

# MCMC Likelihoods for Population Genetics

Eric C. Anderson and Elizabeth A. Thompson  
*Department of Statistics, University of Washington*  
*Box 354322*  
*Seattle, WA 98195 U.S.A.*  
*eriq@stat.washington.edu, thompson@stat.washington.edu*

## 1. The Problem and the Likelihood

It is important to consider the genetic health of small or threatened populations. A key quantity in this regard is the effective number of breeding adults,  $N_a$ , each generation. Population geneticists define  $N_a$  by comparison to an ideal population in which the genes of the next generation are sampled with replacement from the current one.  $N_a$  is usually smaller than the census number of breeding adults,  $C$ . Often  $C$  can be measured, and then it is valuable to know  $\lambda = N_a/C$ . It is difficult to measure  $\lambda$  by demographic methods, particularly for some species of fish and amphibians, which produce thousands of offspring, very few of which survive to adulthood. Instead,  $\lambda$  may be estimated from temporal changes of allele frequencies sampled from the population. We present a Monte Carlo approach for approximating the likelihood curve for  $\lambda$ .

Assume a discrete-generation, semelparous population with  $C_t$  haploid individuals reproducing at  $t$ , giving rise to  $C_{t+1}$  individuals at  $t+1$ . We take genetic samples of size  $S_1, \dots, S_T$  (assume  $S_1 > 0$ ,  $S_T > 0$ ) individuals, and find counts of the  $k$  different allelic types at a locus,  $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_T)$  where  $\mathbf{Y}_t = (Y_{t1}, \dots, Y_{tk})$ . The population at  $t$  is modelled as  $\lfloor \lambda C_t \rfloor$  ideally-reproducing adults, where  $\lfloor x \rfloor$  is the largest integer  $\leq x$ . Underlying the data are latent allele counts  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_T)$  with  $\mathbf{X}_t = (X_{t1}, \dots, X_{tk})$ .  $\{\mathbf{X}_t, t \geq 1\}$  is a first-order Markov chain with transition probabilities  $P_\lambda(\mathbf{X}_{t+1}|\mathbf{X}_t)$  being multinomial with cell probabilities  $\mathbf{X}_t/\lfloor \lambda C_t \rfloor$  and number of trials  $\lfloor \lambda C_{t+1} \rfloor$ . The genetic data at time  $t$  are samples from the gamete pool produced by the  $\lfloor \lambda C_t \rfloor$  adults. Hence, for  $S_t > 0$ ,  $P_\lambda(\mathbf{Y}_t|\mathbf{X}) = P_\lambda(\mathbf{Y}_t|\mathbf{X}_t)$  is multinomial with parameters  $\mathbf{X}_t/\lfloor \lambda C_t \rfloor$  and  $S_t$ . For  $S_t = 0$ ,  $P_\lambda(\mathbf{Y}_t|\mathbf{X}_t) \equiv 1$ . Summing out the nuisance parameters  $\mathbf{X}_1$  over an uninformative prior  $\pi(\mathbf{X}_1)$  gives the likelihood

$$L(\lambda) = P_\lambda(\mathbf{Y}) = \sum_{\mathbf{X}} P_\lambda(\mathbf{Y}, \mathbf{X}) = \sum_{\mathbf{x}_0, \dots, \mathbf{x}_T} \pi(\mathbf{X}_1) P_\lambda(\mathbf{Y}_1|\mathbf{X}_1) \prod_{t=2}^T P_\lambda(\mathbf{X}_t|\mathbf{X}_{t-1}) P_\lambda(\mathbf{Y}_t|\mathbf{X}_t).$$

With  $k = 2$ , the sum over  $\mathbf{X}$  may be evaluated exactly. With larger  $k$ , however, the huge space of possible  $\mathbf{X}_t$ 's makes this infeasible.

## 2. Monte Carlo Likelihood

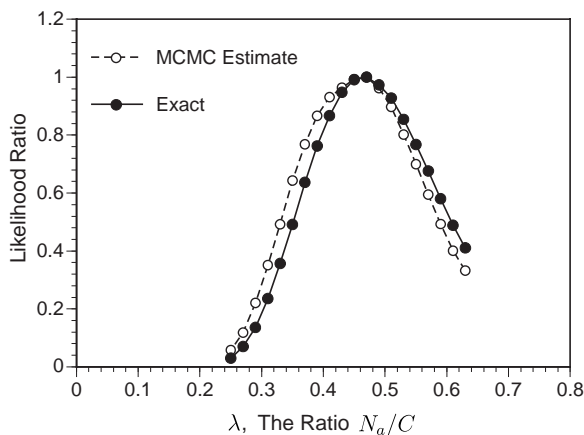
To obtain an efficient Monte Carlo estimate of  $L(\lambda)$ , we consider the likelihood ratios  $L(\lambda)/L(\lambda_0)$ , (Thompson and Guo 1991, Geyer and Thompson 1992)

$$\frac{L(\lambda)}{L(\lambda_0)} = \frac{P_\lambda(\mathbf{Y})}{P_{\lambda_0}(\mathbf{Y})} = \sum_{\mathbf{X}} \frac{P_\lambda(\mathbf{Y}, \mathbf{X})}{P_{\lambda_0}(\mathbf{Y}, \mathbf{X})} P_{\lambda_0}(\mathbf{Y}|\mathbf{X}) = E_{\lambda_0} \left( \frac{P_\lambda(\mathbf{Y}, \mathbf{X})}{P_{\lambda_0}(\mathbf{Y}, \mathbf{X})} \middle| \mathbf{Y} \right)$$

which may be estimated by  $\frac{1}{m} \sum_{i=1}^m P_\lambda(\mathbf{Y}, \mathbf{X}^{(i)})/P_{\lambda_0}(\mathbf{Y}, \mathbf{X}^{(i)})$  where each  $\mathbf{X}^{(i)}$  is realized from  $P_{\lambda_0}(\mathbf{X}|\mathbf{Y})$ . This is an efficient Monte Carlo estimator of the likelihood ratio provided  $\lambda$  is near

to  $\lambda_0$ . Independent samples of  $\mathbf{X}$  are not available because  $P_{\lambda_0}(\mathbf{X}|\mathbf{Y})$  is known only up to scale. Instead,  $\mathbf{X}^{(i)}$  are realized from a Markov chain with limit distribution  $P_{\lambda_0}(\mathbf{X}|\mathbf{Y})$  using a component-wise Metropolis-Hastings algorithm (Hastings 1970): Start with initial values of  $\mathbf{X}$ ; Select a pair  $(X_{tk}, X_{t\ell})$ ,  $k \neq \ell$  at random from  $\mathbf{X}$ ; Propose updating the pair to  $(X_{tk}^*, X_{t\ell}^*) = (X_{tk} - w, X_{t\ell} + w)$ , where  $w$  is a random integer drawn with probability  $q(w; X_{tk}, X_{t\ell})$ ; accept the proposal with probability  $\min\{1, [q(-w; X_{tk}^*, X_{t\ell}^*)P_{\lambda_0}(\mathbf{Y}, \mathbf{X}^*)]/[q(w; X_{tk}, X_{t\ell})P_{\lambda_0}(\mathbf{Y}, \mathbf{X})]\}$ . After initial updates for burn-in,  $\mathbf{X}^{(i)}$ 's are sampled as the state of  $\mathbf{X}$  at a spacing of  $u$  updates.

Estimating a curve for  $L(\lambda)$ , the range of  $\lambda$ 's of interest may be large. In such cases it does not suffice to realize  $\mathbf{X}^{(i)}$ 's under a single  $\lambda_0$ . We sample from several chains, each indexed by a different  $\lambda_0$ ,  $\lambda_0 \in \Lambda$ . Geyer (1994) describes a reverse logistic regression method for reweighting the samples from each chain and estimating the whole likelihood surface.



The figure at left shows the estimated likelihood curve (open circles) from a simulated dataset for which the exact likelihood curve (filled circles) can be computed. The data were simulated for 20 loci with  $k = 2$ ,  $\lambda = .4$ ,  $C_t$  varying between 90 and 130,  $S_t = 200$ , and  $\mathbf{X}_1$ 's drawn from a uniform distribution. The estimated curve is the product of likelihood ratios estimated for each locus with  $\Lambda = \{.25, .27, \dots, .61, .63\}$ ,  $m = 10,000$  and  $u = 1,000$ .

**Figure 1. Estimated and Exact Likelihoods**

We are currently investigating more effective MCMC updates, incorporating uncertainty in census size estimates, and extending the approach to more complex models, including age-structured populations with overlapping generations.

This research was supported in part by NSF grants BIR-9256537 and BIR-9807747.

## REFERENCES

- Geyer, C.J., 1994. Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Tech. Rep. 568r, School of Statistics, University of Minnesota.
- Geyer, C.J. and E.A. Thompson, 1992. Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* 54:657–699.
- Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.
- Thompson, E.A. and S.-W. Guo, 1991. Monte Carlo evaluation of likelihood ratios. *IMA J. Math. Appl. Med. & Biol.* 8:149–169.

## RÉSUMÉ

*Le ratio de l'effectif par rapport au nombre total d'adultes reproducteurs est un paramètre important dans la structure d'une population. Nous présentons une approche de vraisemblance de Monte Carlo permettant d'estimer ce paramètre à partir de données génétiques, au niveau de plusieurs loci multialléliques.*