Biometrika (2012), **99**, 4, *pp.* 981–988
doi: 10.1093/biomet/ass040
Advance Access publication 30 September 2012

Finite population estimators in stochastic search variable selection

BY MERLISE A. CLYDE

Department of Statistical Science, Duke University, Durham, North Carolina 27708-0251, U.S.A.
clyde@stat.duke.edu

AND JOYEE GHOSH

Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, Iowa 52242-1409, U.S.A.
joyee-ghosh@uiowa.edu

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, renormalized 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 that 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 that 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 renormalized estimators, with little increase in computational cost.

Some key words: Bayesian model averaging; Horvitz–Thompson estimator; Inclusion probability; Markov chain Monte Carlo; Median probability model; Model uncertainty; Variable selection.

1. INTRODUCTION

The advent of Markov chain Monte Carlo algorithms greatly expanded Bayesian model selection and model averaging in regression problems that preclude enumeration (Hoeting et al., 1999; Clyde & George, 2004). 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 $\gamma^{(0)}, \gamma^{(1)}, \ldots$ on Γ such that the stationary distribution, π, is the posterior distribution

$$\pi \equiv p(M_\gamma \mid Y) = \frac{p(Y \mid M_\gamma)p(M_\gamma)}{\sum_{\gamma\in\Gamma} p(Y \mid M_\gamma)p(M_\gamma)}, \tag{1}$$

where $p(M_\gamma)$ is the prior probability of model M_γ and the marginal likelihood of model M_γ is proportional to $p(Y \mid M_\gamma) = \int p(Y \mid \theta_\gamma, M_\gamma)p(\theta_\gamma \mid M_\gamma)\,\mathrm{d}\theta_\gamma$, obtained by integrating the sampling distribution of data Y with respect to the prior distribution of model-specific parameters θ_γ. The Monte Carlo or empirical frequencies f_γ of models provide simulation-consistent estimates of posterior model probabilities, $\hat{p}^{\mathrm{MC}}(M_\gamma \mid Y) = T^{-1}\sum_{t=1}^{\mathrm{T}} I\{M_\gamma = M^{(t)}\} = f_\gamma/T$ as the number of iterations of the Markov chain $T \to \infty$. When marginal likelihoods are available, renormalized estimates of posterior probabilities for models may be obtained by replacing Γ in (1) with S_T, the set of unique sampled models. As with empirical

estimators, models not in S_T have estimated probability zero. The renormalized estimators provide exact Bayes factors for comparing any two models and have been used in various contexts by Clyde et al. (1996), George & McCulloch (1997), and Raftery et al. (1997) and more recently by Scott & Carvalho (2008) and Clyde et al. (2011). George (1999a, b) suggests that using the renormalized model probabilities may lead to substantial improvements over the empirical estimates.

Recent simulation studies comparing renormalized and empirical estimates, however, have led to mixed results. Some of the latest stochastic search algorithms, which exclusively use renormalized 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 Markov chain Monte Carlo algorithms, they paradoxically had poorer performance for estimation of inclusion probabilities. Clyde et al. (2011) found that renormalized estimators from the Bayesian adaptive sampling algorithm, which samples without replacement, could have much smaller mean squared errors for estimating inclusion probabilities than either empirical or renormalized estimators from Markov chain Monte Carlo. In contrast, Garcia-Donato and Martínez-Beneito in a technical report available from http://arxiv.org/pdf/1101.4368v1 demonstrated that empirical estimates of inclusion probabilities were preferable to the renormalized estimates from either Bayesian adaptive sampling or Berger & Molina (2005). These results motivate the current work to understand theoretical properties of renormalized estimators in conjunction with the sampling method.

2. ESTIMATION IN BAYESIAN MODEL AVERAGING

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

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

for various choices of $\Delta(M_\gamma)$: $\Delta(M_\gamma) = I(M_j = M_\gamma)$ for posterior model probabilities, $\Delta(M_\gamma) = \gamma_j$ for inclusion probabilities, and $\Delta(M_\gamma) = \hat{Y}_{M_\gamma}$ for prediction. The Hansen–Hurwitz (1943) estimator

$$\hat{\Delta}^{\mathrm{HH}} = \frac{1}{T} \sum_{t=1}^{T} \frac{\tilde{\Delta}_t}{\pi_t^S} = \sum_{\gamma \in \Gamma} \frac{\tilde{\Delta}_\gamma}{\pi_\gamma^S} \frac{1}{T} \sum_{t=1}^{T} I(M_t = M_\gamma) = \sum_{\gamma \in \Gamma} \frac{\tilde{\Delta}_\gamma}{\pi_\gamma^S} \frac{f_\gamma}{T} = \sum_{\gamma \in \Gamma} \frac{\tilde{\Delta}_\gamma}{\pi_\gamma^S} \hat{p}^{\mathrm{MC}}(M_\gamma \mid Y), \tag{3}$$

where $\tilde{\Delta}_t = \Delta(M_t) p(M_t \mid Y)$ and π_t^S is the probability of sampling model M_t, may be used to unify several estimators. If models are sampled independently and with replacement from the posterior distribution, $\pi_t^S = p(M_t \mid Y)$, then the estimator is unbiased and reduces to the Monte Carlo estimator, $\hat{\Delta}^{\mathrm{MC}} = \sum_{\gamma \in \Gamma} \Delta(M_\gamma) \hat{p}^{\mathrm{MC}}(M_\gamma \mid Y)$, as shown by Garcia-Donato and Martínez-Beneito. Under Markov chain Monte Carlo sampling with finite T this is not the case.

PROPOSITION 1. *Given an aperiodic and positive recurrent Markov chain $M^{(0)}, M^{(1)}, \ldots$, an initial distribution α, and matrix of transition probabilities $P(t) \equiv p_{ij}(t) = \mathrm{pr}(M^{(t)} = M_j \mid M^{(0)} = M_i)$ such that the stationary distribution is $\pi = p(\cdot \mid Y)$, let $\hat{p}^{\mathrm{MC}}$ denote the vector of Monte Carlo frequencies (f_γ / T). Then for finite T*

$$E(\hat{p}^{\mathrm{MC}})' = T^{-1} \alpha' \sum_{t=1}^{T} P(t).$$

Proof. Let $M^{(t)}$ denote the state of the model at time t. Then

$$E\{\hat{p}^{\mathrm{MC}}(M_j \mid Y)\} = E\left\{\sum_{t=1}^{\mathrm{T}} \frac{I(M^{(t)} = M_j)}{T}\right\} = T^{-1}\sum_{t=1}^{\mathrm{T}} \mathrm{pr}(M^{(t)} = M_j)$$

$$= T^{-1}\sum_{t=1}^{\mathrm{T}}\sum_{i} \alpha_i \mathrm{pr}(M^{(t)} = M_j \mid M^{(0)} = M_i) = T^{-1}\sum_{t=1}^{\mathrm{T}}\sum_{i} \alpha_i p_{ij}(t).$$

□

The expectation depends on the initial distribution and transition matrix so that the estimator $\hat{p}^{\mathrm{MC}}(M_j \mid Y)$ is biased for finite T. In the Hansen–Hurwitz estimator of (3), it is no longer the case that $\pi_t^S = p(M_t \mid Y)$ for finite T. The ergodic theorem implies that $\hat{p}^{\mathrm{MC}}(M_j \mid Y)$ is asymptotically unbiased as $T \to \infty$ (Roberts, 1996). Because Γ is a finite state space, the chain is uniformly ergodic and hence under existence of a second moment $T^{1/2}(\hat{\Delta}^{\mathrm{MC}} - \Delta)$ converges weakly to a normal distribution with mean zero and variance σ_Δ^2. Bias may be an issue in practice when the chain may have a low probability of transitioning from one high-probability state to another, as in multimodal problems. Modifying the transition kernel (Nott & Green, 2004) and increasing the number of iterations can reduce bias of empirical estimators in finite samples.

The minimal sufficient statistic in finite population sampling is the unordered set of distinct labelled observations (Thompson, 1992, Ch. 3); the Hansen–Hurwitz estimator is not a function of the minimal sufficient statistic, but 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 and rarely used in practice. An alternative estimator that is a function of the minimal sufficient statistic is the Horvitz–Thompson (1952) estimator. For the variable selection problem the minimal sufficient statistics are the model indices M_γ and values $\Delta(M_\gamma)$. By using Horvitz–Thompson 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 ratio Horvitz–Thompson estimators in the context of Markov chain Monte Carlo sampling.

3. Ratio estimators

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

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

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.

Proposition 2. *Let the posterior expectation of function $\Delta(M_\gamma)$ be*

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

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

$$E(R) - \Delta = \rho - \Delta - \mathrm{cov}(R, \bar{b})/\mu_b$$

and the absolute relative bias is

$$\frac{|E(R) - \Delta)|}{\sigma_R} \leqslant \frac{|\Delta - \rho|}{\sigma_R} + \frac{\sigma_{\bar{b}}}{|\mu_b|}. \tag{5}$$

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

$$|E(R) - \Delta| \leqslant |\rho - \Delta| + \frac{|\mathrm{cov}(R, \bar{b})|}{|\mu_b|} \leqslant |\rho - \Delta| + \frac{\sigma_R \sigma_{\bar{b}}}{|\mu_b|}$$

and dividing by σ_R completes the proof, where σ_R and $\sigma_{\bar{b}}$ are the standard deviations of R and $\bar{b}$, respectively. □

The extra term involving $|\rho - \Delta|$ in (5) is zero if $\mu_a = \Delta_{\mathrm{num}}$ and $\mu_b = \Delta_{\mathrm{den}}$, i.e., unbiased estimates of the numerator and denominator are available, in which case a practical estimate of the bound on the relative bias is provided by an estimate of the coefficient of variation for $\bar{b}$. When $\rho \doteqdot \Delta$, the magnitude of the bias may exceed the coefficient of variation. Our goal is to construct a ratio estimator where $\rho = \Delta$, which leads to our main result.

PROPOSITION 3. *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 $a_\gamma^{\mathrm{HT}} = \Delta(M_\gamma)p(Y \mid M_\gamma)p(M_\gamma)/\pi^S(M_\gamma)$ and $b_\gamma^{\mathrm{HT}} = p(Y \mid M_\gamma)p(M_\gamma)/\pi^S(M_\gamma)$. Then the Horvitz–Thompson estimators $\bar{a}^{\mathrm{HT}}$ and $\bar{b}^{\mathrm{HT}}$ are unbiased estimators of Δ_{num} and Δ_{den}, respectively, and the ratio $R^{\mathrm{HT}} = \bar{a}^{\mathrm{HT}}/\bar{b}^{\mathrm{HT}}$ of Horvitz–Thompson estimators defined in* (4) *is approximately unbiased for estimating Δ, with approximate variance*

$$E(R^{\mathrm{HT}} - \Delta)^2 \approx \frac{\mathrm{var}(\bar{a}^{\mathrm{HT}} - \Delta\bar{b}^{\mathrm{HT}})}{\Delta_{\mathrm{den}}^2}. \tag{6}$$

Proof. Since

$$\bar{a}^{\mathrm{HT}} \equiv \sum_{\gamma \in \Gamma} a_\gamma^{\mathrm{HT}} I(M_\gamma \in S_t) = \sum_{\gamma \in \Gamma} \Delta(M_\gamma) \frac{p(Y \mid M_\gamma)p(M_\gamma)}{\pi^S(M_\gamma)} I(M_\gamma \in S_t)$$

$\mu_a = E(\bar{a}^{\mathrm{HT}}) = \sum_{\gamma \in \Gamma} \Delta(M_\gamma)p(Y \mid M_\gamma)p(M_\gamma) = \Delta_{\mathrm{num}}$ as $E\{I(M_\gamma \in S_t)\} = \pi^S(M_\gamma)$. The unbiasedness of $\bar{b}^{\mathrm{HT}}$ follows similarly, and hence $\rho = \Delta$.

Expanding $f(a, b) = a/b$ about the point $(\Delta_{\mathrm{num}}, \Delta_{\mathrm{den}})$ and keeping the first two terms of the Taylor's series expansion leads to the linear approximation

$$R^{\mathrm{HT}} - \Delta \approx \frac{\bar{a}^{\mathrm{HT}} - \Delta_{\mathrm{num}} - \Delta(\bar{b}^{\mathrm{HT}} - \Delta_{\mathrm{den}})}{\Delta_{\mathrm{den}}} = \frac{\bar{a}^{\mathrm{HT}} - \Delta\,\bar{b}^{\mathrm{HT}}}{\Delta_{\mathrm{den}}}. \tag{7}$$

Taking expectations, R is then approximately unbiased for estimating Δ. Squaring (7) and taking expectations leads to the approximate variance of the ratio estimator. □

David & Sukhatme (1974) provide rigorous justification for the approximations used to estimate the mean squared error of ratio estimators.

It is clear from Proposition 3 that the renormalized estimator is a ratio of Horvitz–Thompson estimators and approximately unbiased only in the case of simple random sampling with or without replacement where each model has an equal probability of being selected in the sample, and that in all other cases $\rho \doteqdot \Delta$. We propose a novel method for computing ratio Horvitz–Thompson estimators, in order to provide approximately unbiased estimators that incorporate marginal likelihoods from sampled models using the minimal sufficient statistics.

4. RATIO HORVITZ–THOMPSON ESTIMATORS FOR MARKOV CHAIN MONTE CARLO

Under Markov chain Monte Carlo sampling, direct calculation of $\pi^S(M_\gamma)$, the probability that model M_γ is sampled in T iterations, is possible theoretically, but it involves calculation of the $2^p \times 2^p$ single-step transition matrix, where p is the number of covariates. The order of the computation will be magnitudes higher than enumerating the model space, making exact Horvitz–Thompson estimation impractical.

We instead propose an approximation to $\pi^S(M_\gamma)$, based on thinning the chain so that the remaining T^* samples are approximately independent, and estimating the model inclusion probability as

$$\pi^S(M_\gamma) = 1 - \{1 - p(M_\gamma \mid Y)\}^{T^*} = 1 - \{1 - Cp(Y \mid M_\gamma)p(M_\gamma)\}^{T^*}$$

where C is the normalizing constant. Thinning results in little loss of information, as the Horvitz–Thompson estimate uses the unique marginal likelihoods, and no new information is provided by repeat sampling of a model. Because C is unknown we use an estimate of the normalizing constant

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

where A is the set of unique models based on running a second independent Markov chain (George & McCulloch, 1997). Using $\hat{C}$ in place of the C in the model probabilities provides a simulation consistent estimate $\widehat{\pi}^S(M_\gamma)$ of the model inclusion probabilities, which are then used to construct ratio Horvitz–Thompson estimators for quantities of interest. An estimate of the approximate variance from (6) may be obtained by using the standard Horvitz–Thompson expressions for the variance (Thompson, 1992, p. 69) using the variable $z_\gamma = (a_\gamma - \hat{\Delta} b_\gamma)/\hat{\Delta}_{\text{den}}$. However, our simulation results suggest that better estimates of the variance may be obtained using multi-core or distributed environments to run multiple chains.

While the Horvitz–Thompson estimators for estimating Δ_{num} and Δ_{den} are unbiased and functions of the minimal sufficient statistics, minimal sufficient statistics for finite population sampling are not complete and there is no unique minimum variance estimator. Simulation studies provide evidence that the Horvitz–Thompson estimator provides reductions in mean squared error over the empirical and renormalized estimators.

5. Simulations

We use a simulation design similar to the study in Nott & Kohn (2005), but with dimension $p = 20$. 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 correlation 0·99 with column 19. The response is generated as $Y \sim N(\alpha 1_{50} + X\beta, 2{\cdot}5^2 I)$ where $\alpha = 4$, $\beta = (2, 0, 0, 0, -1, 0, 1{\cdot}5, 0, 0, 0, 1, 0, 0{\cdot}5, 0, 0, 0, 0, -1, 1, 4)'$, and 1_{50} is a vector of ones with length 50. For illustration, we use Zellner's (1986) g-prior with $g = n$ 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 algorithm with add/delete steps and random swap proposals as described in Clyde et al. (2011) to sample models.

For all estimators, we run a Markov chain for 10 000 iterations, and discard the first 1000 samples as burn-in. For Horvitz–Thompson estimation, we take the first 8000 iterations after burn-in and thin the chain by retaining every 8th sample to reduce dependence of draws. To compute $\hat{C}$ for the Horvitz–Thompson estimator, we run a second independent chain of 1000 iterations to determine the set A, so that all estimators are based on an equivalent computational effort of 10 000 iterations. Bias, the standard error, and square root of the mean squared error for estimating posterior marginal inclusion probabilities and posterior model probabilities are summarized in Table 1 and are based on running the Metropolis–Hastings algorithm 100 times with different random starting points generated uniformly. To estimate the bias in estimating Δ for scalar quantities, 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 vector of model indicators, we report the aggregate bias, $\{\sum_{q=1}^{Q} \text{bias}(\hat{\Delta}_q)^2/Q\}^{1/2}$. The mean squared error for a scalar quantity is $\text{MSE}(\hat{\Delta}) = \sum_{i=1}^{100}(\hat{\Delta}^{(i)} - \Delta)^2/100$, while for vectors we report the average mean squared error over the components.

Table 1. *Bias, standard error and square root of average mean squared error, for the estimators of* Δ *given in the first column based on* 100 *simulations*

	Truth	Horvitz–Thompson			Monte Carlo			Renormalized		
Δ	π_j	Bias	s	RMSE	Bias	s	RMSE	Bias	s	RMSE
γ_6	0·13	0·11	1·48	1·48	0·13	1·56	1·57	−5·04	0·46	5·06
γ_{17}	0·13	−0·04	1·60	1·60	−0·24	1·78	1·79	−5·31	0·48	5·33
γ_4	0·14	−0·22	1·67	1·68	−0·09	1·71	1·71	−5·42	0·50	5·44
γ_{16}	0·14	−0·00	1·77	1·77	−0·02	1·71	1·71	−5·20	0·54	5·23
γ_{14}	0·14	0·03	1·73	1·73	−0·08	1·77	1·77	−5·45	0·48	5·47
γ_8	0·14	−0·19	1·54	1·55	−0·11	1·56	1·56	−5·54	0·49	5·56
γ_9	0·15	−0·05	1·69	1·69	−0·08	1·76	1·76	−5·36	0·58	5·40
γ_{10}	0·15	−0·08	1·59	1·59	0·01	1·83	1·83	−5·46	0·49	5·48
γ_{12}	0·16	−0·18	1·81	1·82	−0·11	1·89	1·90	−5·15	0·53	5·18
γ_2	0·16	0·59	1·83	1·93	0·68	2·09	2·20	−5·17	0·56	5·20
γ_5	0·19	−0·31	1·81	1·83	−0·15	2·06	2·07	−4·69	0·68	4·73
γ_3	0·27	0·15	1·77	1·78	0·26	2·36	2·38	−3·33	0·70	3·40
γ_{15}	0·27	−0·18	2·12	2·13	−0·02	2·45	2·45	−5·86	0·73	5·90
γ_{20}	0·38	−0·31	3·26	3·28	−0·81	4·44	4·51	−4·90	1·28	5·07
γ_{13}	0·45	−0·11	2·60	2·60	0·06	3·39	3·39	−2·33	1·01	2·54
γ_{19}	0·72	0·63	3·18	3·24	0·84	4·33	4·41	2·00	1·25	2·35
γ_{11}	0·81	0·67	2·19	2·29	0·24	2·78	2·79	4·74	0·83	4·81
γ_7	1·00	−0·03	0·61	0·61	−0·08	0·68	0·69	0·37	0·03	0·37
γ_{18}	1·00	0·02	0·25	0·25	0·01	0·22	0·23	0·15	0·00	0·15
γ_1	1·00	−0·00	0·03	0·03	0·00	0·02	0·02	0·01	0·00	0·01
$I(\gamma)$	–	0·06	0·10	0·12	0·03	0·30	0·30	0·31	0·32	0·45

s, standard error; RMSE, square root of average mean squared error. Values reported in the table are multiplied by 100 for $\Delta = \gamma_j$ and by 10^4 for $\Delta = I(\gamma)$.

The Horvitz–Thompson estimators are comparable to the empirical estimators in terms of bias, which is negligible in either case. The bias of the renormalized estimators is orders of magnitude higher than for the other methods, with errors for inclusion probabilities often around 5%. There is a suggestion of systematic bias, with inclusion probabilities larger than 0·5 overestimated and inclusion probabilities smaller than 0·5 underestimated. As the sampler visits the same top models over most of the 100 replicates, the renormalized estimates exhibit low variability and their mean squared error in Table 1 is dominated by their bias.

We also compute the bounds on the absolute relative bias in Proposition 2 for inclusion probabilities. For ratio estimators with unbiased estimates of the numerator and denominator, $\rho = \Delta$ and the absolute relative bias for any Δ is bounded by the coefficient of variation for the normalizing constant; the coefficients of variation for the renormalized and Horvitz–Thompson estimators are 0·014 and 0·092, respectively. For the renormalized estimator, there is an extra term $|\rho - \Delta|/\sigma_R$ in the bound that ranged from 1·60 to 11·24 for inclusion probabilities not equal to 1·0. This extra term clearly dominates the absolute relative bias for the renormalized estimator. Running the Markov chain ten times longer, the absolute relative bias for the renormalized estimator remains the same or doubles in 50% of the cases, suggesting that the bias decreases at a slower rate than does the standard deviation. For Horvitz–Thompson estimators, the absolute relative bias generally decreases with longer runs. In both scenarios the bounds on the bias are fairly tight; however, we have a practical sample estimate of the bound only in the case of the Horvitz–Thompson estimator.

Mean squared error may be more important than bias in practice. We find that Horvitz–Thompson has the smallest mean squared error for about 80% of the inclusion probabilities not equal to 1·0, where it is substantially more efficient than the Monte Carlo estimator for inclusion probabilities near 0·5. For estimating model probabilities Horvitz–Thompson clearly has a smaller mean squared error than the other two estimators, which will translate into more efficient estimates for model averaging.

When true population quantities are unknown, the mean squared error may be approximated using estimates from multiple chains with the estimate of the true Δ replaced by its corresponding multi-chain ergodic average. This is computationally practical with today's multi-core or distributed computing environments, and leads to estimates of the mean squared error that are within rounding error of the values in Table 1. This provides the user with more information for choosing among the different estimators.

6. Discussion

Renormalized estimators provide exact posterior quantities under enumeration of model spaces and are consistent, but will lead to biased estimators in finite samples, as the estimator does not account for unequal sampling probabilities. In larger model spaces where a significantly smaller fraction of the model space may be sampled, both bias and variability in renormalized estimates may be much larger than empirical estimators, as seen in the technical report by Garcia-Donato and Martínez-Beneito. Our simulation results suggest that the ratio of Horvitz–Thompson estimators may improve upon both Monte Carlo and renormalized estimators. While both Horvitz–Thompson and renormalized estimators use marginal likelihoods of unique sampled models leading to a reduction in variance, the ratio Horvitz–Thompson estimator takes into account the unequal sampling probabilities of models in order to construct unbiased estimates. This is closely related to reweighting in importance sampling. Hesterberg (1995) discusses alternative methods for constructing weights in importance sampling and found that regression estimators could improve upon the simple ratio estimate usually employed in importance sampling. Adapting regression or calibration estimators (Theberge, 1999) or model-based methods from the sample survey literature to this Markov chain Monte Carlo sampling context may provide additional improvements.

Acknowledgement

The authors thank the editor, associate editor and reviewer for their helpful comments and James Scott and Matthew Heaton for interesting discussions. The first author was supported by the National Science Foundation and the second author was supported by the Old Gold Fellowship, University of Iowa.

References

Berger, J. O. & Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statist. Neer.* **59**, 3–15.

Clyde, M. A., DeSimone, H. & Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *J. Am. Statist. Assoc.* **91**, 1197–208.

Clyde, M. A. & George, E. I. (2004). Model uncertainty. *Statist. Sci.* **19**, 81–94.

Clyde, M. A., Ghosh, J. & Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *J. Comp. Graph. Statist.* **20**, 80–101.

David, I. P. & Sukhatme, B. V. (1974). On the bias and mean square error of the ratio estimator. *J. Am. Statist. Assoc.* **69**, 464–6.

George, E. I. (1999a). Comment on "Bayesian model averaging: A tutorial". *Statist. Sci.* **14**, 409–12.

George, E. I. (1999b). Discussion of "Model averaging and model search strategies" by M. Clyde. In *Bayesian Statist.* 6 – *Proc. 6th Valencia Int. Meeting*.

George, E. I. & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7**, 339–74.

Hansen, M. H. & Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.* **14**, 333–62.

Hartley, H. O. & Ross, A. (1954). Unbiased ratio estimators. *Nature* **174**, 270–1.

Heaton, M. & Scott, J. G. (2010). Bayesian computation and the linear model. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, Ed. M.-H. Chen, D. K. Dey, P. Mueller, D. Sun & K. Ye, pp. 527–45. New York: Springer.

Hesterberg, T. C. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37**, 185–94.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statist. Sci.* **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. *J. Am. Statist. Assoc.* **47**, 663–85.

NOTT, D. J. & GREEN, P. J. (2004). Bayesian variable selection and the Swendsen–Wang algorithm. *J. Comp. Graph. Statist.* **13**, 141–57.

NOTT, D. J. & KOHN, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92**, 747–63.

RAFTERY, A. E., MADIGAN, D. & HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *J. Am. Statist. Assoc.* **92**, 179–91.

ROBERTS, G. O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice*, Ed. W. Gillks, S. Richardson & D. J. Spiegelhalter, pp. 45–57. Boca Raton: Chapman and Hall/CRC.

SCOTT, J. G. & CARVALHO, C. M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *J. Comp. Graph. Statist.* **17**, 790–808.

THEBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *J. Am. Statist. Assoc.* **94**, 635–44.

THOMPSON, S. K. (1992). *Sampling*. Hoboken: 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*, Ed. P. Goel & A. Zellner, pp. 233–43. Amsterdam: North-Holland/Elsevier.

[*Received May* 2010. *Revised April* 2012]