

An ABC/quasi-likelihood approach for linkage/GWAS study of a Sardinian genetic isolate.

Un approccio ABC/quasi-verosimiglianza per l'analisi di linkage e associazione in un isolato genetico della Sardegna

Stefano Cabras, María Eugenia Castellanos, Erlis Ruli, Mario Pirastu, Maria Pina Concas

Abstract In this work we discuss an application of recent developments of Approximate Bayesian Computation (ABC) methods coupled with the quasi-likelihood theory, to the problem of estimating the relation between genotype and phenotype in genetic isolates with known genealogy between subjects from the common founders. The combination of quasi-likelihoods and ABC allows automatic ABC inference when the likelihood function is intractable. Indeed, the quasi-likelihood delivers an approximation of the intractable likelihood that can enter either into an ABC-MCMC algorithm as a proposal density or can be used as a surrogate of the intractable likelihood function directly in the Bayes theorem. The proposed method is applied to a Genome Wide Association Study of the reduced Mean Cell Volume ($MCV < 72$) on a Sardinian genetic isolate for which the genealogy relating all the observed subjects is known.

Abstract *Questo lavoro illustra l'applicazione di tecniche recenti del tipo ABC assieme alle quasi-verosimiglianze nel problema di stima delle varianti genetiche relazionate con un fenotipo in popolazioni chiuse, in cui è noto l'albero genealogico. La combinazione delle tecniche ABC con le quasi-verosimiglianze consente di trattare problemi di inferenza bayesiana per verosimiglianze non disponibili analiti-*

Stefano Cabras

University of Cagliari (Italy), Department of Mathematics and Informatics, e-mail: s.cabras@unica.it and University Carlos III de Madrid (Spain), Department of Statistics, e-mail: stefano.cabras@uc3m.es

María Eugenia Castellanos

Rey Juan Carlos University (Spain), Department of Statistics and O.R. e-mail: maria.castellanos@urjc.es

Erlis Ruli

University of Padova (Italy), Department of Statistics e-mail: ruli@stat.unipd.it

Mario Pirastu

CNR (Italy), Institute of Genetic Populations e-mail: m.pirastu@igp.cnr.it

Maria Pina Concas

CNR (Italy), Institute of Genetic Populations e-mail: m.p.concas@igp.cnr.it

camente. Infatti la quasi-verosimiglianza può essere utilizzata come distribuzione proponente in un algoritmo ABC-MCMC oppure come surrogato della verosimiglianza intrattabile direttamente nel teorema di Bayes. Queste tecniche si applicano alla stima delle varianti genetiche relazionate con il diametro dei globuli rossi nel sangue, su un campione estratto da un isolato genetico della Sardegna per il quale è noto l'albero genealogico.

Key words: Estimatin functions, genealogy trees, likelihood-free inference, objective Bayesian inference, quasi-likelihoods.

1 Introduction: the Sardinian genetic isolate dataset and its potentials

In this paper we address the problem of estimating DNA markers related to a phenotype, e.g. the presence of a certain disease. This problem is dealt with a large sets of DNA sequences, referred to as Single-Nucleotide Polymorphisms (SNPs). Such SNPs are usually observed in millions per individuals and thus fast and reliable statistical methods are needed in order to answer the scientific question enquiring which SNPs are mostly related with the disease. Usually, the data are collected on open populations where the degree of inbreeding is unknown, and is typically negligible. Nonetheless, there is some evidence in the genetic literature that the study of isolate genetics, i.e. human samples with known and relevant inbreeding, provides valuable information on SNPs/disease relation. Such type of samples are very rare and so are the statistical methods to analyse them (Cabras et al., 2011). An example of such data is the Sardinian genetic isolate of Ogliastro region situated in the center of the Sardinian island (Italy).

The presence of highly dependent observations complicates the statistical model. In fact, the probabilistic model for the sample that relates the phenotype to the genotype, must take into account also the genealogy underlying the genetic variant transmission when the SNP related with the phenotype is unknown. Since the genotype is observed for the very last generations only, then the configurations of the SNPs for the previous generations constitute a huge number of random unobservable latent variables. This makes almost impossible to write the likelihood function for the parameters relating the configuration of the SNPs with the phenotype.

The scarcity and the difficulty of analysing such type of data for human diseases might explain why the literature on linkage analysis and Genome Wide Association Studies developed on separate grounds. Here we try to embed them in a common probabilistic framework. This is possible thanks to recent developments on Approximate Bayesian Computation (ABC) that allows to draw posterior inference, on the parameters of interest, even when the likelihood function is not available. This rapidly growing ABC literature leads to a set of methods which do not involve direct calculation of the likelihood, delivering thus Bayesian inference which is approximate in a sense that will be specified later. ABC methods are becoming popular

in genetics (Siegmund et al., 2008; Foll et al., 2008), epidemiology (Blum and Tran, 2010) and in population biology (Ratmann et al., 2007) among others.

We found that ABC can be further improved by using quasi-likelihood functions. The latter is of interest in problems where, for a parametric model, the corresponding likelihood function is not available in a closed form expression and therefore cannot be treated with the usual Bayesian techniques.

As an application, we consider a sample from the Talana village people that is affected by reduced MCV for which it is known that there exists a genetic variant inside the Beta-Globin gene that determines a $MCV < 72$. There are 103 individuals with 96 SNPs. Among these, there is only one variant inside the Beta-Globin gene that we expect to be estimated as the most associated variant with the reduced MCV. An extract of the available sample is represented in Figure 1. We can see that the population is mostly isolated and few founders contributed to genetic variation of the whole sample. Usually, such founders have not been observed, neither for their phenotype and genotype.

2 The Stochastic model and the intractable likelihood

In this analysis we treat each SNP separately and set up a stochastic model for a single SNP. The overall analysis is made by separate analyses for all SNPs. In particular, for individual i let $Y_i \in \{0, 1\}$ represent the indicator of the phenotype, i.e. $Y_i = 1$ if affected and $Y_i = 0$ otherwise and let $X_i \in \{\{aA, Aa\}, aa, AA\}$ be the genotype, e.g. the SNP configuration with three levels. For a genealogy made of N individuals we shall model the corresponding pairs $(Y_1, X_1), \dots, (Y_N, X_N)$. At the phenotype level we assume the usual logistic regression model $(Y_i | X_i, \theta) \sim \text{Bernoulli}(p_i)$, where the linear predictor is

$$\text{logit}(p_i) = \theta_1 \mathbb{I}_{(X_i = \{aA, Aa\})} + \theta_2 \mathbb{I}_{(X_i = aa)} + \theta_3 \mathbb{I}_{(X_i = AA)}.$$

The parameters $\theta = (\theta_1, \theta_2, \theta_3)$ are usually interpreted as the log of the odds ratio for the probabilities of being affected given the SNP configuration and given the affection status of the parents. Concretely, let X_{i_1} and X_{i_2} be the SNP configuration for the ancestors of individual i . Then the probabilistic model for the transmission of a genotype variant is assumed to be governed by the usual Mendelian inheritance law

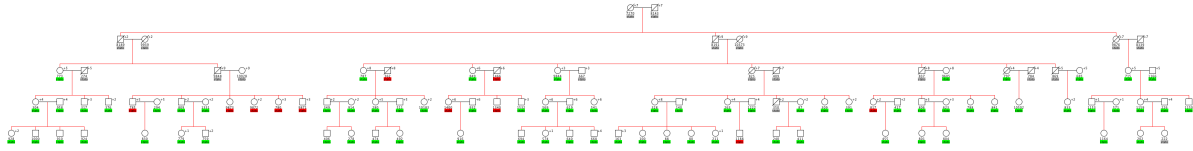


Fig. 1 Extract of the Genealogy tree from the Talana village. Ancestors (white) are not observed; while other individuals are labelled as healthy (green) or affected (red).

of transmission, where the ancestor individuals are assumed to be known according to the genealogy tree (see Figure 1). Therefore, if individual i is a descendent in the genealogy tree, then

$$(X_i | X_{i_1}, X_{i_2}) \sim \text{Mendel's law},$$

while if i is a founder and his/her ancestors are not in the genealogy tree, then

$$X_i \sim \text{Trinomial}(1/3, 1/3, 1/3).$$

As shown in Figure 1, not all configurations have been observed and thus the likelihood for $\theta \in \mathbb{R}^3$, $L_N(\theta)$, given the observed data $\mathcal{D} = (y_i, x_i)_{i=1}^n$, $n < N$ cannot be written in closed form as one must integrate over the SNP configurations of ancestors and founders. Nonetheless, we wish to draw posterior inference on $\pi_N(\theta | \mathcal{D}) \propto L_N(\theta) \pi(\theta)$, based on the improper uninformative prior $\pi(\theta) \propto 1$. In order to approximate $\pi_N(\theta | \mathcal{D})$ we have to resort to ABC methods. We further gain precision by the approximation of $L_N(\theta)$ with the quasi-likelihood function, $L_Q(\theta)$. Hence, we propose to combine ABC and quasi-likelihood. To accomplish this, we need to fix a relation between the observations, through a vector of summary statistics $\mathbf{s}$, and the parameters θ . This is explained in the next section.

3 Bayesian estimation with ABC and quasi-likelihood

The use of ABC for the above model requires a choice of a, not necessarily but hopefully sufficient, summary statistic $\mathbf{s} = (s_1, s_2, s_3) \in \mathbb{R}^3$. In this case we choose s_1 as the observed log-odds for the frequency of $Y = 1 | X = aa$ against $Y = 0 | X = aa$, while s_2 and s_3 are the log odds for $X = \{aA, Aa\}$ and $X = AA$, respectively. Such log-odds are always calculated for the observed individuals in the genealogy tree. In its basic version, ABC consists in sampling parameters from the prior, then generating summary statistics and retaining those generated values of the parameters which lead to summary statistics similar to the observed ones. Such retained values are supposed to be an approximation of the posterior distribution of the parameters.

For the distance $\rho(\mathbf{s}, \mathbf{s}_{obs})$ between $\mathbf{s}$ and its observed value $\mathbf{s}_{obs}$, required in ABC, we chose the Euclidean norm weighted by $1 - \#\{x_{obs} = x_{sim}\}/n$. This weight is the proportion of simulated SNPs in the tree that differs from the observed configurations. The tolerance parameter $\varepsilon > 0$ has been fixed in order to obtain an acceptable jumping probability in the ABC-MCMC algorithm (Marjoram et al., 2003), needed because $\pi(\theta)$ is improper.

The theory and the use of estimating equations and that of the related quasi- and quasi-profile likelihood functions has received a lot of attention in recent years; (see, among others, Barndorff-Nielsen (1995); Adimari and Ventura (2002)). In addition, Ventura et al. (2010); Lin (2006); Greco et al. (2008) discuss the use of quasi-likelihood functions in the Bayesian setting.

In our proposed method the quasi-likelihood enters as a proposal density for the ABC-MCMC algorithm through a reparametrization of θ , namely $f(\theta) = E_{\pi(y|\theta)}(\mathbf{s} |$

$\theta) \in \mathbb{R}^3$ which is informative for the posterior mean of θ for a given choice of $\mathbf{s}$. The function $f(\theta)$ cannot generally be elicited and an estimated version $\hat{f}$ is obtained using a pilot-run simulation. As the pilot-run depends only on the genealogy tree and not on the specific observed sample, it can be used routinely for analysing each SNP separately. Once $\hat{f}$ is obtained, the quasi-likelihood proposal is given by

$$L_Q(\theta) = \Sigma_R^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (\hat{f}(\theta) - \mathbf{s}_{obs})^\top \Sigma_R^{-1} (\hat{f}(\theta) - \mathbf{s}_{obs}) \right), \quad (1)$$

where Σ_R is the estimated conditional covariance matrix for the regression function $\mathbf{S} \mid \theta$ the acceptance probability for the ABC-MCMC becomes

$$\max \left\{ 1, \frac{\pi(\theta^*) L_Q(\theta^{(t-1)}) \mid J(\theta^{(t-1)}) \mid \mathbb{I}_{\rho < \varepsilon}}{\pi(\theta^{(t-1)}) L_Q(\theta^*) \mid J(\theta^*) \mid} \right\}.$$

The obtained simulation output consist of the approximate posterior distributions of θ , one for each SNP. We summarize them marginally by looking at the log of posterior odds for the hypothesis that a SNP is at risk against protective factor for the reduced MCV, that is the approximation of the log of $\Pr(\theta > 0 \mid \mathcal{D}) / (1 - \Pr(\theta > 0 \mid \mathcal{D}))$ which exists as long as the posterior $\pi(\theta \mid \mathcal{D})$ exists. From the obtained weights of evidence we have that the variant inside the Beta-Globin gene exhibits one of the largest weights of evidence. This holds for all the SNPs analysed.

The same analysis has been performed by replacing $L_N(\theta)$ with $L_Q(\theta)$ into the Bayes theorem. We obtained almost the same results, but with an extremely faster computational algorithm that is feasible for very large GWAS analysis. In fact, such algorithm requires to simulate from the multivariate normal in (1) and calculate the inverse of $\hat{f}$. In order to speed-up the calculations, this inverse function can be also interpolated by using the output of the pilot-run.

References

- Adimari, G. and L. Ventura (2002). Quasi-profile log likelihoods for unbiased estimating functions. *Ann. Inst. Stat. Math.* 54, 235–244.
- Barndorff-Nielsen, O. (1995). Quasi profile and directed likelihoods from estimating functions. *Ann. Inst. Stat. Math.* 47, 461–464.
- Blum, M. G. B. and V. C. Tran (2010). Hiv with contact tracing: a case study in approximate bayesian computation. *Biostat.* 11, 644–660.
- Cabras, S., M. E. Castellanos, G. Biino, I. Persico, A. Sassu, L. Casula, S. del Giacco, F. Bertolino, M. Pirastu, and N. Pirastu (2011). A strategy analysis for genetic association studies with known inbreeding. *BMC Genet.* 12:63.
- Foll, M., M. A. Beaumont, and O. Gaggiotti (2008). An approximate bayesian computation approach to overcome biases that arise when using amplified fragment length polymorphism markers to study population structure. *Genet.* 179, 927–939.

- Greco, L., W. Racugno, and L. Ventura (2008). Robust likelihood functions in bayesian inference. *J. Stat. Plann. Inf.* 138, 1258–1270.
- Lin, L. (2006). Quasi bayesian likelihood. *Stat. Methodol.* 3, 444–455.
- Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003). Markov chain monte carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* 100, 15324–8.
- Ratmann, O., O. Jørgensen, T. Hinkley, M. Stumpf, S. Richardson, and C. Wiuf (2007). Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *h. pylori* and *p. falciparum*. *PLoS Comput. Biol.* 3, 2266–2276.
- Siegmund, K. D., P. Marjoram, and D. Shibata (2008). Modeling DNA methylation in a population of cancer cells. *Stat. Appl. Genet. Mol. Biol.* 7, 1–21.
- Ventura, L., S. Cabras, and W. Racugno (2010). Default prior distributions from quasi- and quasi-profile likelihoods. *J. Stat. Plann. Inf.* 140, 2937–2942.