BAYESIAN INFERENCE ON MULTILOCUS GENOTYPIC EFFECTS USING A GIBBS SAMPLER

Junhee An, Younyoung Kim, and Chaeyoung Lee*

Department of Bioinformatics and Life Science, Soongsil University
511 Sangdo-dong, Dongjak-gu, Seoul 156-743, Republic of Korea
**Email: clee@ssu.ac.kr*

Simultaneous analysis of multiple genetic variants is an essential strategy for understanding genetic dissection of complex traits, focusing epistasis along with additive and dominance effects of individual genes. Although phenotypic variation for complex traits might be largely explained by epistasis, most analyses have excluded the possibility of epistasis, especially with lack of individual locus effects. The conventional models for estimating all the possible epistatic effects have a decisively vulnerable point of potentially low power or often nonestimable statistics due to a large number of parameters. Restricted partition method (RPM), a recently developed nonparametric approach for estimating epistasis, overcame the drawback but has both biologically and statistically undesirable properties caused by grouping genotypes. A Bayesian method using a Gibbs sampler for estimating epistasis for complex continuous traits was developed to overcome such problems. This method was devised to draw inferences on multilocus genotypic effects by a Bayesian approach based on their marginal posterior distributions and to attain the marginalization of the joint posterior distribution through Gibbs sampler as a Markov chain Monte Carlo. A simulation study revealed that the Bayesian method using a Gibbs sampler was superior to the currently utilized MDR. Especially, prediction errors substantially decreased under various environmental exposures by the Bayesian method using a Gibbs sampler. The programs would be available for both Gamma and Chi-square prior distributions.

1. INTRODUCTION

Phenotypic variability for complex quantitative traits might be largely explained by epistasis, gene-by-gene interaction effects. Simultaneous analysis of multiple genes by estimating epistasis along with additive and dominance effects of individual genes is an essential strategy for understanding genetic dissection of complex traits. Nevertheless, the potential interaction effects have not been analyzed in many genetic studies of complex traits because of the increasing number of genetic interaction parameters.[1] This problem is serious especially when individual locus effects lack. The conventional models for estimating all the possible epistatic effects have a decisively vulnerable point of potentially low power or often nonestimable statistics because of a large number of interaction parameters. In order to overcome the drawback, a nonparametric approach was recently developed for estimating epistasis, and it is called restricted partition method (RPM).[2] Although the RPM overcame the drawback, it has both biologically and statistically undesirable properties caused by grouping genotypes.[3]

More recently, a Bayesian method using a Gibbs sampler for estimating epistasis for complex continuous traits was proposed to overcome such problems, and we call it Bayesian inference by Gibbs sampling on multilocus genotypic effects (BIGSMGE). In the current study, we generalized the BIGSMGE.

2. BAYESIAN INFERENCE BY GIBBS SAMPLING ON MULTILOCUS GENOTYPIC EFFECTS

In the BIGSMGE, inferences about unknown effects of multilocus genetic interaction was based on their marginal posterior distribution in a Bayesian framework. The marginalization of the joint posterior distribution was attained through Gibbs sampling.

2.1. Posterior distribution

A general formula for the joint posterior distribution of all parameters was first derived using the Bayes theorem. Previously, inverse Gamma distributions (G^{-1}) were assumed for the priors of variance components for both genetic interaction effects and residuals.[3] This was because the use of flat priors for variance components might lead to inferences based on theoretically nonexistent posterior distributions.[4] Inverse chi-square distributions (X^{-2}) as another prior distribution family

[*] Corresponding author.

was also incorporated with the method in the current study. Full conditional posterior distribution was subsequently derived by obtaining the posterior distribution of each parameter given the data and all other parameters.

2.2. Gibbs sampling

Gibbs sampling was applied as a numerical integration method based on a Markov chain Monte Carlo. We conducted the intensive iterative sampling from the consecutively updated full conditional posterior distributions derived in the previous section. For example,

1. Set arbitrary initial values for fixed effects, random multilocus genotypic effects, genetic variance component, and residual variance component.
2. Generate residual variance component using its full conditional posterior distribution (G^{-1} or X^{-2}), and update it.
3. Generate genetic variance component using its full conditional posterior distribution (G^{-1} or X^{-2}), and update it.
4. Generate multilocus genotypic effects using its full conditional posterior distribution (N; normal distribution), and update them. Or generate and update a multilocus genotypic effect, and subsequently generate and update one multilocus genotypic effect at a time.
5. Generate fixed effects using the following full conditional posterior distribution, and update it.
6. Repeat the steps 2, 3, 4, and 5.

An intensive iteration is required to get the reasonable estimates of multilocus genotypic effects. For the default, the Gibbs sampler is run 12,000 iteration rounds. The first 2,000 rounds are discarded as a warming-up period before convergence. The default thinning interval of 10 rounds is applied for retaining sampled values that reduce lag correlation among consecutive samples.

2.3. Estimation of genotypic effect

The posterior mean estimate of the multilocus genotypic effect is calculated as the mean of its values generated from the post warming-up rounds of Gibbs sampling.

3. SIMULATION

Data simulation was conducted to see if BIGSMGE fit the data with epistasis and to compare BIGSMGE to RPM.

3.1. Monte Carlo simulation

Quantitative trait was simulated generated by adding a genotypic mean and an error generated from Normal distributions. The simulation was devised with balanced and unbalanced designs. Regardless of the balance of the design, the average sample size for each genotype was 5, 10, 20, or 50. In another simulation, a fixed effect was additionally included in the phenotypic value. A total of 2,000 data sets were simulated from combinations of within genotype variance, sample size, balance of design, number of loci, and existence of fixed effect. Fifty replicates were simulated for each set. A random number generator based on Box-Muller method was used to generate random Gaussian deviates.[5]

3.2. Results

The simulated data were analyzed by BIGSMGE with the prior of G^{-1}, BIGSMGE with the prior of X^{-2}, and RPM. Mean square prediction errors (MSPEs) were estimated for genotypic means by each method, and the best method was selected for each simulated data set (Tables 1 and 2). The MSPE obtained from BIGSMGE were smaller ($P<0.05$) than the corresponding MSPE from RPM regardless of the simulated designs such as within genotype variance, sample size, and balance of design. Correspondingly, the proportion of RPM selected as the best method was negligible. Comparing the priors, BIGSMGE with G^{-1} was somewhat superior to that with X^{-2}.

Table 1. The best method among BIGSMGE with the prior of G^{-1} (M1), BIGSMGE with the prior of X^{-2} (M2), and RPM (M3) fitting data simulated without fixed effects. D1=balanced and small sized data, D2=unbalanced and small sized data, D3=balanced and large sized data, D4=unbalanced and large sized data

(%)

	D1	D2	D3	D4
M1	51.1	55.8	57.9	62.1
M2	48.9	44.1	42.1	37.9
M3	0.0	0.1	0.0	0.0

Table 2. The best method among BIGSMGE with the prior of G^{-1} (M1), BIGSMGE with the prior of X^{-2} (M2), and RPM (M3) fitting data simulated with fixed effects. D1=balanced and small sized data, D2=unbalanced and small sized data, D3=balanced and large sized data, D4=unbalanced and large sized data

(%)

	D1	D2	D3	D4
M1	54.3	60.2	58.9	59.5
M2	45.7	39.8	41.1	40.5
M3	0.0	0.0	0.0	0.0

4. DISCUSSION

The simulation study revealed a larger MSPE using RPM than using BIGSMGE. This might be due to the information loss from grouping in RPM. Furthermore, estimating epistasis by RPM would not have viable implication to biological epistasis because biology for the grouping is hardly explained.

Using the prior of inverse Gamma distributions was preferable in the comparison of the priors in BIGSMGE. There was, however, any specific trend along with the simulation designs. Especially, this was true with a small degree of belief. Furthermore, the differences of MSPE were not statistically significant in more than 90% of data sets (P>0.05). The programs would be available for both prior distributions at the homepage of the Laboratory of Statistical Genetics, Department of Bioinformatics and Life Science, Soongsil University (http://clee11.cafe24.com/).

Acknowledgments

This study was supported by a grant from BioGreen 21 Program, Rural Development Administration, Republic of Korea (Grant Code No. 20080401034021).

References

1. Frankel WN, Schork NJ. Who's afraid of epistasis?, *Nat Genet* 1996; **14**: 371-373.
2. Culverhouse R, Klein T, Shannon W. Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol* 2004; **27**: 141-152.
3. Lee C, Park J. Estimation of epistasis among finite polygenic loci for complex traits with a mixed model using Gibbs sampling. *J Biomed Inform* 2007; **40**: 500-506.
4. Hobert JP, Casella G. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J Am Stat Assoc* 1996; **91**: 1461-1473.
5. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge, 1992.