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 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 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 $M_\gamma$ 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 $\Gamma$ such that the stationary distribution, $\pi$, is the posterior distribution

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

(1)

where $p(M_\gamma)$ is the prior probability of model $M_\gamma$ and the marginal likelihood of model $M_\gamma$ is proportional to $p(Y \mid M_\gamma) = \int p(Y \mid \theta_\gamma, M_\gamma)p(\theta_\gamma \mid M_\gamma)d\theta_\gamma$, obtained by integrating the sampling distribution of data $Y$ with respect to the prior distribution of model specific parameters $\theta_\gamma$. The Monte Carlo or empirical frequencies $f_\gamma$ of models provide simulation consistent esti-
mates of posterior model probabilities, \( p^{MC}(M_{\gamma} \mid Y) = T^{-1} \sum_{t=1}^{T} I\{M_{\gamma} = M^{(t)}\} = f_{\gamma}/T \) 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 \( \Gamma \) 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 (1999b,a) 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, had much smaller mean squared errors for estimating inclusion probabilities, than either empirical or renormalized estimators from Markov chain Monte Carlo. In a high dimensional simulation study reported in a 2011 technical report available at http://arxiv.org/pdf/1101.4368v1, Garcia–Donato and Martínez–Beneito demonstrated that empirical estimates of inclusion probabilities were preferable to the renormalized estimates from either Bayesian Adaptive Sampling algorithm of Clyde et al. or Berger & Molina. 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

\[
\Delta = \sum_{\gamma \in \Gamma} \Delta(M_{\gamma}) p(M_{\gamma} \mid Y) = \frac{\sum_{\gamma \in \Gamma} \Delta(M_{\gamma}) p(Y \mid M_{\gamma}) p(M_{\gamma})}{\sum_{\gamma \in \Gamma} p(Y \mid M_{\gamma}) p(M_{\gamma})}
\]

(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 predictions. If models are generated independently from \( p(M_{\gamma} \mid Y) \), then the Monte Carlo estimator \( \hat{p}^{MC} \) is shown by Garcia-Donato and Martínez-Beneito to be equivalent to the Hansen–Hurwitz (1943) estimator:

\[
\hat{\Delta}_{HH} = \frac{1}{T} T \sum_{t=1}^{T} \frac{\Delta_{t}}{\pi_{t}^{\gamma}} = \frac{1}{T} \sum_{t=1}^{T} \frac{\Delta_{t}}{\pi_{t}^{\gamma}} 1 \sum_{t=1}^{T} I(M_{t} = M_{\gamma}) = \frac{\sum_{\gamma \in \Gamma} \Delta_{\gamma} f_{\gamma}}{T} = \frac{\sum_{\gamma \in \Gamma} \pi_{\gamma} \hat{p}^{MC}(M_{\gamma})}{T}
\]

where \( \Delta_{t} = \Delta(M_{t}) p(M_{t} \mid Y) \), \( \pi_{t}^{\gamma} \) is the probability of sampling model \( M_{t} \), \( f_{\gamma} \) is the frequency of model \( M_{\gamma} \) in the sample and \( \hat{p}^{MC}(M_{\gamma}) = f_{\gamma}/T \) is the Monte Carlo estimator of the probability of model \( M_{\gamma} \). If \( \pi_{t}^{\gamma} = p(M_{t} \mid Y) \), then the estimator is unbiased and reduces to the Monte Carlo estimator, \( \hat{\Delta}_{MC} = \sum_{\gamma \in \Gamma} \Delta(M_{\gamma}) p^{MC}(M_{\gamma}) \). Under Markov chain Monte Carlo sampling, a standard result is:
Finite Population Estimators in Stochastic Search Variable Selection

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

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

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

$$E(\hat{p}^{MC}(M_{j} | Y)) = E\left\{\sum_{t=1}^{T} \frac{I(M^{(t)} = M_{j})}{T}\right\} = T^{-1} \sum_{t=1}^{T} \text{pr}(M^{(t)} = M_{j})$$

$$= T^{-1} \sum_{t=1}^{T} \sum_{i} \alpha_{i} \text{pr}(M^{(t)} = M_{j} | M^{(0)} = M_{i}) = T^{-1} \sum_{t=1}^{T} \sum_{i} \alpha_{i} p_{ij}^{t}. \quad (3)$$

For finite $T$, the expectation depends on the initial distribution and transition kernel. Ergodicity implies that $\hat{p}^{MC}(M_{j} | Y)$ is asymptotically unbiased as $T \rightarrow \infty$. Furthermore, because $\Gamma$ is a finite state space, the chain is uniformly ergodic and hence under existence of a second moment $T^{1/2}(\Delta_{MC} - \Delta)$ converges weakly to a normal distribution with mean zero and variance $\sigma_{\Delta}^{2}$. For finite samples under an arbitrary initial distribution the estimator is biased, as $\hat{p}^{(T)}_{\gamma} \neq p(M_{t} | Y)$. 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 labeled observations (Thompson, 1992, Chapter 3); the Hansen–Horwitz 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 $M_{\gamma}$ 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

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

for various choices of $a_{\gamma}$ and $b_{\gamma}$. 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 for some function \( \Delta(M_\gamma) \) be
\[
\Delta = \frac{\sum_{\gamma \in \Gamma} \Delta(M_\gamma)p(Y \mid M_\gamma)p(M_\gamma)}{\sum_{\gamma \in \Gamma} p(Y \mid M_\gamma)p(M_\gamma)} \equiv \frac{\Delta_{\text{num}}}{\Delta_{\text{den}}},
\]
and let \( r \) denote a ratio estimator of \( \Delta \) 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
\[
E(r) - \Delta = \rho - \Delta - \text{cov}(r, \bar{b})/\mu_b
\]
and the absolute relative bias is
\[
\frac{|E(r) - \Delta|}{\sigma_r} \leq \frac{|\Delta - \rho|}{\sigma_r} + \frac{\sigma_{\bar{b}}}{|\mu_b|}.
\]

Proof. Starting with the covariance of \( r \) and \( \bar{b} \), \( \text{cov}(r, \bar{b}) = E(r\bar{b}) - E(r)E(\bar{b}) = \mu_a - \mu_bE(r) \) and we have upon rearranging \( E(r) - \Delta = \rho - \Delta - \text{cov}(r, \bar{b})/\mu_b \). Taking absolute values,
\[
|E(r) - \Delta| \leq |\rho - \Delta| + \frac{|\text{cov}(r, \bar{b})|}{|\mu_b|} \leq |\rho - \Delta| + \frac{\sigma_r \sigma_{\bar{b}}}{|\mu_b|}
\]
and dividing by \( \sigma_r \) completes the proof, where \( \sigma_r \) and \( \sigma_{\bar{b}} \) are the standard deviations of \( r \) and \( \bar{b} \).

The extra term involving \( |\rho - \Delta| \) in (5) is zero if \( \mu_a = \Delta_{\text{num}} \) and \( \mu_b = \Delta_{\text{den}} \), e.g. unbiased estimates of the numerator and denominator are available, in which case a practical estimate of the bound on the bias is provided by an estimate of the coefficient of variation for \( \bar{b} \). Note that when \( \rho \neq \Delta \), 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 \( M_\gamma \) is included in the set of unique models \( S_T \) from a sample of size \( T \). Set \( a_\gamma = \Delta(M_\gamma)p(Y \mid M_\gamma)p(M_\gamma)/\pi^S(M_\gamma) \) and \( b_\gamma = p(Y \mid M_\gamma)p(M_\gamma)/\pi^S(M_\gamma) \). Then the Horvitz–Thompson estimators \( \bar{a} \) and \( \bar{b} \) are unbiased estimators of \( \Delta_{\text{num}} \) and \( \Delta_{\text{den}} \), respectively, and the ratio of Horvitz–Thompson estimators \( r = \bar{a}/\bar{b} \) is approximately unbiased for estimating \( \Delta \), with approximate variance (mean-squared-error)
\[
E(r - \Delta)^2 \approx \frac{\text{var}(\bar{a} - \bar{b})}{\Delta_{\text{den}}^2}.
\]

Proof. Since \( \bar{a} = \sum_{\gamma \in \Gamma} a_\gamma I(M_\gamma \in S_T) = \sum_{\gamma \in \Gamma} \Delta(M_\gamma)p(Y \mid M_\gamma)p(M_\gamma)/\pi^S(M_\gamma)I(M_\gamma \in S_T) \), \( E(\bar{a}) = \mu_a = \sum_{\gamma \in \Gamma} \Delta(M_\gamma)p(Y \mid M_\gamma)p(M_\gamma) = \Delta_{\text{num}} \) as \( E(I(M_\gamma \in S_T) = \pi^S(M_\gamma) \). The unbiasedness of \( \bar{b} \) follows similarly, and hence \( \rho = \Delta \).

Expanding \( f(a, b) = a/b \) about the point \( (\Delta_{\text{num}}, \Delta_{\text{den}}) \) and keeping the first two terms of the Taylor’s series expansion leads to the linear approximation
\[
r - \Delta \approx \frac{\bar{a} - \Delta_{\text{num}} - \Delta(\bar{b} - \Delta_{\text{den}})}{\Delta_{\text{den}}} = \frac{\bar{a} - \Delta \bar{b}}{\Delta_{\text{den}}}.
\]
Taking expectations, \( r \) is then approximately unbiased for estimating \( \Delta \). Squaring (7) and taking expectations leads to the approximate mean squared error of the ratio estimator.

David & Sukhatme (1974) provide rigorous justification for the approximations used to estimate the mean squared error of ratio estimators.
As a consequence of Proposition 3, it is clear that the renormalized estimator is 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 \( \rho \neq \Delta \). As an alternative, 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 \( M_\gamma \) is sampled in \( T \) iterations, is possible, 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 estimators 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

\[
\pi^S(M_\gamma) = 1 - \{1 - p(M_\gamma \mid Y)\}^{T^*} = 1 - \{1 - C p(Y \mid M_\gamma) p(M_\gamma)\}^{T^*} \tag{8}
\]

where \( C \) is the normalizing constant. Thinning results in little loss of information as the Horvitz–Thompson estimate uses the unique labels, rather than the number of repeat visits. Because \( C \) is unknown we use an estimate of the normalizing constant

\[
\hat{C} = \frac{\sum_{t=1}^{T^*} I(M^{(t)} \in A)}{T^*} \frac{1}{\sum_{\gamma \in A} p(Y \mid M_\gamma) p(M_\gamma)} \tag{9}
\]

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 \( \hat{\pi}^S(M_\gamma) \) 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, page 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 \( \Delta_{\text{num}} \) and \( \Delta_{\text{den}} \) 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 \( 50 \times 20 \) 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 \( Y \sim N(\alpha 1_{50} + X \beta, 2.5^2 I) \) where \( \alpha = 4, \beta = (2, 0, 0, 0, -1, 0, 1.5, 0, 0, 0, 1, 0, 0.5, 0, 0, 0, -1, 1, 4)' \), and \( 1_{50} \) is a vector of ones with length 50. 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 \( p(M_i) = 1/2^p \). We use a Metropolis–Hastings algorithm with add/delete steps and random swap proposals as described in Clyde et al. (2011) to sample models.

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>( \pi_j )</th>
<th>Bias</th>
<th>s</th>
<th>RMSE</th>
<th>Bias</th>
<th>s</th>
<th>RMSE</th>
<th>Bias</th>
<th>s</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \gamma_6 )</td>
<td>0.13</td>
<td>0.11</td>
<td>1.48</td>
<td>1.48</td>
<td>0.13</td>
<td>1.56</td>
<td>1.57</td>
<td>-5.04</td>
<td>0.46</td>
<td>5.06</td>
</tr>
<tr>
<td>( \gamma_{17} )</td>
<td>0.13</td>
<td>-0.04</td>
<td>1.60</td>
<td>1.60</td>
<td>-0.24</td>
<td>1.78</td>
<td>1.79</td>
<td>-5.31</td>
<td>0.48</td>
<td>5.33</td>
</tr>
<tr>
<td>( \gamma_4 )</td>
<td>0.14</td>
<td>-0.22</td>
<td>1.67</td>
<td>1.68</td>
<td>-0.09</td>
<td>1.71</td>
<td>1.71</td>
<td>-5.42</td>
<td>0.50</td>
<td>5.44</td>
</tr>
<tr>
<td>( \gamma_{16} )</td>
<td>0.14</td>
<td>-0.00</td>
<td>1.77</td>
<td>1.77</td>
<td>-0.02</td>
<td>1.71</td>
<td>1.71</td>
<td>-5.20</td>
<td>0.54</td>
<td>5.23</td>
</tr>
<tr>
<td>( \gamma_{14} )</td>
<td>0.14</td>
<td>0.03</td>
<td>1.73</td>
<td>1.73</td>
<td>-0.08</td>
<td>1.77</td>
<td>1.77</td>
<td>-5.45</td>
<td>0.48</td>
<td>5.47</td>
</tr>
<tr>
<td>( \gamma_8 )</td>
<td>0.14</td>
<td>-0.19</td>
<td>1.54</td>
<td>1.55</td>
<td>-0.11</td>
<td>1.56</td>
<td>1.56</td>
<td>-5.54</td>
<td>0.49</td>
<td>5.56</td>
</tr>
<tr>
<td>( \gamma_9 )</td>
<td>0.15</td>
<td>-0.05</td>
<td>1.69</td>
<td>1.69</td>
<td>-0.08</td>
<td>1.76</td>
<td>1.76</td>
<td>-5.36</td>
<td>0.58</td>
<td>5.40</td>
</tr>
<tr>
<td>( \gamma_{10} )</td>
<td>0.15</td>
<td>-0.08</td>
<td>1.59</td>
<td>1.59</td>
<td>0.01</td>
<td>1.83</td>
<td>1.83</td>
<td>-5.46</td>
<td>0.49</td>
<td>5.48</td>
</tr>
<tr>
<td>( \gamma_{12} )</td>
<td>0.16</td>
<td>-0.18</td>
<td>1.81</td>
<td>1.82</td>
<td>-0.11</td>
<td>1.89</td>
<td>1.90</td>
<td>-5.15</td>
<td>0.53</td>
<td>5.18</td>
</tr>
<tr>
<td>( \gamma_7 )</td>
<td>0.16</td>
<td>0.25</td>
<td>1.35</td>
<td>1.33</td>
<td>0.68</td>
<td>2.09</td>
<td>2.20</td>
<td>-5.17</td>
<td>0.56</td>
<td>5.20</td>
</tr>
<tr>
<td>( \gamma_5 )</td>
<td>0.19</td>
<td>-0.31</td>
<td>1.81</td>
<td>1.83</td>
<td>-0.15</td>
<td>2.06</td>
<td>2.07</td>
<td>-4.69</td>
<td>0.68</td>
<td>4.73</td>
</tr>
<tr>
<td>( \gamma_3 )</td>
<td>0.27</td>
<td>0.15</td>
<td>1.77</td>
<td>1.78</td>
<td>0.26</td>
<td>2.36</td>
<td>2.38</td>
<td>-3.33</td>
<td>0.70</td>
<td>3.40</td>
</tr>
<tr>
<td>( \gamma_{15} )</td>
<td>0.27</td>
<td>-0.18</td>
<td>2.12</td>
<td>2.13</td>
<td>-0.02</td>
<td>2.45</td>
<td>2.45</td>
<td>-5.86</td>
<td>0.73</td>
<td>5.90</td>
</tr>
<tr>
<td>( \gamma_{20} )</td>
<td>0.38</td>
<td>-0.31</td>
<td>3.26</td>
<td>3.28</td>
<td>-0.81</td>
<td>4.44</td>
<td>4.51</td>
<td>-4.90</td>
<td>1.28</td>
<td>5.07</td>
</tr>
<tr>
<td>( \gamma_{13} )</td>
<td>0.45</td>
<td>-0.11</td>
<td>2.60</td>
<td>2.60</td>
<td>0.06</td>
<td>3.39</td>
<td>3.39</td>
<td>-2.33</td>
<td>1.01</td>
<td>2.54</td>
</tr>
<tr>
<td>( \gamma_{19} )</td>
<td>0.72</td>
<td>0.63</td>
<td>3.18</td>
<td>3.24</td>
<td>0.84</td>
<td>4.33</td>
<td>4.41</td>
<td>2.00</td>
<td>1.25</td>
<td>2.35</td>
</tr>
<tr>
<td>( \gamma_{11} )</td>
<td>0.81</td>
<td>0.67</td>
<td>2.19</td>
<td>2.29</td>
<td>0.24</td>
<td>2.78</td>
<td>2.79</td>
<td>4.74</td>
<td>0.83</td>
<td>4.81</td>
</tr>
<tr>
<td>( \gamma_7 )</td>
<td>1.00</td>
<td>-0.03</td>
<td>0.61</td>
<td>0.61</td>
<td>-0.08</td>
<td>0.68</td>
<td>0.69</td>
<td>0.37</td>
<td>0.03</td>
<td>0.37</td>
</tr>
<tr>
<td>( \gamma_{18} )</td>
<td>1.00</td>
<td>0.02</td>
<td>0.25</td>
<td>0.25</td>
<td>0.01</td>
<td>0.22</td>
<td>0.23</td>
<td>0.15</td>
<td>0.00</td>
<td>0.15</td>
</tr>
<tr>
<td>( \gamma_1 )</td>
<td>1.00</td>
<td>-0.00</td>
<td>0.03</td>
<td>0.03</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>( I(\gamma) )</td>
<td>-</td>
<td>0.06</td>
<td>0.10</td>
<td>0.12</td>
<td>0.03</td>
<td>0.30</td>
<td>0.30</td>
<td>0.31</td>
<td>0.32</td>
<td>0.45</td>
</tr>
</tbody>
</table>

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

For all estimators, we run a Markov-chain for 10,000 iterations, and discard the first 1,000 samples as burn-in. For Horvitz–Thompson, we take the first 8,000 iterations after burn-in and thin the chain by retaining every 8th sample to reduce dependence of draws. For computing \( \hat{C} \) for the Horvitz–Thompson estimator, we run a second independent chain of 1,000 iterations to determine the set \( A \), so that all estimators are based on an equivalent computing effort of 10,000 iterations. Bias, the standard deviation, 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 \( \Delta \), for a scalar quantity, e.g. the variable inclusion indicators \( \gamma_j \), 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, \( \sum_{q=1}^{Q} \{ \text{bias}(\Delta_q) \}^2 / Q \}^{1/2} \). The mean squared error for a scalar quantity is

\[
\text{MSE}(\hat{\Delta}) = \sum_{i=1}^{100} \left( \hat{\Delta}^{(i)} - \Delta \right)^2 / 100
\]

while for vectors we report the average mean squared error over the components.

The Horvitz–Thompson estimator is comparable to the empirical estimator in terms of bias, where the bias in either case is negligible. For Markov chain Monte Carlo sampling in the variable selection framework, samples are drawn from the joint posterior distribution of the variable inclusion indicators and models are sampled with probabilities proportional to the marginal likelihood times the prior model probability. As the renormalized estimator does not incorporate
these sampling probabilities, the estimates for the numerator and denominator in the ratio estimates are both biased. This is illustrated in Table 1, with a suggestion of systematic bias where inclusion probabilities larger than 0.5 are overestimated and inclusion probabilities smaller than 0.5 are underestimated. As the sampler visits the same top models over most of the 100 replicates, the renormalized estimates exhibit low variability and the mean squared error in Table 1 is dominated by the 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 \( \Delta \) 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 which 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. 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 is decreasing at a slower rate than the standard deviation. For Horvitz–Thompson, the absolute relative bias generally decreases with longer runs. In both scenarios the bounds on the bias are fairly tight, however, we only have a practical sample estimate of the bound in the case of the Horvitz-Thompson estimator.

While we would like bias to be small, mean squared error may be more relevant 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 \( \Delta \) 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 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 estimates and the renormalized estimator 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,
or model based methods from the sample survey literature to this Markov chain Monte Carlo sampling context may provide additional improvements.

ACKNOWLEDGMENTS

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


