Sawtooth Software

RESEARCH PAPER SERIES

A Comparison of HB-MNL Estimation via Metropolis Hastings (Sawtooth Software's CBC/HB Program), Hamiltonian Monte Carlo (via Stan), and Variational Bayes (via Stan)

Bryan Orme
Sawtooth Software, Inc.

3210 N. Canyon Rd., Provo, Utah
+1 801 477 4700
www.sawtoothsoftware.com

A Comparison of HB-MNL Estimation via Metropolis Hastings (Sawtooth Software's CBC/HB Program), Hamiltonian Monte Carlo (via Stan), and Variational Bayes (via Stan)

Bryan Orme, Sawtooth Software
January 2024

(This short article summarizes main findings from a more complete white paper by Kevin van Horn, Bayesium Analytics. Kevin van Horn has done work in Bayesian inference and Bayesian computation since the late 90's, with applications in speech recognition, market research, and time series forecasting. He spent six years at market research firm The Modellers designing, implementing, and improving a variety of Bayesian models and related simulations and experimental design software.)

Individual-level estimation of utility values is the norm for Choice-Based Conjoint (CBC) since the early 2000s. There are a variety of algorithms for doing so, such as mixed logit, HB-MNL (Hierarchical Bayesian Multinomial Logistic Regression) using Metropolis Hastings, Hamiltonian Monte Carlo, and Variational Bayes. Most methods provide very similar results, but there are interesting questions around speed, convergence, and predictive validity for the competing methods.

Kevin van Horn of Bayesium Analytics recently compared Hamiltonian Monte Carlo (HMC, with no u-turn sampling—NUTS), Metropolis Hastings (MH), and Variational Bayes (VB) for estimating individual-level MNL utilities for Choice-Based Conjoint (CBC) experiments (van Horn, 2024). For his MH runs, van Horn used Sawtooth Software's CBC/HB program. For VB and HMC, van Horn used Stan.

Hamiltonian Monte Carlo uses what van Horn describes as an approximate Hamilton dynamics simulation. Variational Bayes computes a multivariate normal approximation to the posterior and is extremely fast. MH is the algorithm incorporated into HB-MNL that Sawtooth Software has been applying since the late 1990s, as originally recommended to us by leading academics Greg Allenby and Peter Lenk.

Van Horn examined six datasets derived from three CBC datasets (Cruise, HDTV, and an alternative-specific CBC). He created a second dataset derived from each of the original three data sets, leading to six CBC datasets in total. In one case, he expanded the sample size to 6000 by creating artificial respondents drawn from the same (inferred) distribution as the original ones. In two cases he significantly reduced the number of choice tasks to see what would happen in sparse CBC cases.

The basic characteristics of the six CBC datasets were as follows:

	#Resps	**#Tasks**	**Estimated Parameters**	**None Included**
Cruise	600	15	21	No
Cruise-Large	6000	15	21	No
HDTV	2939	12	15	Yes
HDTV-Sparse	2939	4	15	Yes
Alt-Spec	2940	22	27	No
Als-Spec-Sparse	2940	6	27	No

The Cruise dataset features a cell of respondents completing the same version (block) of the questionnaire that may be used for out-of-sample holdout validation. HDTV was fielded with four versions (blocks), allowing for a four-fold out-of-sample validation. The Alt-Spec data set did not have holdout respondents, so van Horn created simulated training and holdout data sets using lower-level part worths estimated from the original data set to evaluate the predictive performance of the different Bayesian algorithms.

Procedures:

Van Horn used defaults in Sawtooth Software's CBC/HB v5.7 software program (prior D.F.=5, prior variance=1, acceptance rate=0.3). Half of the iterations were burn-in, half were "used". The total number of iterations was chosen to approximately match the runtime of the Stan MHC models, with an additional minimal run of 10K burn-in / 10K "used" for comparison.

Van Horn describes HMC as follows: "This is an MCMC algorithm in which each model variable is augmented with a 'momentum' variable. It uses an approximate Hamilton dynamics simulation (which requires the gradient of the density function) which is then corrected by performing a Metropolis acceptance step. These experiments use the HMC variant known as the No-U-Turn Sampler, which automates many of the HMC algorithmic settings." (Van Horn 2024)

Van Horn describes VB as follows: "This computes a multivariate normal approximation to the posterior by minimizing the KL divergence from the approximation, to the posterior. In practice this minimization is carried out by maximizing the ELBO (evidence lower bound), which is equal to the negative KL divergence up to a constant that depends only on the data. The ELBO itself is approximated using a random sample from the approximating distribution." (Van Horn 2024)

Van Horn examined holdout predictive accuracy as well as convergence metrics. Convergence metrics were the Rhat convergence diagnostic and effective sample size (ESS) for HMC and MH, for each element of the population mean estimate of main effects as well as the diagonal of the population-level covariance matrix. Rhat and ESS convergence tests preferably involve running and comparing multiple chains of the estimation from different random starting points, which van Horn did for his analysis.

No utility constraints were applied for this comparison of the three methods of estimating individual-level utilities for CBC experiments. More details are available in van Horn 2024.

Main Findings:

HMC and MH lead to equally good predictions of holdouts and equivalent posterior estimates of parameters given the same amount of runtime. It should be noted that the number of K-1 coded (effects-coded) parameters estimated for these six datasets were 15, 21, and 27. So, van Horn's examination involved modestly sized CBC studies.

HMC requires warm-up iterations where the iterations run much slower than in later stages (iterations) of the algorithm. Thus, for small CBC datasets, HMC could potentially run slower to achieve equivalent results compared to MH (CBC/HB software implementation).

VB is considerably faster than either HMC or MH, with essentially equally good predictions of holdouts for Cruise and HDTV data sets. It performed slightly worse for the alt-spec dataset. However, across multiple runs for the alt-spec dataset, it could lead to anomalous results (get stuck in local minima) with bad predictive accuracy. Thus, with VB, it would be important to run it multiple times to ensure that a local minima was avoided.

Simulating on posterior lower-level draws led to more accurate predictions of holdout shares of choice than simulating on point estimates. Both we and van Horn speculate that this could be due partly to difference in scale factor (simulating on draws typically leads to lower scale factor in predicted shares of choice).

HMC led to better convergence diagnostics according to Rhat and Effective Sample Size (ESS). ESS is a measure of the quality and independence of the posterior draws (lower autocorrelation, better representation of the full posterior distribution, leading to better estimation of the mean). Despite the better formal convergence diagnostics in favor of HMC over MH, the simulated predictions of holdouts were essentially equivalent between the two.

Van Horn also compared posterior estimates (of means and standard deviations) for the three algorithms. He found a good match between HMC and MH. VB didn't match MH and HMC nearly as well, with VB underestimating the posterior standard deviation and smaller scale.

For these small to modest-sized CBC datasets (in terms of number of parameters to estimate) and using sparse and not sparse versions of the datasets, van Horn found surprisingly good (and essentially equivalent) holdout prediction accuracy between shorter and much longer runs (many more iterations); whereas formal diagnostic criteria would suggest the shorter run models hadn't yet converged.

Caveat and Future Research:

Kevin Lattery (SKIM) has commented in personal correspondence with Sawtooth Software that HMC is better than MH for challenging CBC datasets, particularly when the number of estimated parameters is 100 or more. Van Horne's investigation didn't cover such datasets.

References:

Van Horn, Kevin (2024), "Comparison of Algorithms for Estimating Bayesian HMNL: MH vs. HMC vs. VB." Unpublished white paper.