Parallel Computing to Speed up Whole-Genome Analyses Using Independent Metropolis-Hastings Sampling

H. Cheng[1], **D.J. Garrick**[1], **R.L. Fernando**[1]
[1]Iowa State University, Ames, Iowa, US.

ABSTRACT: Bayesian multiple regression methods are widely used in whole-genome analyses by constructing a Markov chain with a stationary distribution equal to the posterior distribution of unknown parameters. In whole-genome analyses, chains of about 50,000 samples are typically used, for which the computation is intensive. Thus, it is desirable if parallel computing, taking advantage of multiple cores on computers, could be used to speed up Bayesian methods. In this paper, a strategy using Independent Metropolis-Hastings (IMH) sampling to parallelize Markov chain Monte Carlo (MCMC) sampling for whole-genome analyses has been shown. We also propose a strategy to construct the proposal distribution in IMH. Addressing the heavy computational burden associated with Bayesian methods by parallel computing will lead to greater use of these methods.

Keywords: **w**hole-genome analyses; parallel computing; independent Metropolis-Hastings sampling.

INTRODUCTION

Bayesian multiple regression methods are widely used in whole-genome analyses. Inferences from most Bayesian methods are based on samples drawn from Markov chains constructed to have a stationary distribution equal to the posterior distribution of unknown parameters. In whole-genome analyses, chains of about 50,000 samples are typically used, which makes the computation intensive. Thus, it is desirable if parallel computing, taking advantage of multiple computer cores, could be used to speed up Bayesian methods. One approach to speed up Bayesian methods is to implement parallel computing within each sampling, where computations are split up and done in parallel on multiple cores (Cheng et al. (2014)). Further, it is often suggested that samples can be drawn in parallel to obtain a large number of short chains. However, the Ergodic theorem of Markov chain theory states that statistics computed from an increasingly long chain, rather than an increasing number of short chains, converge to those from the stationary distribution (Norris (1997)). Thus, combining a large number of short parallel chains may not yield valid results.

An alternative strategy using Independent Metropolis-Hastings (IMH) sampling has been described by Jacob et al. (2011), where a large number of candidate samples are obtained independently using parallel computing, because the proposal used in IMH is not a Markov process. This proposal contrasts to the commonly-used random walk MH where each candidate depends upon the previous sample. In Metropolis-Hastings, including IMH, the candidate samples are accepted or rejected according to the acceptance probability, which results in a Markov chain. In IMH, the time consuming components of this acceptance probability can also be computed in parallel when the candidate samples are drawn. Thus most computations in IMH are done in parallel before the candidate samples are accepted or rejected to obtain a Markov chain. The acceptance probability in IMH is simply computed as the ratio of two known scalars without waiting for them to be evaluated.

The objective of this study is to show how IMH can be used to parallelize MCMC sampling for whole-genome analyses. In this paper, we first described the IMH algorithm for parallel computing in general. Then a numerical example for the BayesA model (Meuwissen *et al*. (2001)) is given.

ALGORITHM

In IMH, the candidates v_t for the unknown quantities are independently sampled from a proposal distribution $q\,(.)$. Thus the sampling is easily parallelized. The samples obtained in parallel are used to construct a Markov chain $(u_0, u_1, u_2 \ldots)$ with a stationary distribution $\Pi(.)$ of interest, by accepting candidates according to the acceptance probability $r(v_t, u_{t-1}) = w\,(v_t)/w\,(u_{t-1})$, where $w(.)$ are computed in parallel as described below.

1. Generate independent candidates v_t from proposal distribution $q\,(.)$ and compute $w(v_t) = \Pi\,(v_t)/q\,(v_t)$. This step for different v_t can be done in parallel. Given p processors are available for computing, the computing time to obtain k>p candidate samples of v_t is the same as that for obtaining k/p samples with single processor.
2. Once all the candidate samples and corresponding $w(.)$ are available:
   a. Take any one from the set of candidate samples as the first value in the Markov chain: u_0.
   b. Each of the remaining candidates is sequentially considered and accepted with probability $r(v_t, u_{t-1})$ to construct a Markov chain. This acceptance probability $r(v_t, u_{t-1})$ is obtained with little effort because the two scalars in this ratio have already been computed in parallel.

NUMERICAL EXAMPLE

Here we present how IMH can be used to parallelize MCMC sampling for the BayesA model:

$y_i = \mu + \sum X_{ij}\alpha_j + e_i$,

where y_i is the phenotype for individual i, μ is the overall mean, X_{ij} is the genotype covariate at locus j for animal i (coded as 0,1,2), α_j is the average allele substitution effect for locus j and e_i is the random residual effect for individual i.

The prior for μ is a constant. The prior for e_i is a normal distribution with mean zero and variance σ_e^2, where σ_e^2 follows a scaled inverted chi-square distribution with degree of freedom υ_e and scale s_e^2. The prior for α_j is a normal distribution with mean zero and variance σ_j^2, where σ_j^2 follows a scaled inverted chi-square distribution with degree of freedom υ_α and scale s_α^2.

Here the target distribution $\Pi(.)$ is the joint posterior distribution of all unknown parameters $f(\boldsymbol{\alpha}, \mu, \boldsymbol{\xi}, \sigma_e^2|\boldsymbol{y})$, where $\boldsymbol{\xi} = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots)$. This target distribution can be written as the product of the Gaussian likelihood and priors for unknown parameters (Sorensen and Gianola. (2002)). The proposal distribution $q(.)$ we used is constructed as

$$q(.) = f\left(\mu, \boldsymbol{\alpha}|\tilde{\boldsymbol{\xi}}, \widetilde{\sigma_e^2}, \boldsymbol{y}\right) \prod f\left(\sigma_j^2|\boldsymbol{\alpha}, \upsilon_\alpha, s_\alpha^2, \boldsymbol{y}\right) \times f(\sigma_e^2|\mu, \boldsymbol{\alpha}, \upsilon_e, s_e^2, \boldsymbol{y}),$$

where $\tilde{\boldsymbol{\xi}}$ is estimated using a few iterations of the EM algorithm (Sun et al. (2012)), $\widetilde{\sigma_e^2}$ is the mean of the prior for σ_e^2. Above, the proposal distribution $f\left(\mu, \boldsymbol{\alpha}|\tilde{\boldsymbol{\xi}}, \widetilde{\sigma_e^2}, \boldsymbol{y}\right)$ for μ and $\boldsymbol{\alpha}$ is a multivariate normal distribution with mean $\boldsymbol{C^{-1}W'y}$ and variance $\boldsymbol{C^{-1}}\widetilde{\sigma_e^2}$, where $\boldsymbol{C}$ is $\boldsymbol{W'W} + \boldsymbol{D^{-1}}\widetilde{\sigma_e^2}$, $\boldsymbol{W} = [\boldsymbol{1}, \boldsymbol{X}]$ and $\boldsymbol{D^{-1}}$ is a diagonal matrix of elements $0, 1/\widetilde{\sigma_1^2}, 1/\widetilde{\sigma_2^2}, 1/\widetilde{\sigma_3^2} \ldots$. The proposal distributions of σ_j^2 and σ_e^2 are all scaled inverted chi-square distributions; for σ_j^2, the scale parameter is $(\alpha_j^2 + \upsilon_\alpha s_\alpha^2)/\upsilon_\alpha$ and the degrees of freedom parameter is $\upsilon_\alpha + 1$; for σ_e^2, the scale parameter is $(\boldsymbol{e'e} + \upsilon_e s_e^2)/\upsilon_e$ and the degrees of freedom parameter is $\upsilon_e + n$, where n is the number of observations and $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{1}\mu - \sum \boldsymbol{X_j}\alpha_j$. A large number p of candidate samples from this proposal distribution were obtained in parallel, and for each sample v_t, the quantity $w(v_t) = \Pi(v_t)/q(v_t)$ was also computed in parallel. Once all the candidate samples and corresponding $w(.)$ were obtained, a Markov chain with target distribution $f(\boldsymbol{\alpha}, \mu, \boldsymbol{\xi}, \sigma_e^2|\boldsymbol{y})$ was constructed by accepting a candidate into the chain with probability $r(v_t, u_{t-1}) = w(v_t)/w(u_{t-1})$ in each iteration.

Simulated data were used to test the performance of this approach. The data used for training consisted of 1,000 observations with 1,000 SNP genotypes and phenotypes simulated to have a heritability of 0.5. A chain of length 1,000 was used to estimate marker effects. An independent dataset of equal size was used for testing. The prediction accuracy of the simulated breeding value in the test dataset was 0.90 using IMH. In contrast, prediction accuracy from single-site Gibbs sampling was only 0.7 even after a chain of length 10,000.

DISCUSSION

One approach to speedup Bayesian methods is to parallelize the computations such as dot products and vector additions that are needed to draw samples (Cheng et al. (2014)). In this paper, we have proposed another approach where samples are obtained in parallel. The key principle in this approach is the use of an independent proposal to draw candidate samples in the Metropolis-Hastings algorithm. In most Metropolis-Hastings implementations, the proposal is constructed using a random walk. While such proposals are easy to construct, they cannot be used in parallel. This includes the Gibbs sampler. In contrast, while independent proposals can be used in parallel, it is not straightforward to construct an independent proposal that will result in a chain with good mixing performance. Ideally, one would like to use a proposal close to the target distribution. In general, this may be difficult, especially for high-dimensional problems. However, as described below, the "cycle" in the Gibbs sampler can be broken to produce an independent sampler when marginal modes for certain parameters can be computed and sufficient data are available.

Suppose we want to draw samples from $f(x_1, x_2, x_3)$, which does not have a closed form. To devise an independent sampler, it is instructive to begin with the Gibbs sampler. In the Gibbs sampler, samples for x_1, x_2, x_3 are drawn from $f(x_1^t|x_2^{t-1}, x_3^{t-1}), f(x_2^t|x_3^{t-1}, x_1^t)$ and $f(x_3^t|x_1^t, x_2^t)$. However, this is not an independent sampler. One strategy to turn this into an independent sampler is to draw x_1 from $f(x_1|\tilde{x}_2, \tilde{x}_3), x_2$ from $f(x_2|x_1, \tilde{x}_3)$ and x_3 from $f(x_3|x_1, x_2)$, where $\tilde{x}_2, \tilde{x}_3$ are the posterior mode of x_2, x_3. This can be justified by showing that $f(x_1|\tilde{x}_2, \tilde{x}_3)$ is approximately equal to $f(x_1)$ and $f(x_2|x_1, \tilde{x}_3)$ is approximately equal to $f(x_2|x_1)$ if $f(x_2, x_3)$ and $f(x_3)$ is reasonably peaked (Gianola et al. (1986)). This leads to a proposal that approximates the target, namely $f(x_3|x_1, x_2)f(x_2|x_1)f(x_1)$. In our numerical example, this strategy is used to construct the proposal distribution, where components of the posterior mode of $\tilde{\boldsymbol{\xi}}$ approximated from a few iterations of the EM algorithm is used to break the "cycle" in the Gibbs sampler.

Instead of using these independent proposals just once serially, different permutations of the same samples could be used serially in parallel, say p times. Statistics computed from p chains of k samples might have smaller variances than from a single chain of k samples. A more elaborate block IMH with different choices of permutations has been studied by Jacob et al. (2011).

An added benefit of MCMC-based Bayesian methods over methods such as the EM algorithm is that they provide posterior distributions for parameters of interest, which are used to make inferences about these parameters. For example, the posterior distribution of breeding values can be used to obtain theirs accuracies rather than approximations that are typically used. In Bayesian whole-genome analyses, the posterior distributions can be obtained for the proportion of variance attributed to any genomic region to detect causal loci (Fernando et al. (2013)).

CONCLUSION

In this paper, we have described how IMH can be used for parallel computing. Since the key principle in this approach is the use of an independent proposal, we have discussed a strategy to construct an independent proposal. A numerical example following this strategy to construct the proposal distribution has been shown to perform better than the single-site Gibbs sampler. Reducing the computing time associated with MCMC-based Bayesian methods by parallel computing will lead to greater use of these methods, which have many attractive features for whole-genome analyses.

ACKNOWLEDGMENTS

This work was supported by the US Department of Agriculture, Agriculture and Food Research Initiative National Institute of Food and Agriculture Competitive grant no. 2012-67015-19420 and by National Institutes of Health grant R01GM099992.

LITERATURE CITED

Cheng, H., Fernando, R., Garrick, D. (2014). Plant and Animal Genome XXII. P1061.

Fernando, R.L., Garrick, D.J. (2013). Human press, chapter 10. ISBN 978-1-62703-446-3

Gianola, D., Foulley, J.L., Fernando, R.L. (1986). Genet Sel Evol. 18(4):485-498.

Jacob, P., Robert, C.P., Smith, M.H. (2011). Journal of Computational and Graphical Statistics, 20(3):616–635.

Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). Genetics, 157:1819–1829.

Norris, J.R. (1997). Cambridge University Press, New York. ISBN 0-521-63396-6.

Sorensen, D., Gianola, D. (2002). Springer. ISBN 0-387-954406.

Sun, X., Qu, L., Garrick, D.J., Dekkers, J.C.M., Fernando, R.L. (2012). PlOS One. 7(11):e49157(doi: 10.1371/journal.pone.0049157).