Journal of Theoretical Biology 459 (2018) 119–129

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtb

Allelic frequency estimation in presence of uncertain priors

Ali Karimnezhad[a,b,*], Fahimeh Moradi[b]

[a] Ottawa Hospital Research Institute, Ottawa, ON, Canada
[b] Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

ARTICLE INFO

Article history:
Received 31 October 2017
Revised 19 August 2018
Accepted 24 September 2018
Available online 25 September 2018

Keywords:
Admissibility
Allelic frequency
Bayesian analysis
Minimaxity
Robust Bayesian analysis

ABSTRACT

In this paper, we assume that allele frequencies are random variables and follow certain statistical distributions. However, specifying an appropriate informative prior distribution with specific hyperparameters seems to be a major issue. Assuming that prior information varies over some classes of priors, we develop the concept of robust Bayes estimation into the context of allele frequency estimation. We first assume that the region of interest is a single locus and the prior information is represented in terms of a class of Beta distributions, and present explicit forms of the resulting Bayes and robust Bayes estimators. We then extend our results to biallelic k-loci and multi-allelic k-loci cases within the region of interest. We perform a simulation study to measure performance of the proposed robust Bayes estimators against some Bayes estimators associated with specific hyperparameters. The simulations reflect satisfactory performance of the proposed robust Bayes estimators when there is no evidence implying the actual prior distribution.

1. Introduction

Allelic frequencies are one of the basic terms in several areas of population genetics and bioinformatics including linkage and association analysis, calculation of linkage disequilibrium, admixture mapping and somatic point mutation detection. In fact, many genetic epidemiological analyses are quite sensitive to estimates of allele frequencies (Lockwood et al., 2001). Mandal et al. (2006) emphasize that before one is able to apply model-dependent linkage analysis appropriately, allele frequencies have to be known. As well, they refer to some outstanding studies in the literature discussing the effects of using wrong allele frequencies. According to Lockwood et al. (2001), under-estimation of allele frequencies can lead to false linkage, whereas over-estimation can lead to reduced power.

Allele frequencies can also be different among populations within a given geographical region. Weir and Hill (2002) remark that even if two populations are maintained under the same evolutionary conditions, the corresponding allele frequencies will be different due to the stochastic nature of the conditions. Lockwood et al. (2001) report that when studies involve multiple populations with different evolutionary histories, it is difficult to obtain accurate estimates of allele frequencies. It is of interest to point out that, a lot of effort has been devoted to study the changes in allelic frequencies from one population to another, known as genetic drift. For example, Corander et al. (2003) discuss that genetic drift results in divergence of gene frequencies between populations of a common origin when migration and mutation rates are low. A huge number of studies in the literature deal with the development of statistical methods for estimation of the degree of population differentiation and the related topic of genetic population structure. Readers may refer to Bhatia et al. (2013), Holsinger (1999), Holsinger and Weir (2009) and Leinonen et al. (2013), among many others.

The literature is abundant of many works that have treated the allelic frequencies using the maximum likelihood (ML) approach (e.g. Adrianto and Montgomery, 2012; Boehnke, 1991; Holsinger, 1999; Lange, 1995; 2003, among many others). Although the ML estimators have often good properties, the ML procedure treats the underlying parameter of interest as an unknown and fixed parameter. In contrast to the ML approach, quite a number of studies in the literature follow the Bayesian strategy. Referring back to Crow and Kimura (1970) and Wright (1931, 1937), Martínez et al. (2015) point out that allelic frequencies under certain scenarios have random variation, and assume that allelic frequencies follow a Beta distribution with possibility of a genetic interpretation. As well, many others including Holsinger (1999), Lange (1995, 2003) and Lockwood et al. (2001) assume that allelic frequencies follow the Beta prior. Lange (1995, 2003) remarks that the primary drawback of being Bayesian in the allelic fre-

[*] Corresponding author at: Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada..
E-mail address: a.karimnezhad@uottawa.ca (A. Karimnezhad).

https://doi.org/10.1016/j.jtbi.2018.09.029
0022-5193/© 2018 Elsevier Ltd. All rights reserved.

quency estimation problem is that there is no obvious way of selecting reasonable hyperparameters. He overcomes this problem by estimating the hyperparameters using the Newton's approximation method. Holsinger (1999) considers the Beta prior, as well. He also remarks that if there is no prior information about allele frequencies, a uniform distribution might be chosen, suggesting that every possible allele frequency is equally likely. Lockwood et al. (2001) propose a Bayesian hierarchical model that allows for explicit inclusion of prior information about both allele frequencies and inter-population divergence. Thus, it is necessary to treat the allelic frequencies as random variables and decision theory, especially the Bayesian decision theory, should be followed in order to derive optimal estimates of the allelic frequencies.

As reviewed above, a major criticism in the Bayesian decision theory is the uncertainty in prior elicitation. It is usually unknown if a given prior $\pi(.)$ is best and yields a reliable solution. That would be ideal if one could specify a prior from relevant sources of information, but such sources are not always present. In such situations, Bayesians may refer to some objective (or nonsubjective) priors. There are quite a number of valuable references reviewing several criteria for the construction of objective priors. Berger (2013) describes different methods of prior elicitation and extensively discusses the determination of non-informative priors, maximum entropy priors and right (or left) invariant Haar priors. As well, he argues a number of criticisms concerning non-informative priors and refers to Jeffreys (1961) prior as the most popular non-informative prior. Ghosh (2011) reviews some specific criteria for the selection of priors, and provides a "tools box" containing many objective priors including Bernardo's reference prior (Bernardo, 1979), Jeffreys' prior (Jeffreys, 1961), probability matching priors, etc. See also Irony and Singpurwalla (1997) for an overview of the conceptual aspects of using non-informative priors. According to Berger (2013), perhaps a negative feature of objective priors is that they are often numerous. For example, in the problem of estimating a binomial parameter θ, Berger (2013) refers to four non-informative priors, namely, $\pi_1(\theta) = 1$, $\pi_2(\theta) = [\theta(1-\theta)]^{-1}$, $\pi_3(\theta) \propto [\theta(1-\theta)]^{-1/2}$, $\pi_4(\theta) \propto \theta^{\theta}(1-\theta)^{(1-\theta)}$. However, using objective priors overcomes the problem of defining a prior distribution arbitrarily.

The uncertainty problem might also happen when two or more statisticians agree on a prior distribution but there is a significant difference between their chosen hyperparameters. To exemplify, suppose a random variable X follows a *Bernoulli*(θ)-distribution in which θ represents the allele frequency. A biologist might assign *Beta*$(2, 2)$-prior distribution for the desired parameter θ while another biologist might assert that *Beta*$(2, 4)$-prior would suit the data much better. In this situation, we have to give credit to both priors to make a reliable decision. To overcome such an uncertainty in prior elicitation, robust Bayesian methodology has been introduced in the literature. It solves the problem by minimizing some functionals giving credit to the underlying prior varying in a prespecified class of priors, say Γ. In fact, robust rules are aimed at global prevention against bad choices of prior or hyperparameters (Karimnezhad et al., 2017; Karimnezhad and Parsian, 2014; 2018). For more discussion, refer to Arias-Nicolás et al. (2009), Berger (1990, 2013), Berger et al. (1994) and Insua et al. (1992), among many others.

In this paper, we address the use of Bayesian inference in the allele frequency estimation problem when there is an uncertainty in choosing a prior distribution. To provide a solution, we follow the robust Bayesian methodology and derive some PRGM rules. Section 2 provides some basic materials. Section 3 is the main focus of our paper in which different Bayes and robust Bayes rules for estimating an allele frequency in a specific locus are presented. In Section 4, we extend our results to k-loci and multi-allelic cases. A simulation study along with a comparison of performance of the Bayes and robust Bayes estimators is presented in Section 5. Finally, discussion and concluding remarks are provided in Section 6.

2. Materials and methods

In this paper, we follow the Hardy–Weinberg equilibrium principle at every locus. Assuming that the capital letter "B" stands for a "reference allele", suppose X_1, X_2 and X_3 are random variables indicating the number of individuals with the genotypes AA, AB and BB. Denoting the frequency of the reference allele B by θ, where $\theta \in [0, 1]$, the Hardy–Weinberg equilibrium principle leads to the assumption that $\boldsymbol{X} = (X_1, X_2, X_3)^T$ conditional on θ follows a trinomial distribution with the frequencies $(1-\theta)^2$, $2\theta(1-\theta)$ and θ^2. The purpose here is to estimate the frequency θ by some decision rule $\delta = \delta(\boldsymbol{X}) \in \mathcal{D}$, where $\mathcal{D}$ refers to the class of all decision rules. It is necessary to remark that we distinguish between an estimator (or a decision rule) and an estimate. An "estimator" is a rule $\delta(\boldsymbol{X})$ which is based on the random variable $\boldsymbol{X}$, and an "estimate" is a value $\delta(\boldsymbol{x})$ which is based on a realization of $\boldsymbol{X}$, i.e., $\boldsymbol{x}$.

Perhaps the most popular method in estimating a desired parameter is the ML approach in which an estimator is derived by maximizing the likelihood function of a given sample over the parameter space.

Following the Hardy–Weinberg equilibrium principle, the likelihood function based on the observation $\boldsymbol{x}$ is given by

$$L(\theta|\boldsymbol{x}) = \frac{(2n)!}{(2x_1+x_2)!(x_2+2x_3)!}\theta^{x_2+2x_3}(1-\theta)^{2x_1+x_2}, \quad 0 \le \theta \le 1,$$

provided that $x_1 + x_2 + x_3 = n$. It is easy to verify that the ML estimator of the allelic frequency θ is given by $\delta_{ML}(\boldsymbol{X}) = \frac{X_2+2X_3}{2n} = 1 - \frac{X_2+2X_1}{2n}$.

Notice that, as Martínez et al. (2015) point out, estimating the allelic frequency θ can be performed by only counting the two alleles A and B, instead of counting the three genotypes AA, AB and BB, in a given sample. To do so, simply define $\boldsymbol{Y} = (Y_1, Y_2)^T$, where $Y_1 = X_2 + 2X_3$ and $Y_2 = X_2 + 2X_1$. By this transformation, the random variables Y_1 and Y_2 are the number of B and A alleles, respectively. The likelihood function on the basis of the observation $\boldsymbol{y}$ is then changed to

$$L(\theta|\boldsymbol{y}) = \frac{(2n)!}{y_1!y_2!}\theta^{y_1}(1-\theta)^{y_2}, \quad 0 \le \theta \le 1,$$

provided that $y_1 + y_2 = 2n$. Therefore, the allelic frequency θ is estimated by $\delta_{ML}(\boldsymbol{Y}) = \frac{Y_1}{2n} = 1 - \frac{Y_2}{2n}$.

Using either the counts $\boldsymbol{x}$ or $\boldsymbol{y}$ leads to the same ML estimates which in fact roots from using the same likelihood functions, see Appendix B. This is true in general and both counts $\boldsymbol{x}$ and $\boldsymbol{y}$ should yield the same ML estimates. Nevertheless, sometimes working with the counts $\boldsymbol{y}$ makes calculations more simple and faster rather than the counts $\boldsymbol{x}$. However, since it is not always possible to make a one-to-one map between $\boldsymbol{y}$ and $\boldsymbol{x}$, we stick to using the counts $\boldsymbol{x}$, unless otherwise stated.

Now, as it is common in the decision theory, let the loss function $\mathcal{L}(\theta, \delta)$ measure penalty of incorrect estimation of the parameter of interest θ by a decision rule (estimator) $\delta \in \mathcal{D}$. Following the Bayesian inference, let the random variable θ have the probability density function (pdf) $\pi(.)$. Once the observation $\boldsymbol{x}$ is available, the prior density $\pi(\theta)$ is updated and replaced by the posterior pdf $\pi(.|\boldsymbol{x})$, and then a Bayes point estimator $\delta_\pi = \delta_\pi(\boldsymbol{X})$ of the parameter of interest θ is derived by minimizing the posterior risk

$$\rho_\delta(\pi, \boldsymbol{X}) = E[\mathcal{L}(\theta, \delta)|\boldsymbol{X}] = \int_\Theta \mathcal{L}(\theta, \delta)\pi(\theta|\boldsymbol{X})d\theta,$$

which in fact averages losses incurred when estimating θ by δ based on the posterior density $\pi(.|\boldsymbol{X})$. For more details see