

Comparing Hierarchical Bayes and Latent Class Choice: Practical Issues for Sparse Data Sets

By Paul Richard McCullough

ABSTRACT

Choice-based Conjoint is the most often used method of conjoint analysis among today's practitioners. In commercial studies, practitioners frequently find it necessary to design complex choice experiments in order to accurately reflect the marketplace in which the client's product resides. The purpose of this paper is to compare the performance of several HB models and LCC models in the practical context of sparse real-world data sets using commercially available software. The author finds that HB and LCC perform similarly and well, both in the default and more advanced forms (HB with adjusted priors and LCC with C factors). LCC may estimate parameters with slightly less bias and HB may capture more heterogeneity. Sample size may have more potential to improve model performance than using advanced forms of either HB or LCC.

INTRODUCTION

Choice-based Conjoint is the most often used method of conjoint analysis among today's practitioners. In commercial studies, practitioners frequently find it necessary to design complex choice experiments in order to accurately reflect the marketplace in which the client's product resides. Studies with large numbers of attributes and/or heavily nested designs are common. However, the number of tasks available for an individual respondent is limited not only by respondent fatigue but also project budgets. Estimating a fairly large number of parameters with a minimum number of choice tasks per respondent can, and in practice often does, create a sparse data set.

Disaggregate choice utility estimation is typically done using Hierarchical Bayes (HB), even with sparse data sets. Jon Pinell and Lisa Fridley (2001) have shown that HB performance can degrade when applied to some partial profile designs. Bryan Orme (2003) has shown that HB performance with sparse data sets can be improved by adjusting the priors.

Latent Class Choice Models (LCC) are an alternative to HB that, for sparse data sets, may offer the practitioner potentially significant managerial as well as statistical advantages:

More managerial insight:

- Powerful market segmentation
- Identification of insignificant attributes and/or levels
- Identification of class independent attributes

More parsimonious; less overfitting in the sense that statistically insignificant parameters can be easily identified and omitted

MAEs and hit rates equal or nearly equal to that of HB

Further, Andrews, et al. (2002) raise the possibility that LCC models may capture more respondent heterogeneity than HB models given sufficiently sparse data at the respondent level.

However, in commercial practice, LCC may have some limitations:

- Computation time and computer capacity
- Real-time to final results (despite some claims to the contrary, LCC models can be both computationally and real-time intensive)
- Required expertise

OBJECTIVES

The purpose of this paper is to compare the performance of HB models and LCC models in the practical context of sparse real-world data sets using commercially available software. Of particular interest is whether or not LCC models capture more heterogeneity than HB models with these data sets.

The software used for this analysis was Sawtooth Software's CBC/HB Version 4.6.4 (2005) and Statistical Innovation's Latent Gold Choice 4.5 (Vermunt and Magidson, 2005).

STUDY DESIGN

Using three commercial data sets, model performance for utilities based on versions of HB and LCC will be compared. One data set will be based on a partial profile choice design. Another data set will be based on a heavily nested attribute-specific design.

Utilities will be estimated using the following techniques:

Default HB- Sawtooth's HB module with default settings

MAE Adjusted Priors HB- Sawtooth's HB module with prior variance and prior degrees of freedom of covariance matrix tuned to optimize holdout MAE

Hit Rate Adjusted Priors HB- Sawtooth's HB module with prior variance and prior degrees of freedom of covariance matrix tuned to optimize holdout hit rate

Default LCC- Statistical Innovation's Latent Gold Choice

CFactor LCC - Statistical Innovation's Latent Gold Choice with one continuous factors

Aggregate Logit Model- Estimated within Sawtooth's SMRT module

Note: While the utilities for default HB were estimated by simply running the choice data through Sawtooth's HB program, the default LCC estimation routine included various manual adjustments to the model based on output diagnostics, e.g., the omission of all statistically insignificant parameters, designating certain attributes as class independent, merging attribute effects across selected classes, constraining certain class parameters to zero, etc. Thus, the amount of effort and expertise for the two "default" approaches differs substantially.

Adjusted Priors HB is HB with prior variance and degrees of freedom of the covariance matrix adjusted either up or down from the default. Adjusting the priors has the effect of either increasing or decreasing the influence of the upper model over the lower or subject-level models. With sparse data sets, it has been shown (Orme, 2003) that increasing the weight of the upper model improves model performance. A grid search is undertaken to find the optimal combination of prior variance and degrees of freedom weights to input as priors. Optimal is defined to be either minimizing holdout MAE or maximizing hit rate.

Statistical Innovation’s Latent Gold Choice program allows the user to introduce one or more continuous factors (Cfactors) to overlay on parameter estimates (Magidson and Vermunt, 2007). Cfactors have the effect of distributing heterogeneity continuously across respondents. A Cfactor on which only the intercept loads, for example, creates a unique random intercept for each respondent. For the models reported here, one Cfactor was used.

Aggregate Choice Model is defined to be the choice model estimated when all respondents are pooled to estimate one set of parameters. For this paper, the aggregate choice model was estimated using Sawtooth Software’s SMRT program. No additional parameters, such as interaction effects or cross effects, were included in the aggregate model. The purpose of the aggregate model was not to build the best aggregate model possible but to be a “worst case” reference point.

Models were compared using these diagnostics:

Tuned MAEs

Disaggregate MAEs-fixed tasks

Disaggregate MAEs-random tasks

Hit Rates-fixed tasks

Hit Rates-random tasks

Average Holdout Variance Ratio-Variance per alternative averaged across holdout tasks; actual divided by predicted

Fixed tasks refer to holdout tasks. Each of the three data sets had at least one holdout task which was not used in the estimation of the utilities. Random tasks are the choice tasks that were used in the estimation of utilities. It was the case for all three data sets that the random tasks varied across respondents. All of the data collected in the three data sets reported here were collected online, using Sawtooth Software’s SSI Web software. The fixed tasks did not vary across respondents.

MAEs

Mean Absolute Error (MAE) is an aggregate measure of how well a model predicts choices. As illustrated in Table 1 below, MAE is calculated as the absolute difference between actual choice task alternative share and predicted share, averaged across all alternatives.

Table 1.

	Raw	Predicted	Delta
Alt #1	20%	33%	13%
Alt #2	30%	33%	3%
Alt #3	50%	33%	17%
Sum of Errors			33%
MAE			11%

Disaggregate MAEs is a disaggregate measure of model performance. The calculation is similar to that of MAE except the calculation is done at the respondent level rather than aggregate.

Table 2.

Resp#1	Raw	Predict	Delta
Alt #1	0	33%	33%
Alt #2	0	33%	33%
Alt #3	1	33%	67%
Sum of Errors			133%
MAE			44%

Exponential tuning is an aggregate method of empirically adjusting for the net effect of scale factors from within the simulator. With exponential tuning, all utilities are multiplied by a constant. Constants less than 1 flatten simulated preference shares and constants greater than 1 heighten simulated preference share differences. The constant is adjusted to minimize MAEs of holdout tasks.

HIT RATES

Hit rates are defined to be the percentage of times the alternative in a task (fixed or random) with the largest predicted purchase probability is the alternative selected by the respondent.

AVERAGE HOLDOUT VARIANCE RATIO

Average Holdout Variance Ratio is defined to be the average variance across the population for each alternative in the holdout task(s) divided by the average variance of the predicted choices. These average variances are calculated by first calculating the variance across the population for each alternative in the holdout task(s). These alternative-specific variances are then averaged (see Table 3 below).

Table 3.

	Actual	Predicted
A	33	45
B	28	38
C	35	50
AHV	32	44
AHVR	0.73	

The purpose of this diagnostic is to measure captured heterogeneity. The goal of any predictive model is not to manufacture a large amount of variance. The goal of the model is to reflect and replicate the true variance from the population. If the model is working well, that is, if the model is capturing heterogeneity, the average holdout variance ratio should be near 1.

Of the four model diagnostic measures, MAE, DMAE, hit rate and AHVR, all but MAE will reflect captured heterogeneity to some degree.

DATA SETS

Data set # 1 is from the biotech industry and is B2B:

634 respondents

Assumed heterogeneous:

- All scientists who analyze organic materials
- Variety of analytic tools used
- Variety of research areas of interest (pharma, environ, etc.)
- Variety of company types (research lab, QC, contract, etc.)
- Purchasing authority/non-authority
- Large budgets/small (\pm \$100,000)

9 attributes, 34 levels, full profile, no prohibitions or nesting

8 choice tasks per respondent; 3 alternatives per task; 9 attributes per alternative

Data set # 2 is from the consumer electronics category and is B2C:

1,231 respondents

Assumed heterogeneous (study purpose was segmentation):

- Broad consumer profile:
 - 16-64 years of age
 - \$45,000 +
 - Own or lease a car

27 attributes, 69 levels, attribute-specific design

12 choice tasks per respondent; 6 alternatives per task; up to 12 attributes per alternative

Data set # 3 is also from the consumer electronics category and is B2C:

301 respondents

Assumed heterogeneous:

- 28-54 years of age
- White collar, professional, educator, business owner
- Work on a laptop
- Income \$80,000 or more

15 attributes, 45 levels, partial profile design

12 choice tasks per respondent; 3 alternatives per task; 7 attributes per alternative

To characterize these data sets:

Data set # 1 is not sparse

Data set # 2 is sparse with a large sample

Data set # 3 is sparse with a small sample

RESULTS

Overall, all HB models and all LCC models performed similarly and well. Referring to Table 4., all disaggregate models outperformed the aggregate model, with the exception of default HB, data set # 3 and the MAE measure. Recall data set # 3 was the most “sparse” in the sense that there were a large number of attribute and levels, the design was partial profile and sample size was relatively small. Also recall that Pinnell and Fridley (2001) got a similar result for some partial profile designs.

All disaggregate models had excellent MAEs and acceptable hit rates. From a practical perspective, if a practitioner estimated any one of these disaggregate models, including default HB, and saw these MAEs and hit rates, he/she would likely be pleased.

Overall, the two LCC models had superior MAEs and two of the HB models (default and hit rate-tuned priors HB) had superior hit rates. This may indicate that LCC parameter estimates have less bias and HB may capture more heterogeneity.

Also note that tuning the priors to MAEs and tuning to hit rates sometimes yielded different results. In both data sets # 1 and # 2, hit rate-tuned HB performed better than MAE-tuned HB. MAEs for the two techniques were similar but hit rates were substantially better for hit rate-tuned HB. In the third data set, the priors were the same for the two approaches.

Sample size appears to have a significant impact on model performance, particularly hit rate, a measure of captured heterogeneity. Data set # 2 had a sample size of 1,231. Hit rates for all disaggregate models were significantly higher than for the other two data sets. Further, DMAEs were substantially lower and AHVRs were noticeably closer to 1 (Table 5).

Finally, LCC models, while demonstrating comparable performance to the HB models, did so occasionally with much more parsimony. The Cfactor LCC model for data set # 2 used only 8 of the 27 attributes. For data set # 3, the most sparse of the three data sets, the LCC models used 13 of the 15 total attributes.

Table 4.

		Aggr HB	Default HB	Adjusted Priors HB (MAE)	Adjusted Priors HB (Hit Rate)	Default LC	Cfactor LC
Data Set #1	MAE tuned	2.53	2.44	1.98	2.19	1.88	1.75
	Hit Rate- Fixed Tasks	53.9%	64.0%	55.8%	65.0%	60.1%	65.0%
	Attributes/ evels	9/34	9/34	9/34	9/34	9/34	9/34
Data Set #2	MAE tuned	1.30	0.91	0.79	0.86	0.61	0.82
	Hit Rate- Fixed Task	32.4%	75.7%	69.1%	76.9%	68.9%	73.0%
	Attributes/ evels	27/69	27/69	27/69	27/69	16/47	8/28
Data Set #3	MAE tuned	1.52	2.09	0.95	0.95	0.62	0.22
	Hit Rate- Fixed Task	50.8%	61.8%	65.1%	65.1%	62.5%	62.8%
	Attributes/ evels	15/45	15/45	15/45	15/45	13/40	13/40

Table 5 below lists hit rates, DMAEs and Average Holdout Variance Ratios; three measures of captured heterogeneity. For all three data sets, both default HB and hit rate-tuned HB capture more heterogeneity than either LCC model, relative to the three measures listed, with the exception of default HB, hit rate and data set # 3.

Table 5.

		Aggr HB	Default HB	Adjusted Priors HB (MAE)	Adjusted Priors HB (Hit Rate)	Default LC	Cfactor LC
Data Set #1	Hit Rate- Fixed Tasks	53.9%	64.0%	55.8%	65.0%	60.1%	65.0%
	DMAE- Fixed Tasks	29.07	19.03	27.03	19.58	25.51	22.82
	Variances Ratio	n/a	1.38	4.11	1.43	2.54	1.56
Data Set #2	Hit Rate- Fixed Task	32.4%	75.7%	69.1%	76.9%	68.9%	73.0%
	DMAE- Fixed Task	21.70	7.26	13.70	7.34	12.45	14.29
	Variances Ratio	n/a	1.03	1.15	1.03	1.02	1.03
Data Set #3	Hit Rate- Fixed Task	50.8%	61.8%	65.1%	65.1%	62.5%	62.8%
	DMAE- Fixed Task	33.35	20.49	21.84	21.84	27.02	26.17
	Variances Ratio	n/a	1.03	1.09	1.09	1.81	1.58

Table 6 compares hit rates for holdout tasks and hit rates for random tasks. Random task hit rates for the HB models approach 100% and are dramatically higher than holdout task hit rates. Random task hit rates for the LCC models are comparable to holdout task hit rates.

Table 7 compares DMAEs for holdout tasks and DMAEs for random tasks. Similarly to Table 6 data, random task DMAEs for the HB models are dramatically lower than holdout task DMAEs. Random task DMAEs for the LCC models are comparable to holdout task DMAEs.

For data set # 1, the LCC models used all available attributes. However, three attributes were class independent and three others had one or more class parameters dropped. For data set # 3, the LCC models used 13 of 15 available attributes. Additionally, one attribute was class independent and one class parameter was omitted.

Given the dramatic improvement in hit rate and DMAE for the HB random task measures relative to holdout task measures and the relative parsimony of the LCC models, it appears that HB may overfit the data.

**Table 6.
Hit Rate**

Hit Rates		Aggr HB	Default HB	Adjusted Priors HB (MAE)	Adjusted Priors HB (Hit Rate)	Default LC	Cfactor LC
Data Set #1	Hit Rate-Fixed Tasks	53.9%	64.0%	55.8%	65.0%	60.1%	65.0%
	Hit Rate-Random Tasks	47.3%	99.3%	71.9%	98.2%	61.0%	69.4%
	Attributes/levels	9/34	9/34	9/34	9/34	9/34	9/34
Data Set #2	Hit Rate-Fixed Task	32.4%	75.7%	69.1%	76.9%	68.9%	73.0%
	Hit Rate-Random Tasks	35.9%	96.3%	80.1%	94.6%	74.4%	78.1%
	Attributes/levels	27/69	27/69	27/69	27/69	16/47	8/28
Data Set #3	Hit Rate-Fixed Task	50.8%	61.8%	65.1%	65.1%	62.5%	62.8%
	Hit Rate-Random Tasks	45.7%	98.6%	88.7%	88.7%	55.2%	61.2%
	Attributes/levels	15/45	15/45	15/45	15/45	13/40	13/40

Table 7.
DMAEs

DMAEs		Aggr HB	Default HB	Adjusted Priors HB (MAE)	Adjusted Priors HB (Hit Rate)	Default LC	Cfactor LC
Data Set #1	DMAE-Fixed Tasks	29.07	19.03	27.03	19.58	25.51	22.82
	DMAE-Random Tasks	32.75	1.78	25.86	4.89	26.72	22.30
	Attributes/levels	9/34	9/34	9/34	9/34	9/34	9/34
Data Set #2	DMAE-Fixed Task	21.70	7.26	13.70	7.34	12.45	14.29
	DMAE-Random Tasks	21.49	1.89	11.81	2.85	10.77	9.15
	Attributes/levels	27/69	27/69	27/69	27/69	16/47	8/28
Data Set #3	DMAE-Fixed Task	33.35	20.49	21.84	21.84	27.02	26.17
	DMAE-Random Tasks	33.51	3.23	14.45	14.45	28.17	25.79
	Attributes/levels	15/45	15/45	15/45	15/45	13/40	13/40

CONCLUSIONS

Default HB is by far the easiest of the examined models to build yet it performed nearly as well as more sophisticated models even though two of the three data sets used in the analysis were quite sparse.

HB and LCC perform similarly and well, both in the default and more advanced forms (HB with adjusted priors and LCC with Cfactors).

At least for these sparse data sets, LCC may estimate parameters with slightly less bias and HB may capture more heterogeneity.

Sample size may have more potential to improve model performance than using advanced forms of either HB or LCC.

Tuning HB priors to hit rates, rather than MAEs, appears to be the more productive approach.

DISCUSSION

HB and LCC may be reaching the limit of their potential. Both models, despite sophisticated adjustments performed similarly to each other and also similarly to their default versions. Further advances may need to come from a different source. For example, perhaps changing the way questions are asked may yield higher quality data which would, in turn, improve model performance.

For naive users, default HB seems clearly to be the preferred method. It requires virtually no tweaking, it is extremely simple to run and generates adequate results.

If however, the user is more advanced and either requires a segmentation as well as choice utilities, or is interested in building a parsimonious model (and the managerial insight that parsimony yields), LCC offers a viable alternative.

A word of caution, however, regarding LCC models. LCC models run fairly quickly if no Cfactors are included. Including Cfactors increases computation substantially. Models that may have taken a couple minutes to run might take a couple hours with Cfactors included. If the modeler wishes to additionally model scale factor lamda, which can be done with the Latent Gold Choice Syntax Module, run times might increase to 10-12 hours. In the commercial world, these run times may occasionally prove impractical.

ADDITIONAL READING

Andrews, Rick L. Andrew Ainslie and Imran S. Currim (2002), *An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity*, Journal of Marketing Research (November), 479-87

Magidson, Jay and Tom Eagle (2005), *Using Parsimonious Conjoint and Choice Models to Improve the Accuracy of Out-of-Sample Share Predictions*, ART Forum, American Marketing Association, Coeur D'Alene, ID

Magidson, Jay and Jeroen Vermunt (2007), *Use of a Random Intercept in Latent Class Regression Models to Remove Response Level Effects in Ratings Data*, Bulletin of the International Statistical Institute, 56th Session, paper #1604, 1-4

Orme, Bryan (2003), *New Advances Shed Light on HB Anomalies*, Sawtooth Software Research Paper Series, Sawtooth Software, Inc., Sequim, WA

Orme, Bryan and Peter Lenk (2004), *HB Estimation for "Sparse" Data Sets: The Priors Can Matter*, ART Forum, American Marketing Association, Whistler, BC

Pinnell, Jon and Lisa Fridley (2001), *The Effects of Disaggregation with Partial Profile Choice Experiments*, Sawtooth Software Conference, Victoria, BC

Sawtooth Software (2005), *The CBC/HB System for Hierarchical Bayes Estimation Version 4.0 Technical Paper*, accessible from www.sawtoothsoftware.com/download/techpap/hbtech.pdf.

Vermunt, J. K. and J. Magidson (2005), *Technical Guide for Latent GOLD Choice 4.0: Basic and Advanced*, Belmont Massachusetts: Statistical Innovations Inc.