On the Similarity of Classical and Bayesian Estimates of Individual Mean Partworths

by

Joel Huber

Fuqua School of Business

Duke University

and

Kenneth Train

Department of Economics

University of California, Berkeley

August 3, 2000

Abstract: An exciting development in modeling has been the ability to estimate reliable individual-level parameters for choice models. Individual partworths derived from these parameters have been very useful in segmentation, identifying extreme individuals, and in creating appropriate choice simulators. In marketing, hierarchical Bayes models have taken the lead in combining information about the aggregate distribution of tastes with the individual's choices to arrive at a conditional estimate of the individual's parameters. In economics, the same behavioral model has been derived from a classical rather than a Bayesian perspective. That is, instead of Gibbs sampling, the method of maximum simulated likelihood provides estimates of both the aggregate and the individual parameters. This paper explores the similarities and differences between classical and Bayesian methods and shows that they result in virtually equivalent conditional estimates of partworths for customers. Thus, the choice between Bayesian and classical estimation becomes one of implementation convenience and philosophical orientation, rather than pragmatic usefulness.

On the Similarity of Classical and Bayesian Estimates of Individual Mean Partworths

Introduction

More than any other discipline, marketing is concerned with predicting individual choice. After all, choice is what consumers do in selecting among alternatives in the marketplace. A focus on the individual values provides a critical foundation of segmentation, identification of prospects and as input to market simulators (Wedel et al. 1999). Initially, choice experiments were estimated at an aggregate level. Ratings-based conjoint methods (Green and Srinivasan 1990) were needed to predict individual choices since there is much more information in ratings of four alternatives compared with a selection of one.

Recently hierarchical Bayes has provided a way to estimate remarkably stable individual choice models (Allenby and Ginter 1995; Lenk, Desarbo, Green and Young 1996, Sawtooth Software 1999). Within a Bayesian framework, these models estimate the distribution of coefficients across the population and combine information with the individual's choices to derive posterior or conditional estimates of the individual's values. At the same time, random or mixed coefficient choice models arising from a classical framework have permitted a similar analysis by combining maximum likelihood estimates of the population distribution with individual choices (Revelt and Train, 1999). In this paper we examine the empirical differences between these classical and hierarchical Bayes estimates. Both methods share the same behavioral assumptions, but derive from quite different estimation techniques and interpretive philosophies.

It is not the intent of this paper to bridge the cultural chasm between Bayesian and classical statisticians. Indeed, divisions between these frameworks have resisted twenty years of efforts to bring them together (Gelman et al. 1995). The goal of this paper is instead to show that the

distinction is irrelevant for the purposes of estimating the mean partworths of individuals. Since these estimates are critical for those developing segments, identifying outliers and simulating market choices, it does not matter which method is used for the majority of our practical and theoretical research.

The two procedures are related numerically.[1] When the same model is specified under the two approaches, estimates from the classical and Bayesian procedures converge asymptotically (Lindsey 1996). In small samples, the two procedures provide numerically different results, due to the different ways of treating uncertainty in the parameters of the population distribution.[2] The relevant question is then an assessment of how different the results are in small samples.

We investigate this question with a sample of 361 customers, using both classical and Bayesian procedures to estimate the mean of the conditional distribution for each customer. We use a mixed or random coefficients logit specification of behavior, though other behavioral representations could be used instead. We find that the results are remarkably similar for the two approaches, even with this relatively small sample.

The similarity of results, asymptotically and in our small-sample example, means that in many contexts the differences between the two approaches arise in how the results are interpreted more than in the numerical estimates themselves. As has been found and exploited for other kinds of models, the advantages of Bayesian numerical procedures for mixed logits can be utilized while retaining a classical perspective, and classical procedures can be applied in situations where they provide numerical conveniences, without abandoning a Bayesian perspective.

[1] The relation is primarily due to the fact that the posterior distribution is proportional to the likelihood function times the prior distribution. For a flat prior (as is usually specified) or asymptotically for any prior that nowhere vanishes, the posterior distribution, which is the basis for Bayesian estimation, is therefore proportional to the likelihood function, which is used for classical estimation. Also, the mean of the posterior, taken as an estimator, is asymptotically equivalent to the maximum likelihood estimator.

[2] The Bayesian approach represents the uncertainty in terms of a posterior distribution, while the classical approach represents it with the asymptotic sampling distribution of the maximum likelihood estimator.

In section 1, we provide the specification of mixed logits and describe how they are conceptualized and estimated in both the classical and Bayesian perspectives. Section 2 presents our application of a mixed logit model on 361 customers, comparing the results from the two procedures. Section 3 concludes with a discussion of the pragmatic reasons to choose one system over the other.

1. Specification of Mixed Logits

We first describe the behavioral specification, which is shared by both approaches, and then briefly describe the conceptualization and techniques of estimation, which are different for the two approaches. We assume that the partworths are normally distributed in the population. This assumption is consistent with the commercially available software for the Bayesian approach (Sawtooth Software 1999). However, it is not required for either approach, and we later discuss the use of non-normal distributions.

1.1 Behavioral Specification

Assume each customer faces a choice among J alternatives in each of T choice situations.[3] Customer n is assumed to choose the alternative in choice situation t with the highest utility. The utility of alternative i as faced by customer n in situation t is modeled as:

$$U_{nit} = \beta_n' X_{nit} + e_{nit,} \qquad (1)$$

where X_{nit} is a vector of independent variables that are observed by the researcher, such as attributes of the alternative i in choice situation t. These independent variables are considered non-stochastic. By contrast, the terms β_n and e_{nit} are not observed by the researcher and are considered stochastic. The coefficient vector, β_n, is assumed to be distributed normally across

the population, independent of e and X, with mean vector b and covariance matrix W. The term e_{nit} is assumed to be distributed iid extreme value. The assumption of an iid extreme value distribution for this additive error term makes the model a mixed logit instead of another type of choice model, such as random coefficient probit. In each choice situation, each customer chooses the alternative that provides the greatest utility.

1.2 *Classical Estimation*

The parameters b and W are considered fixed, representing the true mean and covariance of the β_n's in the population. These parameters are estimated on a sample of customers drawn from the population. The estimators, denoted $\hat{b}$ and $\hat{W}$, are stochastic due to sampling and are obtained using maximum simulated likelihood estimation. The likelihood function is $L(b,W)=\prod_n L_n(b,W)$ where $L_n(b,W)$ is the probability of customer n's sequence of choices given b and W. Since e_{nit} is iid extreme value, this probability is an integral over β_n of a product of logits. Simulation approximates this integral, using draws of β_n from a normal distribution with mean b and covariance W. The estimator has an asymptotic distribution that is used as the approximate sampling distribution in finite samples, as given by McFadden and Train (forthcoming). We denote this distribution as $f(\hat{b},\hat{W})$.

For any b and W, the density of β_n conditional on customer n's sequence of choices is $G_n(\beta_n \mid b,W) = L_n(\beta_n)\, N(\beta_n \mid b,W) \,/\, L_n(b,W)$, where $N(\beta_n \mid b,W)$ is the normal population density with mean b and covariance W, and $L_n(\beta_n)$ is the probability of the customer's sequence of choices conditional on β_n. The expectation of this density, labeled $E(\beta_n \mid b,W)$, is approximated through simulation, by taking draws of β_n from $N(\beta_n \mid b,W)$, weighting each draw by the ratio $L_n(\beta_n) / L_n(b,W)$, and averaging the results. The estimator for b and W provides the estimator $\hat{E}(\beta_n) = E(\beta_n \mid \hat{b},\hat{W})$. The sampling distribution of $\hat{E}(\beta_n)$ can be approximated by taking

[3] T can be as low as 1. J and T can vary over customers, though we suppress the notation for this possibility.

draws of $\hat{b}$ and $\hat{W}$ from their sampling distribution $f(\hat{b},\hat{W})$ and calculating $E(\beta_n|\hat{b},\hat{W})$ for each draw.

$\hat{E}(\beta_n)$ can be viewed in either of two ways in classical estimation. First, $G_n(\beta_n | b,W)$ can be considered the density of β_n in the subpopulation of customers who, when facing the sequence of choice situations described by X_{nit} for all i and t, make the choices that customer n made. Then $E(\beta_n| b,W)$ is the mean β_n within this subpopulation, and $\hat{E}(\beta_n)$ is an estimator of this mean. Under this interpretation, the number of choice situations, as well as their characteristics as defined by the X_{nit}'s, are considered fixed. Second, one can consider the choice situations to be sampled from a universe of possible choice situations, such that T can rise in the same way that the number of sampled customers can rise. Under this view, $\hat{E}(\beta_n)$ is an estimator of customer n's coefficient vector, β_n. Importantly, the mean of a likelihood function that is expressed as a density is asymptotically equivalent to the maximum likelihood estimator of that likelihood function. Therefore, $\hat{E}(\beta_n)$ is asymptotically equivalent to the maximum likelihood estimator of β_n.

1.3 *Bayesian Estimation*

Under a Bayesian framework, b and W are considered stochastic from the researcher's perspective. The researcher has a prior distribution on b and W, denoted $p(b,W)$, and combines this prior with the likelihood function of the data to obtain a posterior distribution. The joint posterior for b, W, and β_n for all n is proportional to $\prod_n L_n(\beta_n)\, N(\beta_n| b,W)\, p(b,W)$. Draws from this joint posterior distribution are obtained through Gibbs sampling. That is, a sequence of conditional draws is obtained, where each parameter is drawn conditional on a draw from the other parameters. For a draw of b and W, the conditional posterior for density of β_n is $G_n(\beta_n | b,W) = L_n(\beta_n)\, N(\beta_n| b,W) / L_n(b,W)$. Note that this conditional posterior is the same as the conditional density of β_n used in the classical approach. The two approaches use the same

density for β_n given values of b and W. The two approaches only differ in the values of b and W from which the density of β_n is derived.

In Gibbs sampling, draws of b are obtained from its posterior conditional on draws of W and β_n for all n. When β_n is normally distributed, as we have assumed, then the population conditional posterior for b is normal with mean equal to the average of β_n over n and covariance W/N, where N is the sample size. Drawing from this distribution is easy. Similarly, draws of W are obtained from its posterior conditional on b and β_n for all n. When β_n is normally distributed, and the prior on W is inverted Wishart, the conditional posterior for W is also inverted Wishart, which is easy to simulate (Gelman et al. 1995). The simplicity of drawing from these posteriors is one of the main computational advantages of Gibbs sampling in the Bayesian approach. In fact, the reason β_n is assumed to be normally distributed is that, with this assumption, priors on b and W can be specified that give easy-to-draw-from posteriors.

The Gibbs sampling provides a set of draws of β_n from its posterior. The mean of these draws is denoted $\tilde{E}(\beta_n)$. It is the Bayesian analog of $\hat{E}(\beta_n)$ under the classical approach. In the next section, we calculate and compare $\tilde{E}(\beta_n)$ and $\hat{E}(\beta_n)$.

2. Application

Many states, including California, Pennsylvania, and Massachusetts, allow households to choose the company from which they buy electricity. We examine the factors that affect 361 residential customers' choice of electricity supplier, using choice-based conjoint. Surveyed customers made 12 choices each from among four suppliers that differed on the basis of five relevant attributes:

- Fixed price, in cents per kilowatt-hour (7 or 9 cents per kWh)
- Length of contract (0, 1, 2 or 3 years)

- Type of company (the local utility, a "well-known company other than the local utility", or "an unfamiliar company".)
- Time-of-use rates, with 11 cents per kWh from 8am-8pm and 5 cents per kWh from 8pm-8am
- Seasonal rates, with 10 cents per kWh in summer, 8 cents in winter, and 6 cents in spring and fall.

Details on the sample and survey are provided Electric Power Research Institute (1998)[4]

Partworths are estimated for price, contract length, and indicator variables for whether the company was the local utility, a well-known company other than the local utility, time-of-use rates, and seasonal rates. The "unfamiliar company" is taken as the base for normalization, so that the partworths for the local utility and a well-known company are the values of these kinds of companies relative to it. Price and contract length are "linearized," in that the same partworth is applied for each one-unit increase in the variable (i.e., one-cent increase in price, or one-year increase in contract length.) The prices under time-of-use and seasonal rates did not vary in the experiments; consequently, their partworths indicate the values of these rates, including the negative value of the specified prices. As stated in section 1, we assume that all partworths are normally distributed across the population, with a full covariance matrix.

We estimate the expected partworths for each customer, conditional on that customer's observed choices. In the Bayesian procedure, this statistic is the mean of the draws of β_n for that customer. In the classical procedure, this statistic is the mean of the conditional distribution of β_n for the customer, based on the maximum likelihood estimates of the population distribution. With both procedures, the twelfth choice situation was excluded from estimation and used to test forecasts based on the expected partworths for each customer.

[4] We are grateful to Ahmad Farqui, of EPRI, for allowing us to use these data.

Table 1 displays the sample average of the expected partworths under the two procedures. Column 1 gives the average over customers of the classical estimate for each customer and column 2 gives the corresponding average of the Bayesian estimates. The two sets of estimates are quite similar. The scale of the Bayesian estimates is slightly higher than that of the classical estimates. To account for the scale difference, the third column of Table 1 gives the average for the classical estimates scaled such that the average price coefficient is the same for the two procedures. With this rescaling, the two sets of averages are remarkably close.

Table 2 displays the standard deviation of the expected partworths over sampled customer. These standard deviations are similar, and when the scale difference is accounted for, the two sets of standard deviations are very close.

Table 3 arrays the correlation matrix for the vector of expected partworths under both approaches. The upper-triangular portion of the matrix gives the correlations for the classical estimates, and the lower-triangular portion gives the correlations for the Bayesian estimates. The corresponding upper and lower figures are quite similar. For example, the correlation between the price coefficient and the contract length coefficient is 0.106 for the classical estimates and 0.104 for the Bayesian estimates. Note that the price coefficient is very highly correlated with the time-of-use and seasonal coefficients. This correlation is expected since these variables are all price-related. Both the Bayesian and classical procedures are more prone to simulation noise when coefficients are highly correlated; stated alternatively, both methods require more draws to obtain a given level of accuracy when there is high correlation among coefficients. It is particularly noteworthy, therefore, that the two methods provide such close estimates in our study despite such high correlations.

Table 4 displays the correlation between estimates under the two methods. For example, the first row gives the correlation over the sampled customers between their expected price coefficient estimated by the Bayesian method and their expected price coefficient estimated by

the classical method. The correlation is 0.975. The correlation is even higher for the other partworths. These correlations are remarkably high, particularly given that each approach is subject to simulation noise. The two methods are giving essentially the same results, aside from the small scale factor.

As stated above, the last choice experiment was retained for testing. We forecasted each customer's choice in their last experiment using the expected partworths. The results are given in Table 5. Each customer's expected partworths were used in a logit formula to calculate the probability for each alternative in the customer's last choice. The average probability for the chosen alternative (where the average is over the sampled customers) is practically the same for both methods: 0.6299 for the Bayesian estimates and 0.6293 for the classical estimates. We also examined the hit rates, the ability of each model to predict the alternative actually chosen out of the four alternatives. Again, the results were nearly the same: the Bayesian estimates resulted in a 71% hit rate, virtually identical to the 72% for the classical procedure. Importantly, the Bayesian and classical methods provide the same prediction for more than 96% of the respondents.

3. Discussion

We have applied both hierarchical Bayes and maximum likelihood estimation procedures to the same random parameter structure. Despite substantial algorithmic differences, the Bayesian and the classical estimates of the individual partworths are virtually identical. Our focus here has been on the estimated parameters for each person, as those are the critical measures used in customer selection, segmentation, and market simulations.

With normal distribution in a mixed logit model, both methods are fairly straightforward to implement. With other specifications, there can be differences in the convenience attached to the computational procedures employed by each approach. The main differences as we see among them are the following.

(1) For the classic case, it can sometimes be difficult to locate the maximum of the likelihood function with some distributions and behavioral models. The likelihood function can have multiple local maxima, and assuring oneself that a local maximum is indeed the global maximum can be computationally difficult. Also, the likelihood function might not be well approximated by a quadratic. The standard numerical maximization procedures work best when the function is close to a quadratic. We have found that the maximization procedures can often fail to find an increase even though a maximum has not been located. This problem does not seem to arise with mixed logits using normal distributions for the coefficients, as we are using in the current paper. However, we have found it to arise with other distributions, particularly log-normals. (See the discussion below about non-normal distributions.) Bayesian procedures have an advantage in these circumstances because the maximum of the likelihood function does not need to be located under these procedures. Rather, draws from the posterior are taken. The average of these draws can be used as a classical estimator that is asymptotically equivalent to the maximum likelihood estimator.

(2) When the dimension of β_n is large, its covariance W has numerous elements. In classical estimation, each element of the upper diagonal of W generates a parameter that utilizes a degree of freedom. Computationally, the derivative of the likelihood function with respect to each element of W must be calculated, such that, with a full W, computation time rises exponentially with the number of coefficients. To maintain a manageable number of parameters, off-diagonal elements of W are often constrained to zero under classical approaches. In contrast, the Bayesian approach can handle a full W almost as easily as a restricted W, as computation time rises much more moderately with the number of parameters.

(3) Identification is less of an issue in Bayesian compared with classical approaches. In the classical estimation, unidentified parameters cannot be estimated. In the Bayesian approach, the prior can provide needed identification, or a flat prior can be specified such that unidentified parameters manifest themselves as flat areas of the posterior.

Reasons (1)-(3) provide motivation for estimation of mixed logits with Bayesian procedures even if a classical perspective is maintained. The computational disadvantage of Bayesian procedures arises from their need to draw from conditional posteriors. When the coefficients are jointly normally distributed, priors can be specified that make the conditional posteriors for b and W easy to draw from. However, changes in the distributional assumptions can be difficult to implement in the Bayesian procedures. In the classical procedure, alternative distributions can easily be specified for the coefficients. Further, some coefficients can be assumed to be fixed while others vary, and different distributions can be specified for different coefficients. For simulated maximum likelihood the only change in the estimation procedure occurs in the line of code that specifies the draws of the coefficients. With Bayesian procedures, the situation is not as easy. For example, suppose one coefficient is assumed to be fixed while the others are jointly normal. This change cannot be implemented by simply setting the variance of the fixed coefficient to zero within the sampling algorithm for drawing the coefficients for each person. The algorithm used with mixed logits (i.e., Metropolis-Hastings) accepts new draws for some customers and rejects new draws for other customers such that the fixed coefficient would differ over customers in one iteration of draws rather than being the same for all customers. Instead, a new layer of conditioning is required in the Gibbs sampling, requiring draws of the fixed coefficient conditional on the mean, covariance, and values of the random coefficients. Similarly, if non-normal distributions are specified for the coefficients, the conditional posteriors for the parameters of these distributions are not normally distributed. By contrast, when β_n is normally distributed, the conditional posterior of b is normal and the conditional posterior of W is inverted Wishart, under appropriate priors. With other distributions, its is not as easy to specify priors that give easy-to-draw-from conditional posteriors. Work in this area is proceeding (e.g., Boatwright, McCulloch and Rossi, forthcoming), but the requirement that the conditional posterior be derivable and easy to draw from is a necessary requirement and a limitation of Bayesian methods.

Our finding of equivalence in expected partworths at the individual level probably won't bring Bayesian and classic thinkers together, because they fundamentally disagree on the meaning of uncertainty. Still, for most marketing applications, our results provide sufficient convergence to permit one to use either method. Those most comfortable with maximum likelihood can bask in its familiar statistics, while those most comfortable with Bayesian concepts of uncertainty may frolic in its waves. From a predictive perspective it does not matter.

References

Allenby, Greg M., and Peter E. Rossi, 1999, "Marketing Models of Customer Heterogeneity," *Journal of Econometrics*, 89(1-2), 57-78.

_____ and James L. Ginter (1996), "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 32, 392-403.

Boatwright, Peter, Robert M. McCulloch, and Peter Rossi, forthcoming, "Account-Level Modeling for Trade Promotions: An Application of a Constrained Parameter Hierarchical Model," *Journal of the American Statistical Association*.

Electric Power Research Institute (1998), "Predicting Customer Choice Among Electricity Pricing Options," Report RR-108864-V2. EPRI, Palo Alto, CA.

Gelman, Andrew; John B. Carlin, Hal S. Stern and Donald B. Rubin (1995), *Bayesian Data Analysis.* Chapman and Hall, London.

Green, Paul E. and V. Srinivasan (1990) "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing, 54,* (October) 3-19.

Lindsey, Jim K. (1996) *Parametric Statistical Inference,* Oxford: Claredon Press.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, and Martin R. Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, Vol. 15, No. 2, 173-91.

McFadden, Daniel L., and Kenneth E. Train, forthcoming, "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics,* forthcoming.

Research Triangle Institute (1997) "Predicting Retail Customer Choice Among Electricity Pricing Alternatives" RTI project # 6773.

Revelt, D., and Kenneth E. Train, (1999), "Customer-Specific Taste Parameters and Mixed Logit," working paper, Department of Economics, University of California, Berkeley. Paper and estimation software on http://elsa.berkeley.edu/users/train/index.html

Sawtooth Software Inc., 1999, "The CBC/HB Module for Hierarchical Bayes Estimation," at Sawtooth Software's website, www. Sawtoothsoftware.com.

Wedel, Michel; Wagner Kamakura, Neeraj Arora, Albert Bemmaor, Jeongwen Chiang, Terry Elrod, Rich Johnson, Peter Lenk, Scot Neslin and Carsten Stig Poulsen (1999) "Discrete and Continuous Representations of Unobserved Heterogeneity in Choice Modeling," *Marketing Letters* 10:3 219-232.

Table 1

Average of Expected Partworths

	Classical	Bayesian	Scaled Classical
Price	-1.044	-1.120	-1.120
Contract Length	-0.260	-0.286	-0.279
Local Utility	2.714	2.821	2.91
Well-known Company	2.090	2.193	2.24
Time-of-use Rates	-9.894	-10.54	-10.61
Seasonal Rates	-10.11	-10.84	-10.84

Table 2
Standard Deviation of Expected Partworths

	Classical	Bayesian	Scaled Classical
Price	0.693	0.784	0.743
Contract Length	0.380	0.419	0.407
Local Utility	1.983	2.044	2.127
Well-known Company	1.381	1.468	1.481
Time-of-use Rates	6.643	7.061	7.125
Seasonal Rates	6.054	6.615	6.493

Table 3
Correlation Matrix of E (b)
Classical estimates in upper triangle. Bayesian estimates in lower triangle

	Price	Contract Length	Local Utility	Well-known Company	Time-of-use Rates	Seasonal Rates
Price	1	0.106	0.649	0.523	0.903	0.942
Contract Length	0.104	1	0.335	0.245	0.083	0.036
Local Utility	0.667	0.290	1	0.828	0.625	0.615
Well-known Company	0.466	0.205	0.789	1	0.488	0.466
Time-of-Use Rates	0.898	0.116	0.651	0.433	1	0.937
Seasonal Rates	0.932	0.057	0.644	0.411	0.943	1

Table 4
Correlation between Classical and Bayesian Estimates

Price	.975
Contract length	.988
Local utility	.985
Well-known company	.978
Toll rates	.981
Seasonal rates	.979

Table 5: Prediction for Last Choice Situation

	Classical	Bayesian
Average Probability of the chosen alternative	0.6293	0.6299
Number of Matches out of 361	261	256