

Finite Population Estimators in Stochastic Search Variable Selection

BY MERLISE A. CLYDE* AND JOYEE GHOSH†

SUMMARY

Monte Carlo algorithms are commonly used to identify a set of models for Bayesian model selection or model averaging. Because empirical frequencies of models are often zero or one in high dimensional problems, posterior probabilities calculated from the observed marginal likelihoods, re-normalized over the sampled models are often employed. Such estimates are the only recourse in several newer stochastic search algorithms. In this paper, we prove that renormalization of posterior probabilities over the set of sampled models generally leads to bias which may dominate mean squared error. Viewing the model space as a finite population, we propose a new estimator based on a ratio of Horvitz-Thompson estimators which incorporates observed marginal likelihoods, but is approximately unbiased. This is shown to lead to a reduction in mean squared error compared to the empirical or re-normalized estimators, with little increase in computational costs.

Key words: Bayesian model averaging; Inclusion probability, Markov chain Monte Carlo; Median probability model; Model uncertainty; Variable Selection

1. INTRODUCTION

The advent of Markov chain Monte Carlo (MCMC) algorithms greatly expanded Bayesian model selection and model averaging (BMA) in regression problems that precluded enumeration (see Hoeting et al. (1999) and Clyde & George (2004), and references therein). For variable selection, a model M_γ may be represented by a binary vector $\gamma \in \Gamma \equiv \{0, 1\}^p$ of indicators specifying the inclusion/exclusion of the p potential predictors. Posterior inference is based on constructing an aperiodic and positive recurrent Markov chain $\gamma^{(0)}, \gamma^{(1)}, \dots$ on Γ such that the stationary distribution, π , is the posterior distribution

$$\pi \equiv p(M_\gamma | Y) = \frac{p(Y | M_\gamma)p(M_\gamma)}{\sum_{\gamma \in \Gamma} p(Y | M_\gamma)p(M_\gamma)} \quad (1)$$

where $p(M_\gamma)$ is the prior probability of model M_γ and the marginal likelihood of model M_γ is proportional to $p(Y | M_\gamma) = \int p(Y | \theta_\gamma, M_\gamma)p(\theta_\gamma | M_\gamma)d\theta_\gamma$, obtained by integrating the sampling distribution of data Y with respect to the prior distribution of model specific parameters θ_γ . The Monte Carlo (MC) frequencies f_γ of models provide simulation consistent estimates of posterior model probabilities, $\hat{p}^{\text{MC}}(M_\gamma | Y) = \frac{1}{T} \sum_{t=1}^T I\{M_\gamma = M^{(t)}\} = f_\gamma/T$ as the number of iterations of the Markov chain $T \rightarrow \infty$. When marginal likelihoods

* Merlise A. Clyde is Professor of Statistics, Duke University, Durham, NC 27705. Email clyde@stat.duke.edu

† Joyee Ghosh is Assistant Professor, Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242-1409. Email joyee-ghosh@uiowa.edu

are available re-normalized estimates (RN) of posterior probabilities for models may be obtained by replacing Γ in (1) with S_T , the set of unique sampled models. As with MC estimators, models not in S_T have estimated probability zero. The RN estimates provide exact Bayes factors for comparing any two models and have been used in various contexts by Clyde et al. (1996); George & McCulloch (1997); Raftery et al. (1997) and more recently by Scott & Carvalho (2008) and Clyde et al. (2011) in search algorithms where MC estimators are not available. George (1999b,a) suggest that using the re-normalized model probabilities may lead to substantial improvements over the MC estimates.

Recent simulation studies comparing RN and MC estimates, however, have lead to mixed results. Some of the latest stochastic search algorithms which exclusively use RN estimators use adaptive estimates of marginal inclusion probabilities to guide the search for models with high posterior probability, but without ensuring that samples are generated according to the posterior distribution over models (Berger & Molina, 2005; Scott & Carvalho, 2008; Clyde et al., 2011). Heaton & Scott (2010) note that while these search algorithms typically find models with higher marginal likelihoods than standard MCMC algorithms, they paradoxically had poorer performance for estimation of inclusion probabilities than the MC estimates from MCMC. In the context of sampling without replacement using the BAS algorithm, Clyde et al. (2011) found that RN estimates from BAS had much smaller mean squared errors for estimating inclusion probabilities, than either MC or RN estimators from MCMC sampling. Garcia-Donato and Martínez-Beneito (GD-MB) in a 2011 technical report¹ demonstrated in a larger scale simulation that inclusion probabilities based on MC frequencies however were preferable to the RN estimates from either BAS or Berger & Molina.

These results motivate the current work to better understand theoretical properties of the RN estimator in conjunction with the sampling method. We provide a formal proof of why RN estimates from most stochastic search algorithms are biased. We propose an alternative estimator based on the Horvitz-Thompson estimator from finite population sampling, which incorporates marginal likelihoods of visited models as in the RN estimator, but also enjoys an approximate unbiasedness property by taking into account the unequal probability of sampling models. We demonstrate that this estimator may lead to a smaller mean squared error than both MC and RN estimates.

2. ESTIMATION IN BMA

Our goal is to estimate quantities under model averaging of the form

$$\Delta = \sum_{\gamma \in \Gamma} \Delta(M_\gamma) p(M_\gamma | Y) = \frac{\sum_{\gamma \in \Gamma} \Delta(M_\gamma) p(Y | M_\gamma) p(M_\gamma)}{\sum_{\gamma \in \Gamma} p(Y | M_\gamma) p(M_\gamma)} \quad (2)$$

where various choices of $\Delta(M_\gamma)$ lead to posterior model probabilities ($\Delta(M_\gamma) = I(M_j = M_\gamma)$), inclusion probabilities ($\Delta(M_\gamma) = \gamma_j$), predictions ($\Delta(M_\gamma) = \hat{Y}_{M_\gamma}$), etc. If models are generated independently from $p(M_\gamma | Y)$, then the MC estimator \hat{p}^{MC} is shown by GD-BM to be equivalent to the Hansen-Hurwitz (HH) estimator in the sample survey literature (Hansen & Hurwitz, 1943), or more generally

$$\hat{\Delta}^{HH} = \frac{1}{T} \sum_{t=1}^T \frac{\tilde{\Delta}_t}{\tilde{\pi}_t^S} = \sum_{\gamma \in \Gamma} \frac{\tilde{\Delta}_\gamma}{\tilde{\pi}_\gamma^S} \frac{1}{T} \sum_{t=1}^T I(M_t = M_\gamma) = \sum_{\gamma \in \Gamma} \frac{\tilde{\Delta}_\gamma}{\tilde{\pi}_\gamma^S} \frac{f_\gamma}{T} = \sum_{\gamma \in \Gamma} \frac{\tilde{\Delta}_\gamma}{\tilde{\pi}_\gamma^S} \hat{p}^{MC}(M_\gamma)$$

¹ Available at <http://arxiv.org/pdf/1101.4368v1>

where $\tilde{\Delta}_t = \Delta(M_t)p(M_t | Y)$, π_t^S is the probability of sampling model M_t and f_γ is the frequency of model M_γ in the sample; if $\pi_t^S = p(M_t | Y)$, then the estimate is unbiased and reduces to the Monte Carlo estimator, $\hat{\Delta}^{MC} = \sum_{\gamma \in \Gamma} \Delta(M_\gamma) \hat{p}^{MC}(M_\gamma)$. For finite T with MCMC, the expectation depends on the initial design and transition kernel.

PROPOSITION 1. *Given an aperiodic and positive recurrent Markov chain $\gamma^{(0)}, \gamma^{(1)}, \dots$ on Γ , an initial distribution α and transition matrix $P = (p_{ij})$ such that the stationary distribution is $\pi = p(\cdot | Y)$, let \hat{p}^{MC} denote the vector of Monte Carlo frequencies (f_γ/T). Then*

$$E(\hat{p}^{MC})' = \frac{1}{T} \alpha' \sum_{t=1}^T P^t \quad \text{where } p_{ij}^t \equiv P^t.$$

Proof. Let $M^{(t)}$ denote the state of the model $\gamma^{(t)}$ at time t , then

$$\begin{aligned} E\{\hat{p}^{MC}(M_j | Y)\} &= E\left[\sum_{t=1}^T \frac{I\{M^{(t)} = M_j\}}{T}\right] = \frac{1}{T} \sum_{t=1}^T \text{pr}(M^{(t)} = M_j) \\ &= \frac{1}{T} \sum_{t=1}^T \sum_i \alpha_i \text{pr}(M^{(t)} = M_j | M^{(0)} = M_i) = \frac{1}{T} \sum_{t=1}^T \sum_i \alpha_i p_{ij}^t. \end{aligned} \quad (3)$$

Ergodicity implies that $\hat{p}^{MC}(M_j | Y)$ is asymptotically consistent (hence unbiased) as $T \rightarrow \infty$. Furthermore, because Γ is a finite state space, the chain is uniformly ergodic and hence under existence of a second moment $\sqrt{T}(\hat{\Delta}^{MC} - \Delta)$ converges weakly to a normal distribution with mean zero and variance σ_Δ^2 . For finite samples under an arbitrary initial distribution the MCMC estimator is not unbiased, as $\pi_t^S \neq p(M_t | Y)$. While trace plots may suggest “convergence” to the stationary distribution, bias may still be present because the chain may have a low probability of transitioning from one high probability state to another given the initial state; a problem in practice with highly correlated variables in high dimensional spaces and one coordinate at a time update schemes (Nott & Green, 2004). Modifying the transition kernel and increasing the number of iterations of the MCMC algorithm can be used to reduce bias of MC estimates in finite samples.

In finite population sampling, the minimal sufficient statistic is the unordered set of distinct labeled observations (Thompson, 1992, Chapter 3), in this case, the model indices and values $p(M_\gamma | Y)$ or $p(Y | M_\gamma)p(M_\gamma)$. The Hansen-Hurwitz or MC estimator is not a function of the minimal sufficient statistic, however, a Rao-Blackwell estimator with smaller mean squared error may be obtained by taking conditional expectations given the minimal sufficient statistic. Unfortunately, the resulting estimator is difficult to compute (even with independent sampling), and rarely used in practice. An alternative estimator that is a function of the minimal sufficient statistic is the Horvitz-Thompson (HT) estimator (Horvitz & Thompson, 1952). By using HT estimators for the numerator and denominator of (2), we may construct an estimator that is approximately unbiased, a function of the observed marginal likelihoods, and with smaller mean squared error. We first consider ratio estimators to unify the different methods and then discuss how to construct HT estimators in the context of MCMC sampling.

3. RATIO ESTIMATORS

Using the set of unique sampled models S_T , estimators of (2) may be expressed as:

$$r = \frac{\sum_{\gamma \in \Gamma} a_\gamma I(\gamma \in S_T)}{\sum_{\gamma \in \Gamma} b_\gamma I(\gamma \in S_T)} \equiv \frac{\bar{a}}{\bar{b}} \quad (4)$$

for various choices of a_γ and b_γ . We generalize the result of Hartley & Ross (1954) to obtain an exact expression for the bias of ratio estimators as in 4) under general sampling.

THEOREM 1. Let Δ denote the posterior expectation for some function $\Delta(M_\gamma)$

$$\Delta = \frac{\sum_{\gamma \in \Gamma} \Delta(M_\gamma) p(Y | M_\gamma) p(M_\gamma)}{\sum_{\gamma \in \Gamma} p(Y | M_\gamma) p(M_\gamma)} \equiv \frac{\Delta_{num}}{\Delta_{den}}$$

and let r denote a ratio estimator of Δ given by equation (4) with $E(\bar{a})/E(\bar{b}) \equiv \mu_a/\mu_b \equiv \rho$. The bias of the ratio estimator r is

$$E(r) - \Delta = \rho - \Delta - \frac{\text{cov}(r, \bar{b})}{\mu_b}$$

and the absolute relative bias (ARB) is

$$\frac{|E(r) - \Delta|}{s_r} \leq \frac{|\Delta - \rho|}{s_r} + \frac{s_{\bar{b}}}{|\mu_b|}.$$

Proof. Starting with the covariance of r and \bar{b} , $\text{cov}(r, \bar{b}) = E(r\bar{b}) - E(r)E(\bar{b}) = \mu_a - \mu_b E(r)$ we have upon rearranging that $E(r) - \Delta = \rho - \Delta - \text{cov}(r, \bar{b})/\mu_b$. Taking absolute values,

$$|E(r) - \Delta| \leq |\rho - \Delta| + |\text{cov}(r, \bar{b})/\mu_b| \leq |\rho - \Delta| + \frac{s_r s_{\bar{b}}}{|\mu_b|}$$

and dividing by s_r completes the proof. \square

As ratio estimators are not unbiased, efficiency of different methods is made on the basis of mean squared error.

LEMMA 1. A first order approximation to the mean squared error of the ratio estimator in (4) is

$$E(r - \Delta)^2 \approx \frac{E(\bar{a} - \Delta \bar{b})^2}{\Delta_{den}^2}. \quad (5)$$

Proof. The first two terms of a Taylor's series expansion of $f(a, b) = a/b$ about the point $(\Delta_{num}, \Delta_{den})$ lead to

$$r - \Delta \approx \frac{\bar{a} - \Delta_{num} - \Delta(\bar{b} - \Delta_{den})}{\Delta_{den}} = \frac{\bar{a} - \Delta \bar{b}}{\Delta_{den}}.$$

David & Sukhatme (1974) provide justification and bounds to the approximate bias and MSE of ratio estimators. This leads to our main result:

THEOREM 2. Let $\pi^S(M_\gamma)$ denote the probability that model M_γ is included in the set of unique models S_T from a sample of size T . Set $a_\gamma = \Delta(M_\gamma) p(Y | M_\gamma) p(M_\gamma) / \pi^S(M_\gamma)$ and $b_\gamma = p(Y | M_\gamma) p(M_\gamma) / \pi^S(M_\gamma)$. Then the Horvitz-Thompson estimators \bar{a} and \bar{b} are

193 unbiased estimators of Δ_{num} and Δ_{den} , respectively, and the ratio estimator $r = \bar{a}/\bar{b}$ is
 194 approximately unbiased for estimating Δ , with approximate variance (MSE)
 195

$$196 \quad s^2 = E(r - \Delta)^2 \approx \frac{\text{var}(\bar{a} - \Delta\bar{b})}{\Delta_{\text{den}}^2}. \quad (6)$$

197
 198 *Proof.* Since $\bar{a} = \sum_{\gamma \in \Gamma} a_\gamma I(M_\gamma \in S_t) = \sum_{\gamma \in \Gamma} \Delta(M_\gamma) \frac{p(Y|M_\gamma)p(M_\gamma)}{\pi^S(M_\gamma)} I(M_\gamma \in S_t)$, $E(\bar{a}) =$
 199 $\sum_{\gamma \in \Gamma} \Delta(M_\gamma) p(Y | M_\gamma) p(M_\gamma) = \Delta_{\text{num}}$ as $E(I(M_\gamma \in S_t)) = \pi^S(M_\gamma)$. The unbiasedness of
 200 \bar{b} follows similarly. As $E(\bar{a} - \Delta\bar{b})/\Delta_{\text{den}} = 0$, r is approximately unbiased and the approx-
 201 imate MSE (variance) of r is obtained from the linear approximation in Lemma 1. \square
 202

203 From Theorem 2, it is clear that the renormalized estimator is approximately unbiased
 204 only in the case of simple random sampling (with or without replacement) where each
 205 model has an equal probability *a priori* of being selected in the sample, however this
 206 design is seldom used. In the next section, we propose a method for computing HT
 207 estimators in practice in order to reduce bias while incorporating marginal likelihoods
 208 from sampled models using the minimal sufficient statistics.
 209

210 4. RATIO HORVITZ THOMPSON ESTIMATORS FOR MCMC SAMPLING

211 For Monte Carlo sampling from the posterior distribution, the HT weights are the
 212 probability that model M_γ is included in a sample of size T ,
 213

$$214 \quad \pi^S(M_\gamma) = 1 - (1 - p(M_\gamma|Y))^T. \quad (7)$$

215 Direct calculation of $\pi^S(M_\gamma)$ is possible under MCMC sampling, however, it involves cal-
 216 culation of the $2^p \times 2^p$ single step transition matrix, where p is the number of covariates.
 217 The order of the computation will be magnitudes higher than enumerating the model
 218 space, making exact HT estimators impractical.
 219

220 We instead propose an approximation to $\pi^S(M_\gamma)$ based on thinning the chain so that
 221 the remaining T^* samples are approximately independent. This results in little loss of
 222 information as the Horvitz-Thompson estimate uses the unique labels, rather than the
 223 number of repeat visits. Because $p(M_\gamma | Y)$ is known up-to a proportionality constant,
 224 we use a simulation consistent estimate of the normalizing constant
 225

$$226 \quad \hat{C} = \frac{\sum_{t=1}^{T^*} I(M^{(t)} \in A)}{T^*} \frac{1}{\sum_{\gamma \in A} p(Y | M_\gamma) p(M_\gamma)} \quad (8)$$

227 where A is the set of unique models based on running a second independent Markov chain
 228 (George & McCulloch, 1997). Replacing $p(M_\gamma | Y)$ by $p(Y | M_\gamma)p(M_\gamma)\hat{C}$ in (7) and T
 229 by the length of the thinned chain T^* provides a simulation consistent estimate $\hat{\pi}^S(M_\gamma)$
 230 of the model inclusion probabilities, which are then used in Theorem 2 to construct
 231 ratio Horvitz-Thompson estimators for quantities of interest. An estimate of the variance
 232 from (6) may be obtained by using the standard Horvitz-Thompson expressions for the
 233 variance (Thompson, 1992, page 69) using the variable $z_\gamma = (a_\gamma - \hat{\Delta}b_\gamma)/\hat{\Delta}_{\text{den}}$.
 234

235 While the Horvitz Thompson estimators for estimating Δ_{num} and Δ_{den} are unbiased
 236 and functions of the minimal sufficient statistics, minimal sufficient statistics for finite
 237 population sampling are not complete and there is no unique minimum variance (mean
 238 squared error) estimator. Simulation studies provide evidence that the HT estimator
 239 provides reductions in MSE over the MC and RN estimators.
 240

5. SIMULATIONS

We use a simulation design similar to the study in Nott & Kohn (2005), but increase the dimensionality from $p = 15$ to $p = 20$ and introduce two variables with a correlation of 0.99. The first 15 columns of our 50×20 design matrix X are generated exactly as in Nott & Kohn (2005). Columns 16-19 are generated using independent $N(0, 1)$ variables and column 20 is generated to have a 0.99 correlation with column 19. The response is generated as $Y \sim N(\alpha 1 + X\beta, 2.5^2 I)$ where $\alpha = 4$, $\beta = (2, 0, 0, 0, -1, 0, 1.5, 0, 0, 0, 1, 0, 0.5, 0, 0, 0, -1, 1, 4)'$, 1 is a column of ones. For illustration, we use Zellner's g -prior with $g = n$ (Zellner, 1986) for model-specific parameters, which leads to a closed form expression for the marginal likelihood of a model and set $p(M_\gamma) = 1/2^p$. We use a Metropolis-Hastings (MH) algorithm with add/delete steps and random swap proposals as described in Clyde et al. (2011) to sample models.

Quantity	Truth	Horvitz Thompson			Monte Carlo			Re-normalized		
Δ	π_j	Bias	RMSE	\hat{s}	Bias	RMSE	\hat{s}	Bias	RMSE	$\widehat{\text{RMSE}}$
γ_6	0.13	0.12	1.41	1.42	0.13	1.57	1.56	-5.04	5.06	3.43
γ_{17}	0.13	-0.13	1.72	1.71	-0.24	1.79	1.78	-5.31	5.33	3.50
γ_4	0.14	-0.18	1.48	1.45	-0.09	1.71	1.71	-5.42	5.44	3.49
γ_{16}	0.14	-0.00	1.54	1.54	-0.02	1.71	1.71	-5.20	5.23	3.49
γ_{14}	0.14	-0.19	1.59	1.62	-0.08	1.77	1.77	-5.45	5.47	3.53
γ_8	0.14	-0.04	1.44	1.46	-0.11	1.56	1.56	-5.54	5.56	3.65
γ_9	0.15	-0.16	1.61	1.65	-0.08	1.76	1.76	-5.36	5.40	3.50
γ_{10}	0.15	-0.15	1.58	1.57	0.01	1.83	1.83	-5.46	5.48	3.55
γ_{12}	0.16	-0.13	1.66	1.68	-0.11	1.90	1.89	-5.15	5.18	3.40
γ_2	0.16	0.58	1.94	1.90	0.68	2.20	2.09	-5.17	5.20	3.88
γ_5	0.19	-0.21	1.73	1.73	-0.15	2.07	2.06	-4.69	4.73	3.09
γ_3	0.27	0.15	1.97	2.03	0.26	2.38	2.36	-3.33	3.40	2.60
γ_{15}	0.27	-0.18	2.00	2.03	-0.02	2.45	2.45	-5.86	5.90	3.88
γ_{20}	0.38	-0.57	3.23	3.21	-0.81	4.51	4.44	-4.90	5.07	3.53
γ_{13}	0.45	-0.14	2.47	2.53	0.06	3.39	3.39	-2.33	2.54	2.21
γ_{19}	0.72	0.77	3.14	3.09	0.84	4.41	4.33	2.00	2.35	2.25
γ_{11}	0.81	0.57	2.36	2.33	0.24	2.79	2.78	4.74	4.81	3.09
γ_7	1.00	-0.03	0.59	0.62	-0.08	0.69	0.68	0.37	0.37	0.46
γ_{18}	1.00	0.01	0.22	0.22	0.01	0.23	0.22	0.15	0.15	0.17
γ_1	1.00	0.00	0.03	0.03	0.00	0.02	0.02	0.01	0.01	0.02
$I(\gamma)$	-	0.05	0.09	0.11	0.02	0.30	0.30	0.30	0.32	0.20

Table 1. Average bias, average square root of mean squared error (RMSE), and estimated standard error (\hat{s}) or RMSE for the simulated data. Values reported in the table for bias, RMSE, and s are multiplied by 10^2 for $\Delta = \gamma_j$ and by 10^4 for $\Delta = I(\gamma)$.

We run the MH algorithm for 10,000 iterations, discarding the first 1,000 samples as burn-in for MC and RN. For HT we thin the chain by retaining every 10th sample to reduce dependence of draws. For computing \hat{C} for the HT estimator, we run a second chain of 1,000 iterations to determine the set A . Bias and square root of the mean squared error (RMSE) for estimating posterior marginal inclusion probabilities and posterior model probabilities are summarized in Table 1 and are based on running the MH algorithm 100 times with different random starting points generated uniformly. To estimate the bias in estimating Δ , for a scalar quantity, e.g. the variable inclusion indicators γ_j , we use the average bias over 100 replicates. For a Q dimensional vector, e.g. the 2^{20} dimensional vec-

289 tor of model indicators, we report the aggregate bias, given by $(\sum_{q=1}^Q (\text{bias}(\hat{\Delta}_q))^2 / Q)^{1/2}$.
 290
 291 The mean squared error for a scalar quantity is $\text{MSE}(\hat{\Delta}) = \sum_{i=1}^{100} (\hat{\Delta}^{(i)} - \Delta)^2 / 100$ while
 292 for vectors we report the average mean squared error over the components.

293 The ratio HT estimator appears to be comparable to the MC estimator in terms of
 294 bias (Table 1), where the bias in either case is negligible (roughly 1% for inclusion proba-
 295 bilities). In MCMC models are sampled according to their posterior probabilities, so that
 296 predictors with posterior inclusion probability greater than 0.5 will be oversampled and
 297 similarly those with posterior inclusion probability less than 0.5 will be undersampled. As
 298 the RN estimator does not take into account the sampling procedure; it systematically
 299 overestimates the larger inclusion probabilities and underestimates the smaller inclusion
 300 probabilities. As the MCMC sampler visits the same top models over most of the 100
 301 replicates, the RN estimates of inclusion probabilities exhibit low variability and the
 302 RMSE in Table 1 is dominated by the bias term.

303 We also compute the bounds on the absolute relative bias provided by Theorem 1 for
 304 inclusion probabilities. For ratio estimators with unbiased estimates of the numerator
 305 and denominator, $\rho = \Delta$ and the ARB from Theorem 1 for any Δ is bounded by the
 306 coefficient of variation CV for the normalizing constant; for RN the CV = 0.014, while
 307 for HT the CV = 0.082 (based on a 10% thinned MCMC). For RN and the approximate
 308 ratio HT estimator, there is an extra term $|\rho - \Delta|/s_r$ in the bound; the bounds range
 309 from 0.10 to 0.44 for HT and from 1.60 to 11.26 for RN for inclusion probabilities not
 310 equal to 1.0. This term is of the same order of magnitude of the CV for HT, but clearly
 311 dominates the ARB for RN. Running the MCMC ten times longer, the ARB for RN
 312 remains the same or doubles (50% of the cases) suggesting that the bias is decreasing
 313 at a slower rate than the standard deviation. For HT, the ARB generally decreases with
 314 longer runs. In both scenarios the bounds on ARB are fairly tight.

315 While we would like bias to be small, MSE is more important in practice. We find that
 316 the estimates of approximate variance (MSE) from Lemma 1 are in close agreement with
 317 the MSE for HT. HT typically has the smallest RMSEs for inclusion probabilities, where
 318 it is approximately 50% more efficient than MC for inclusion probabilities near 0.5. RN
 319 is slightly better in terms of MSE for inclusion probabilities near 1, where the numerator
 320 and denominator are highly correlated reducing the bias as suggested by Theorem 1. For
 321 estimating model probabilities HT clearly has a smaller MSE than either MC or RN,
 322 which will translate into more efficient estimates for BMA.

323 324 325 6. DISCUSSION

326 Renormalized estimators provide exact posterior quantities under enumeration of
 327 model spaces and are consistent, but will lead to biased estimators in finite samples
 328 as the estimator does not account for unequal sampling probabilities. In larger model
 329 spaces where a significantly smaller fraction of the model space may be sampled, both bias
 330 and variability in RN estimates may be much larger than MC estimators (as seen in GD-
 331 MB). Our results suggest that the ratio of Horvitz-Thompson estimators may improve
 332 upon both MC and RN estimators. While both HT and RN use marginal likelihoods of
 333 unique sampled models (leading to a reduction in variance), the ratio Horvitz-Thompson
 334 estimator takes into account the unequal sampling probabilities of models in order to
 335 construct unbiased estimates. This is closely related to reweighting in importance sam-
 336 pling. Hesterberg (1995) discusses alternative methods for constructing weights in im-

portance sampling and found that regression estimators could improve upon the simple ratio estimate usually employed in importance sampling. Adapting regression estimators (calibration estimators (Theberge, 1999)) or model based methods from the sample survey literature may provide additional improvements for BMA.

7. ACKNOWLEDGMENTS

The authors would like to thank the editor, associate editor and reviewer for their helpful comments and James Scott and Matthew Heaton for interesting discussions on this topic. The first author was supported by National Institutes of Health grant NL 1R01-HL090559 and the second author was supported by the National Institutes of Health grants NIH/NIEHS 5T32ES007018 and NIH/NIEHS P30 ES10126 and the Old Gold Fellowship, University of Iowa.

REFERENCES

- BERGER, J. O. & MOLINA, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica* **59**, 3–15.
- CLYDE, M., DESIMONE, H. & PARMIGIANI, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* **91**, 1197–1208.
- CLYDE, M. & GEORGE, E. I. (2004). Model uncertainty. *Statistical Science* **19**, 81–94.
- CLYDE, M. A., GHOSH, J. & LITTMAN, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics* **20**(1), 80–101.
- DAVID, I. P. & SUKHATME, B. V. (1974). On the bias and mean square error of the ratio estimator. *Journal of the American Statistical Association* **69**, pp. 464–466.
- GEORGE, E. (1999a). Discussion of “Model averaging and model search strategies” by M. Clyde. In *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*.
- GEORGE, E. I. (1999b). Comment on “Bayesian model averaging: A tutorial” (Pkg: p382-417). *Statistical Science* **14**, 409–412.
- GEORGE, E. I. & MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–374.
- HANSEN, M. H. & HURWITZ, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics* **14**, 333–362.
- HARTLEY, H. O. & ROSS, A. (1954). Unbiased ratio estimators. *Nature* **174**, 270–271.
- HEATON, M. & SCOTT, J. (2010). Bayesian computation and the linear model. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, M.-H. Chen, D. K. Dey, P. Mueller, D. Sun & K. Ye, eds.
- HESTERBERG, T. C. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37**, 185–194.
- HOETING, J. A., MADIGAN, D., RAFTERY, A. E. & VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science* **14**, 382–401. Corrected version at <http://www.stat.washington.edu/www/research/online/hoeting1999.pdf>.
- HORVITZ, D. & THOMPSON, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.
- NOTT, D. J. & GREEN, P. J. (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics* **13**, 141–157.
- NOTT, D. J. & KOHN, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92**, 747–763.
- RAFTERY, A. E., MADIGAN, D. & HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179–191.
- SCOTT, J. G. & CARVALHO, C. M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics* **17**, 790–808.
- THEBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association* **94**, pp. 635–644.
- THOMPSON, S. K. (1992). *Sampling*. Wiley Interscience.
- ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g -prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. North-Holland/Elsevier.