BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 4 2009, pages 552–554
doi:10.1093/bioinformatics/btn665

Genetics and population analysis

Correcting for ascertainment bias in the inference of population structure

Gilles Guillot[1,2,3,*], Matthieu Foll[4,5]

[1]Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo, P.O. Box 1066, Blindern 0316, Oslo Norway, [2]Gothenburg Stochastic Centre, Chalmers University of Technology, Gothenburg, Sweden, [3]Department of Applied Mathematics and Computer Science, INRA/Agro-ParisTech, Paris, France, [4]Computational and molecular population genetics lab, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland and [5]Swiss Institute of Bioinformatics, Berne, Switzerland

Received on September 19, 2008; revised and accepted on December 29, 2008

Advance Access publication January 9, 2009

Associate Editor: Martin Bishop

ABSTRACT

Background: The ascertainment process of molecular markers amounts to disregard loci carrying alleles with low frequencies. This can result in strong biases in inferences under population genetics models if not properly taken into account by the inference algorithm. Attempting to model this censoring process in view of making inference of population structure (i.e. identifying clusters of individuals) brings up challenging numerical difficulties.

Method: These difficulties are related to the presence of intractable normalizing constants in Metropolis–Hastings acceptance ratios. This can be solved via an Markov chain Monte Carlo (MCMC) algorithm known as single variable exchange algorithm (SVEA).

Result: We show how this general solution can be implemented for a class of clustering models of broad interest in population genetics that includes the models underlying the computer programs STRUCTURE, GENELAND and GESTE. We also implement the method proposed for a simple example and show that it allows us to reduce the bias substantially.

Availability: Further details and a computer program implementing the method are available from http://folk.uio.no/gillesg/AscB/

Contact: gilles.guillot@bio.uio.no

1 BACKGROUND

The ascertainment process of molecular markers is most of the time performed on a discovery panel of limited size. As a result, some polymorphic loci do not display any variability on this particular panel and are hence disregarded in any subsequent data analysis. Inferences carried out on such censored subset of loci are therefore subject to a bias known as ascertainment bias. Its qualitative effect and magnitude depend on the ascertainment strategy and on the particular model considered, (Clark *et al.*, 2005; Nielsen and Signorovitch, 2003; Rosenblum and Novembre, 2007; Wakeley *et al.*, 2001). To avoid this bias, it is necessary to inject information in the inference algorithm about the way data have been censored. This amounts to specifying a model of the ascertainment process.

In the case of Bayesian clustering models of population structure, there is a rather natural choice for the prior and the likelihood for the non-censored data. Working on censored data can be accounted for by working with an appropriately modified prior and likelihood. In order to make the presentation more explicit, we base our presentation on the multinomial-Dirichlet model with correlated allele frequencies first introduced by Balding and Nichols (1995) and reworked by Falush *et al.* (2003) and Guillot (2008) among others. This model is a good working example since (i) its generality makes it a versatile model in population genetics, (ii) it has been recently shown to be highly sensitive to the ascertainment bias (Foll *et al.*, 2008), (iii) earlier solutions proposed suffered from a bias (Holsinger *et al.*, 2002; Nicholson *et al.*, 2002), (iv) the problems encountered and the solution proposed here are common to the models underlying the programs STRUCTURE (Falush *et al.*, 2003), GENELAND (Guillot *et al.*, 2005) and GESTE (Foll and Gaggiotti, 2006). Note that the ascertainment bias is currently not taken into account in these programs, although they are increasingly used with markers especially prone to it, such as SNPs and AFLPs.

2 DIFFICULTIES WITH MCMC ALGORITHMS ACCOUNTING FOR ASCERTAINMENT BIAS

The multinomial-Dirichlet model arises in a wide range of demographic models, in particular the infinite-island model (Wright, 1943). In this model, each individual has a known population of origin, and populations exchange genes with a unique and common migrant pool, see Balding (2003) for assumptions and derivations. Each population k receives migrants from the pool at a rate λ_k. Under these assumptions, allele frequencies $\tilde{p_k}$ at each locus follow a Dirichlet distribution with parameters $p\lambda_k$, p representing the allele frequencies in the migrant gene pool. The coefficient $F_{ST\,k}$, defined as $1/\lambda_k - 1$, measures how divergent each local population k is from the metapopulation as a whole. We treat p as unknown and place on it a Dirichlet prior with on top of the hierarchical model a single scalar parameter a.

Assuming a data set n consisting of genotypes at L loci of individuals belonging to K groups, the likelihood is multinomial. Non-zero inbreeding coefficients F_{IS} in the various populations is accounted for at the likelihood level through an extra set of parameters. See Section A of Supplementary Material for details. In practice, population memberships of individuals might be known

*To whom correspondence should be addressed.

or part of the unknown quantities to infer. This task brings up no additional difficulty and we assume hereafter that population memberships of individuals are known.

For easier notation, we write $\phi=(a,F_{ST},F_{IS})$ and $\psi=(p,\tilde{p})$. We denote the prior distributions for ϕ and ψ, and the likelihood for the observed data n in the non-censored case, respectively, by $f(\phi)$, $g(\psi|\phi)$ and $h(n|\phi,\psi)$. Since the Bayesian model is defined in terms of a joint prior probability distribution and a likelihood which are known quantities, they do not bring up any particular difficulty in view of MCMC inference in the non-censored case. Still assuming that the biological/measurement process governing parameters and data has a joint distribution given by $\pi(\phi,\psi,n)=f(\phi)g(\psi|\phi)h(n|\phi,\psi)$ in the non-censored case, but assuming now that n is a dataset arising after censoring, then, the joint distribution of (ϕ,ψ,n) is given by

$$\pi_c(\phi,\psi,n)=f(\phi)\,g(\psi|\phi)\,h(n|\phi,\psi)I_{n\in A}/K_\phi \tag{1}$$

where $I_{n\in A}$ is the indicator function corresponding to the censoring and $K_\phi=\int g(\psi|\phi)h(n|\phi,\psi)I_{n\in A}d\psi\,dn$. See Section B.1 of Supplementary Material for a justification.

The posterior distribution $\pi_c(\phi,\psi|n)$ being proportional to $\pi_c(\phi,\psi,n)$, it involves the unknown normalizing factor K_ϕ. Let us assume that one wants to make a move from (ϕ,ψ) to (ϕ^*,ψ^*) in a Metropolis–Hastings (MH) scheme with proposal distribution $q((\phi^*,\psi^*)|(\phi,\psi))$. The MH ratio is

$$R_{(\phi,\psi),(\phi^*,\psi^*)}=\frac{h(n|\phi^*,\psi^*)}{h(n|\phi,\psi)}\frac{K_\phi}{K_{\phi^*}}\frac{f(\phi^*)}{f(\phi)}\frac{g(\psi^*|\phi^*)}{g(\psi|\phi)}\times\frac{q((\phi,\psi)|(\phi^*,\psi^*))}{q((\phi^*,\psi^*)|(\phi,\psi))} \tag{2}$$

The ratio K_ϕ/K_{ϕ^*} consists of two unknown terms, this thwarts the implementation of the standard MH algorithm.

3 SOLUTION

The cause of difficulties is clear: accounting for the ascertainment process (censoring of the dataset) in a hierarchical model brings up an unknown normalizing constant. This does not allow us to implement standard versions of the MH algorithm. This constant shows up because the model mixes parameters that are locus specific (allele frequencies $\tilde{p}$ and p) and parameters that are not (F_{IS} and F_{ST}) with complex relations of conditional dependences. The same unknown normalizing constant (and issue) would therefore show up in all population genetics clustering models based on the correlated allele frequency model.

For a problem originating from Spatial Statistics, Møller *et al.* (2006) proposed an auxiliary variable method to perform MCMC simulations when the likelihood in the target distribution involves a ratio of unknown normalizing constants. This algorithm has been reworked and simplified subsequently by Murray *et al.* (2006). Both algorithms can be adapted to the present context. See also Beaumont (2003) and Andrieu and Robert (2009) for related approaches. For the sake of conciseness, we present only the single variable exchange algorithm (SVEA) due to Murray *et al.* (2006).

A general MCMC algorithm attempts to make moves from the current state (ϕ,ψ) to a new state (ϕ^*,ψ^*). A natural strategy consists in alternating block updates from (ϕ,ψ) to (ϕ,ψ^*) then from (ϕ,ψ) to (ϕ^*,ψ). The moves of first type do not bring any difficulty as the ratio of unknown terms vanishes. We therefore focus on moves from (ϕ,ψ) to (ϕ^*,ψ). For a proposal $q(\phi^*|\phi)$, the MH ratio is:

$$R_{\psi,\psi^*}=\frac{h(n|\phi^*,\psi)}{h(n|\phi,\psi)}\frac{f(\phi^*)}{f(\phi)}\frac{g(\psi|\phi^*)}{g(\psi|\phi)}\frac{K_\phi}{K_{\phi^*}}\frac{q(\phi|\phi^*)}{q(\phi^*|\phi)} \tag{3}$$

The implementation of the SVEA adapted to the multinomial-Dirichlet model with correlated allele frequencies consists in substituting an importance sampling estimate of K_ϕ/K_{ϕ^*} with its unknown value in R_{ψ,ψ^*} in a way that preserves the invariant distribution. In the present case, this estimate of K_ϕ/K_{ϕ^*} is $g(\nu|\phi)h(m|\phi,\nu)/g(\nu|\phi^*)h(m|\phi^*,\nu)$ where (ν,m) is a pair of auxiliary variables sampled jointly from the distribution of $(\psi,n|\phi^*)$ under censoring. In particular, ν is sampled over the same space as ψ and m over the same space as n. Intuitively, the integration constant involves enumerating the space for which $I_{n\in A}$ is 1, then the MCMC needs to sample from this, and that is essentially what is happening with ν and m. See Section C of Supplementary material for details.

The SVEA update of ϕ in an MCMC algorithm consists in the steps given in Table 1.

Table 1. Steps in SVEA updates of ϕ

(a)	Propose ϕ^* from an arbitrary proposal distribution $q(\phi^*\|\phi)$.
(b)	Propose (ν,m) from distribution $g(\nu\|\phi)h(m\|\phi,\nu)I_{m\in A}/K_\phi$.
(c)	Accept ϕ^* with probability $\min(1,R')$ with $R'=\frac{h(n\|\phi^*,\psi)}{h(n\|\phi,\psi)}\frac{f(\phi^*)}{f(\phi)}\frac{g(\psi\|\phi^*)}{g(\psi\|\phi)}\frac{g(\nu\|\phi)}{g(\nu\|\phi^*)}\frac{h(m\|\nu,\phi)}{h(m\|\nu,\phi^*)}\frac{q(\phi\|\phi^*)}{q(\phi^*\|\phi)}$.

4 RESULTS IN A TOY EXAMPLE

We have implemented the algorithm proposed for a simplified version of the model described above. We simulated genotypes at $L=20$ bi-allelic loci for $N=60(30+30)$ individuals belonging to $K=2$ populations at Hardy–Weinberg equilibrium ($F_{IS}=0$). We assumed a Beta(2,20) distribution for F_{ST} and a Beta(1/2,1/2) distribution for the frequencies p in the ancestral population. Each dataset was simulated so as to mimic an ascertainment strategy where all loci with minor allele frequency lower than 5% were discarded. We simulated independently 1000 such datasets. For each dataset, we carried out inference for $(F_{ST},p,\tilde{p})$ with a standard MH algorithm based on the uncorrected prior-likelihood model π. Then we carried out inference with the prior-likelihood model accounting for censoring (namely π_c), which is made possible through a combination of MH and SVEA type steps.

Inferences of F_{ST}s by the MH algorithm that does not account for ascertainment bias are subject to a bias that amount to 10% of the average value of F_{ST}. With the inference method proposed, this bias drops down to 2%.

5 DISCUSSION

The message of this note is twofold: accounting for the ascertainment bias in MCMC inferences of population structure brings up numerical difficulties and these difficulties can be bypassed by suitable adaptation of recent methods in computational statistics.

We have tried to keep the description of the problem at a general level so that the idea can be adapted to other contexts, in

particular to more complex ascertainment strategies. For example, our algorithm requires to be able to draw random samples from an auxiliary variables whose distribution depends on the ascertainment strategy. This may be challenging in case of complex ascertainment strategies (e.g. SNPs discovered in sister species). If simulating data according to the ascertainment strategy is not possible, a recent extension proposed by Atchade *et al.* (2008) can be used. With this approach, it is not necessary to be able to draw samples from the exact distribution under the ascertainment strategy. However, having a statistical model of the ascertainment strategy remains a requirement. Therefore, it is still problematic to account for the ascertainment bias if poor records were kept of the strategies of discovery and we insist on the need to keep an accurate record of the various strategies used.

To conclude, let us recall again that (i) accounting for ascertainment bias in MCMC inferences would bring-up the difficulty described here for all hierarchical Bayesian models based on the correlated allele frequencies model, and presumably many other kinds of models, (ii) all steps involved in our solution are straightforward as long as the prior distributions and the likelihood are known in the non-censored case, (iii) the proposed algorithm is general and allows one to make MCMC inferences that account for the ascertainment bias in a wide class of population genetics models that include the models underlying the softwares STRUCTURE, GENELAND and GESTE.

ACKNOWLEDGEMENT

This work benefited from comments of Nicola Barson, Oscar Gaggiotti, John Novembre and Arnaud le Rouzic.

Funding: ANR (grant No NT05-4-42230 to G.G.); Swiss NSF (grant No 3100A0-112072 to M.F.).

Conflict of Interest: none declared.

REFERENCES

Andrieu,C. and Roberts,G. (2009) The pseudo-margnal approach for efficient Monte-Carlo computations. *Ann. Stat.* (in press).

Atchade,Y. *et al.* (2008) Bayesian computation for statistical models with intractable normalizing constants. *Research report, arXiv:0804.3152v1*.

Balding,D. (2003) Likelihood-based inference for genetic correlation coefficients. *Theor. Popul. Biol.*, **63**, 221–230.

Balding,D. and Nichols,R. (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

Beaumont,M. (2003) The estimation of growth or decline in genetically monitored populations. *Genetics*, **164**, 1139–1160.

Clark,A. *et al.* (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.*, **15**, 1496–1502.

Falush,D. *et al.* (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Foll,M. and Gaggiotti,O. (2006) Identifying the environmental factors that determine the genetic structure of populations. *Genetics*, **174**, 875–891.

Foll,M. *et al.* (2008) An approximate Bayesian computation approach to overcome biases that arise when using AFLP markers to study population structure. *Genetics*, **179**, 927–939.

Guillot,G. (2008) Inference of structure in subdivided populations at low levels of genetic differentiation.The correlated allele frequencies model revisited. *Bionformatics*, **24**, 2222–2228.

Guillot,G. *et al.* (2005) A spatial statistical model for landscape genetics. *Genetics*, **170**, 1261–1280.

Holsinger,K. *et al.* (2002) A Bayesian approach to inferring population structure from dominant markers. *Mol. Ecol.*, **11**, 1157–1164.

Møller,J. *et al.* (2006) An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, **93**, 451–458.

Murray,I. *et al.* (2006) MCMC for doubly-intractable distributions. In *Proceedings of the 22nd annual conference on uncertainty in artificial inteligence*, Cambridge, MA, USA.

Nicholson,G. *et al.* (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Stat. Soc. B*, **64**, 695–715.

Nielsen,R. and Signorovitch,J. (2003) Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage quilibrium. *Theor. Popul. Biol.*, **63**, 245–255.

Rosenblum,E. and Novembre,J. (2007) Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. *J. Hered.*, **98**, 331–336.

Wakeley,J. *et al.* (2001) The discovery of single-nucleotide polymorphisms and inferences about human demographic history. *Am. J. Hum. Genet.*, **69**, 1332–1347.

Wright,S. (1943) Isolation by distance. *Genetics*, **28**, 114–138.