasks per respondent to 9, and by about 10% again when dropping from 9 tasks to 4. Similarly, mean absolute error in predicting choice shares increases by about 10% each time the number of tasks per respondent is halved. Even with as few as four questions per respondent, hit rates are much higher than the 20% that would be expected due to chance.

Typical HB Results for ACA Data:

This data set was reported by Orme, Alpert, and Christensen (1997) in which 80 MBA students considered personal computers, using 9 attributes, each with two or three levels. Several kinds of data were collected in that study, but we now consider only ACA data plus first choices from five holdout tasks, each of which contained three concepts.

ACA provides information about each respondent's "self-explicated" part-worths, as well as answers to paired-comparison tradeoff questions. The HB user has the option of fitting just the paired comparison data, or the paired comparison data plus the self-explicated information. Also, there is an option of constraining the part-worths to obey order relations corresponding to the self-explicated information. Here are hit rates for several methods of estimating part-worths:

	MBA Data Hit Rate %
ACA Version 4 ("optimal weights")	66.25
Pairs Only With Constraints	71.50
Pairs + Self-Explicated With Constraints	70.75
Pairs + Self-Explicated, No Constraints	69.35
Pairs Only, No Constraints	69.00

It is noteworthy that all four of the HB combinations are more successful than the conventional method of estimation offered by ACA. Similar results have been seen in several other comparisons of HB with conventional ACA estimation methods.

Typical HB Results for Regression Data:

In the Orme et al. study each respondent also did a full-profile card-sort in which 22 cards were sorted into four piles based on preference, and then rated using a 100-point scale. Those ratings were converted to logits, which were used as the dependent variable,

both for OLS regression and also by HB. In these regressions each respondent contributed 22 observations and a total of 16 parameters were estimated for each, including an intercept. Here are hit rates for predicting holdout choices:

MBA Data

Ordinary Least Squares	72.00%
HB	73.50%

HB has a 1.5% margin of superiority. This is typical of results seen when HB has been compared to individual least squares estimation.

Remaining Questions:

Although HB users have seemed pleased with their results, there are two important problems yet to be solved. The first problem is that of enforcing **order constraints**. Conjoint analysis is usually more successful if part-worths are constrained so that values for more desirable levels are greater than values for less desirable levels. Not only is this true for attributes where everyone would agree that "more is better", but it also appears to be true in ACA when each respondent's part-worths are constrained to have the same order relations as his self-explicated information.

Our HB module for ACA does permit enforcing such constraints, but to accomplish this we have had to employ a procedure which is not strictly correct. We are working on this problem, and hope soon to include a statistically-correct capability of enforcing order constraints in each of our HB software products.

The second problem is that of knowing **how many iterations to specify**. After the iterations have been completed, it is not difficult to look at their history and decide whether the process appeared to converge during the "burn-in" period. But not much information has been available about how many iterations are likely to be required. To be on the safe side, we have suggested between 10 and 20 thousand preliminary iterations. But if fewer are really required, it may be possible to cut run times dramatically. The authors of the next paper have examined this question, and I look forward to hearing what they have to report.

Should You Use HB?

There is no denying that HB takes longer than other methods of individual estimation. We have seen run times ranging from a few minutes to a few days. Here are some timing examples for HB-Reg, the package for general regression applications such as customer satisfaction or scanner data. The time required is approximately proportional to number of respondents, the average number of answers per respondent, and (for large numbers of variables) the square of the number of variables. Using a computer with a 500 MHz

Pentium III processor, we have observed these times for 20,000 iterations with 300 respondents and 50 answers per respondent:

Representative Times for Computation

Number of Variables	Time for 20,000 Iterations
10	1 hour
40	3 hours
80	14 hours

To summarize our experience with HB, part-worths estimated by HB have usually been better and almost never worse at predicting holdout choices than part-worths estimated by previously available methods.

HB also has the valuable property of being able to produce useful individual estimates even when few questions are asked of each respondent, in situations where previously the researcher had to settle for aggregate estimates.

I have not said much about customer satisfaction research, but there is also evidence that HB produces better estimates of attribute importances in the face of colinearity, such as often found in customer satisfaction data.

Given our favorable experience with HB estimation, we recommend using HB whenever the project schedule permits doing so.

References

Huber, J., Arora, N., and Johnson, R. (1998) "Capturing Heterogeneity in Consumer Choices," *ART Forum*, American Marketing Association.

Orme, B. K., Alpert, M. I. & Christensen, E. (1997) "Assessing the Validity of Conjoint Analysis – Continued," *Sawtooth Software Conference Proceedings*, Sawtooth Software, Sequim.

Sawtooth Software (1998) "The CBC/HB Module for Hierarchical Bayes Estimation," Technical Paper accessible from sawtoothsoftware.com web site.

Sawtooth Software (1999) "The ACA/HB Module for Hierarchical Bayes Estimation," Technical Paper accessible from sawtoothsoftware.com web site.

Sawtooth Software (1999) "HB-Reg for Hierarchical Bayes Regression," Technical Paper accessible from sawtoothsoftware.com web site.