Composite likelihood estimation applied to Single Nucleotide Polymorphism (SNP) data.

Rasmus Nielsen
University of Copenhagen, Center for Bioinformatics
Universitetsparken 15, Building 10
2100 Kbh , Denmark
rasmus@binf.ku.dk

Carsten Wiuf
University of Aarhus, Bioinformatics Research Center
Høegh-Guldbergsgade 10, Building 090
8000 Aarhus C, Denmark
wiuf@birc.au.dk

Single Nucleotide Polymorphisms (SNPs) are positions in a DNA sequence that are variable among different individuals of a population. Many of the large scale genomic efforts have currently switched from direct DNA sequencing to the production of SNP data in humans. To human geneticists and population geneticists, the availability of large scale SNP data is exciting because such data can be used to answer many important scientific questions regarding recombination and mutation in the human genome, and regarding the demographics and ancestry of human populations. Unfortunately, there are very few appropriate statistical tools available for the population genetic analysis of SNP data. Let ψ be a vector of parameters, and let X be the binary data matrix with columns corresponding to individual SNPs and rows corresponding to DNA sequences. The matrix is assumed binary because in human SNP data more than two different nucleotides (among $\{A, C, T, G\}$) are rarely observed in a specific site. The full likelihood function is then given by

$$p(X \mid \psi) = \int_{\mathcal{G}} p(X \mid G, \psi) p(G \mid \psi) dG, \tag{1}$$

where $G \in \mathcal{G}$ is a stochastic graph (called the *ancestral recombination graph*) specifying the genealogical (ancestral) relationship of the DNA sequences. The genealogy of a single site in the DNA sequences is described by a stochastic binary tree, the coalescent (Kingman 1982), and the graph G can thus be seen as a collection of trees, one for each site in the sequences (Hudson 1983, Griffiths and Marjoram 1996, Wiuf and Hein 1999). The trees might vary between sites becuase of genetic recombination. For some important models, known as *neutral models*, the parameters specifying G are disjoint from the parameters specifying X conditional on G. For an overview of coalescent theory, see Hudson (1991), and Hein et al (2005).

Even under the most simple (reasonable) models, the likelihood function cannot be evaluated analytically for just a few SNPs. Instead population geneticists have been using computationally intensive methods to evaluate the likelihood function such as MCMC (e.g. Nielsen 2000, Kuhner et al. 2000) and sequential importance sampling (e.g. Fearnhead and Donnelly 2001, Griffiths and Marjoram 1996, Stephens and Donnelly 2000). Unfortunately, these methods have not proven fast and efficient enough to scale up to real data. The fundamental problem is that the expected number of nodes in G increases exponentially with the recombination rate parameter, which is approximately linearly dependent on the length of the DNA sequences. This makes methods that stochastically integrate over $\mathcal{G}$ inapplicable to large scale genomic data. To overcome this problem, there has been several recent efforts to devise approximate methods for inference. Among the most promising methods are the composite likelihood methods suggested by Hudson (2001), Kim and Stephan (2002) and McVean et al. (2002). In these

methods the likelihood function is calculated marginally for one or a small number of SNPs, or for a small region of the DNA sequences. A composite likelihood function is then formed by taking the product of these marginal likelihood functions, thereby treating SNPs or regions as being independent. In some cases it is known that composite likilihood methods provide consistent estimators of population parameters (e.g., Fearnhead 2003). We will here give two examples of the use of composite likelihood estimators for neutral models.

Estimation of θ under Kingman's coalescent

We are interested in estimating the scaled mutation rate θ ($= 4N\mu$, where N is the effective population size and μ is the mutation rate per site per generation, Watterson 1975). The likelihood in a variable site (i) is given by

$$E_\psi((1 - e^{-\theta T_i/2})e^{-\theta(T-T_i)/2}), \tag{2}$$

where ψ is a vector of parameters not containing θ, T is the length of the tree (G) relating the sequences in site i, and T_i is the sum of the length of all branches in G in which a single mutation would induce the observed data pattern in site i. Because the per site mutation rate is very low, it is common to consider equation (2) for $\theta \approx 0$. As θ becomes small, equation (2) tends to $\theta E_\psi(T_i)/2$. Likewise, the probability that a site is invariable is

$$E_\psi(1 - e^{-\theta T/2}) \tag{3}$$

which tends to $1 - \theta E_\psi(T)/2$ as θ becomes small. The composite likelihood function, the product over all sites, is for small scaled mutation rates approximately proportional

$$\theta^S \left(1 - \frac{\theta E_\psi(T)}{2}\right)^{k-S} \tag{4}$$

where S is the number of variable sites and k the total number of sites. The maximum composite likelihood estimate $\hat{\theta}_\psi$ of θ is then

$$\hat{\theta}_\psi = \frac{2S}{E_\psi(T)k}, \tag{5}$$

and an estimator of the scaled mutation rate for the whole DNA sequences is $k\hat{\theta}_\psi$ (also known as Watterson's estimator). It is usually derived as a method of moments estimator. In the case of Kingman's coalescent, ψ is zero-dimensional and $E_\psi(T) = 2\sum_{j=1}^{n-1} 1/j$, while for generalizations of Kingman's coalescent, the expectation typically depends on parameters describing the shape of the genealogy. In the latter case $\hat{\theta}_\psi$ is a profile estimator.

The expectation of $\hat{\theta}_\psi$ is (Watterson 1975, Hudson 1991)

$$E_\psi(\hat{\theta}_\psi) = \frac{2E_\psi(S)}{kE_\psi(T)} = \theta, \tag{6}$$

and a useful approximation to the variance is given by (Kaplan and Hudson 1985, Hudson 1991)

$$\mathrm{Var}_\psi(\hat{\theta}_\psi) \approx \frac{\theta}{kE_\psi(T)} + \frac{2\theta^2}{k^2 E_\psi(T)^2} \int_0^k (k - x) f_{n,\psi}(x) dx, \tag{7}$$

where $f_{n,\psi}(x)$ is the covariance of the total tree lengths in two positions x sites away. It depends on the sample size n. Equation (6) and (7) rely on the Poisson nature of the mutation process. For many interesting models (e.g. Kingman's coalescent) $f_{n,\psi}(x)$ decays like $1/x$, and

$$\mathrm{Var}_\psi(\hat{\theta}_\psi) \approx \frac{C_{n,\psi} \log(k)}{k} \tag{8}$$

for large k (Kaplan and Hudson 1985, Wiuf and Nielsen, unpublished results), where $C_{n,\psi}$ is a constant depending on n and ψ. Thus, $\hat{\theta}_\psi$ is unbiased and consistent for $k \to \infty$. It is possible to show that $\hat{\theta}_\psi$ is as good as possible, i.e. its variance decays at the same rate as the maximum likelihood estimator (Wiuf and Nielsen, unpublished results).

Application to the estimation of demographic parameters

In this section we discuss the properties of the the composite likelihood estimator of the scaled migration rate between a pair of populations (e.g. Hudson 1991). In a single SNP site, the data consists of the vector $x = (x_{00}, x_{01}, x_{10}, x_{11})$, where x_{00} is the number of copies of the first allele in the first subpopulation, x_{01} is the number of copies of the second allele in the first subpopulation, and so forth.

The basic model for G is a model of symmetric migration between two subpopulations that may have different sizes. Technically, the model of G is two coupled coalescent processes, one describing the ancestral relationships between sequences in the first subpopulation, the other describing the ancestral relationship between sequences in the second subpopulation. The two processes are coupled through migration: Each sequence migrates to the other subpopulation at rate $M = Nm$, where m is the probability of migration per sequence per generation. M is thus the expected number of migrants entering a subpopulation in a given generation. Additionally, the model has a parameter, f, describing the ratio of the two subpopulation sizes; i.e. $\psi = (M, f)$.

Many data sets that are generated today consist only of variable sites. This is a consequence of SNP technology and the way SNPs normally are being typed in large samples. It has the further important statistical consequence that equation (1) should be interpreted as being *conditional* on all sites being variable. If only variable sites are observed (or typed), it is part of the sampling strategy and must be reflected in the likelihood.

The effect of this procedure is to eliminate θ, which here is considered a nuisance parameter. As in Nielsen (2000), the likelihood function calculated for a single SNP site (i) can then be expressed as

$$\frac{E_\psi(T_i)}{E_\psi(T)} \tag{9}$$

(again assuming θ is small). These expectations cannot be calculated analytically for the present model, but they can be approximated very fast using simulation. Simulate k trees under the previously defined coalescence process for a particular value of ψ, then a simulation consistent estimate of the likelihood function can be obtained as

$$\frac{\sum_{j=1}^{k} T_i(j)}{\sum_{j=1}^{k} T(j)} \tag{10}$$

where $T(j)$ and $T_i(j)$ are the values of T and T_i, respectively, for simulated tree j. The likelihood surface can then be estimated by repeating this procedure for different values of ψ and/or by the application of various importance sampling schemes for generating surfaces from single value simulations (not shown). Notice that the same set of simulations can be used for multiple SNPs making this approach very computationally attractive. Wiuf and Nielsen (unpublished results) have shown that the maximum composite estimators of M and f are consistent for large number of SNPs. Other properties of the estimators, such as the precision, can easily be addressed through simulation.

This simulation procedure was applied to 12,836 SNPs from the Seattle SNP database (NHLBI Program for Genomic Applications, UW-FHCRC, Seattle, WA, http://pga.gs.washington.edu).

Each SNP was typed for 46 Caucasian Americans chromosomes and 48 African American chromosomes. For the purpose of this study, these two groups are considered to be subpopulations in the population genetic sense. The likelihood surface was estimated on a grid of 400 values of ψ using a value of $k = 10^6$ for each value of ψ. Composite maximum likelihood estimates of $\hat{M} = 2.3$ and $\hat{f} = 5.8$ were obtained, indicating that the effective African (American) population size is considerably larger than the European (American) population and that there is considerable amounts of gene flow between these subpopulations. The estimates are in good accordance with previous estimates of demographic parameters for these populations.

The composite likelihood procedure is relative fast and easy to implement for many population genetic models of interest. The fact that the composite likelihood methods often have desirable statistical properties (such as consistency) and are easy and fast to implement make them attractive vehicles for statistical inference of population genetic parameters based on SNP data.

REFERENCES

1. Fearnhead, P. 2003 *Theor. Pop. Biol.* **64:** 67-79.
2. Fearnhead, P. and Donnelly, P. 2001 *Genetics* **159:** 1299-1318.
3. Griffiths, R. C. and Marjoram, P. 1996 *J. Comp. Biol* **3:** 479-502.
4. Hein, J., Scheirup, M. H., and Wiuf, C. 2005 *Gene Genealogies, Variation, and Evolution.* Oxford University Press.
5. Hudson, R. R. 1983 *Theor. Pop. Biol.* **23:** 183-201.
6. Hudson, R. R. 1991 *Oxford Surveys in Evolutionary Biology* **7:** 1-49.
7. Hudson, R. R. 2001 *Genetics* **159:** 1805-1817.
8. Kaplan, N. and Hudson, R. R. 1985 *Theor. Pop. Biol.* **28:** 382-396.
9. Kim, Y. and Stephan, W. 2002 *Genetics* **155:** 1415-1427.
10. Kingman, J. F. C. 1982 *Stoch. Proc. Appl.* **13:** 235-248.
11. Kuhner, M. K., Yamato, J., and Felsenstein, J. 2000 *Genetics* **156:** 1393-1401.
12. McVean, G., Awadalla, P., and Fearnhead, P. 2002 *Genetics* **160:** 1231-1241.
13. Nielsen, R. 2000 *Genetics* **154:** 931-942.
14. Stephens, M. and Donnelly, P. 2000 *J. Roy. Stat. Soc* B **62:** 605-635.
15. Watterson, G. A. 1975 *Theor. Pop. Biol.* **7:** 256-276.
16. Wiuf, C. and Hein, J. 1999 *Theor. Pop. Biol.* **55:** 248-259.

RÉSUMÉ

Composite likelihood methods are used in population genetics in models where full likelihood approaches are computationally intractable. The composite likelihood methods often have desirable statistical properties (such as consistency) and can be used for the estimation of mutation parameters and demographic parameters. We discuss some examples of composite likelihood methods that can be used for the analysis of Single Nucleotide Polymorphism (SNP) data and discuss their statistical properties.