Statistical models for inferring the structure and history of populations from genetic data

Jukka Sirén

Department of Mathematics and Statistics
Faculty of Science
University of Helsinki

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public examination in auditorium CK112, Exactum (Gustaf Hällströmin katu 2B) on 23rd of April 2012, at 12 o'clock noon.

Helsinki 2012

ISBN 978-952-10-7895-8 (paperback)
ISBN 978-952-10-7896-5 (PDF)
http://ethesis.helsinki.fi
Helsinki University Print
Helsinki 2012

Supervisor
Professor Jukka Corander
Department of Mathematics and Statistics
University of Helsinki
Finland

Pre-examiners
Professor Pekka Pamilo
Department of Biosciences
University of Helsinki
Finland

Professor Ziheng Yang
Department of Genetics, Evolution and Environment
University College London
United Kingdom

Custos
Professor Jukka Corander
Department of Mathematics and Statistics
University of Helsinki
Finland

Opponent
Professor David Balding
UCL Genetics Institute
University College London
United Kingdom

Abstract

Population genetics has enjoyed a long and rich tradition of applying mathematical, computational and statistical methods. The connection between these fields has deepened in the last few decades as advances in genotyping technology have led to an exponential increase in the amount of genetic data allowing fundamental questions involving the nature of genetic variation to be asked. The massive quantities of data have necessitated the development of new mathematical and statistical models along with computational techniques to provide answers to these questions.

In this work we address two problems in population genetics by constructing statistical models and analyzing their performance with simulated and real data. The first one concerns the identification of genetic structure in natural populations from molecular data, which is an important aspect in many fields of applied science, including genetic association mapping and conservation biology. We frame it as a problem of clustering and classification and utilize background information to achieve a higher accuracy, when the genetic data is sparse. We develop a computationally efficient method for taking advantage of geographical sampling locations of the individuals. The method is based on the assumption that the spatial structure of the populations correlates strongly with the genetic structure, which has been proven reasonable for human populations.

In the assignment of individuals into known populations, we also show how improvements in the efficiency of the inference can be obtained by considering all of the individuals jointly. The result is derived in the context of classification, which is major field of study in machine learning and statistics, making it applicable in a wide range of situations outside population genetics.

The other problem involves the reconstruction of evolutionary processes that have resulted in the structure present in current populations. The genetic variation between populations is caused to large extent by genetic drift, which corresponds to random fluctuations in the distribution of a genetic type due to demographic processes. Depending on the genetic marker under study, mutation has only a minor or even negligible role, in contrast with traditional phylogenetic methods, where mutational processes dominate as the time scales are longer. We follow the change in the relative frequencies of different genetic types in populations by deriving approximations to widely used models in population genetics. The direct modeling of population level properties allows the method to be applied data sets harboring thousands of samples, as demonstrated by the analysis of global population structure of *Streptococcus pneumoniae*.

Preface

L.J. Savage wrote in the preface of his 1954 book The Foundations of statistics about his concerns over the impact of the work:

> Again, what he has written is far from perfect, even to his biased eye. He has stopped revising and called the book finished, because one must sooner or later.
>
> Finally, he fears that he himself, and still more such public as he has, will forget that the book is tentative, that an author's most recent word need not to be his last word.

While the objectives of this work are much more modest than those of Savage's book, which laid decision theoretic foundations for Bayesian approach to statistics, my own feelings are closely reflected in the above quotation. The work behind the articles that constitute the main part of the thesis spread over many years and the first one was published already four years ago. Should I be addressing the problem described in article (I) now, the resulting methods would probably be quite different. However, such considerations serve only as a thought experiment as the relevance of applied statistical methods rely on their ability to provide meaningful answers to scientific questions. Whether a possibly better method could be devised is a question for possible future research, but it does not diminish the importance of earlier work.

I would like to thank my supervisor Jukka Corander for giving an opportunity to work with a range of interesting problems. He has provided me a great freedom and support in pursuing the subjects I have found interesting. Many times during this work when I have felt that there are insurmountable barriers for the method under development, Jukka has quickly provided a wide array of ideas and techniques how these can be circumvented.

Two book reading seminars organized by Elja Arjas have helped me to deepen my knowledge of population genetics and understand better the foundations of statistics. In the first seminar on coalescent theory, the interesting discussions by Elja, Anders, Siru, Matti and Jukka K put the theory in context and opened up the many details in a way, which would not have been possible by independent reading of the book. The other seminar on the book Probability theory by E.T. Jaynes made me to study the philosophy and practice of statistics.

This work has not been made in isolation and I would like to thank the collaborators who have not been mentioned earlier. The articles (I-V) greatly benefited from the contributions made by the coauthors Pekka Marttinen, Jing Tang, Yaqiong Cui, Timo Koski and Bill Hanage. I also acknowledge the many interesting discussions on Bayesian statistics and other topics with my fellow PhD students: Paul, Riku, Väinö, Elina, Lu, Alberto, Jie and others.

Finally, I am thankful to my wife Milka and my daughter Emmi. They have helped me to concentrate on the most important aspects of my life, when the work for this thesis has been the most stressful.

This work was made possible by funding from COMBI graduate school and Finnish Doctoral Programme in Computational Sciences FICS. Financial help has also been received from the Finnish Centre of Excellence in Computational Inference Research (COIN).

Contents

List of original articles

I Corander, J., Sirén, J. and Arjas, E. 2008. Bayesian spatial modeling of genetic population structure. Computational Statistics. 23:111–129.

II Corander, J., Marttinen, P., Sirén, J. and Tang, J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics. 9:539.

III Corander, J., Cui, Y., Koski, T. and Sirén, J. 2011. Have I seen you before? Principles of Bayesian predictive classification revisited. Statistics and Computing. Published online Oct 4, 2011. doi:10.1007/s11222-011-9291-7

IV Sirén, J., Marttinen, P. and Corander, J. 2011. Reconstructing Population Histories from Single Nucleotide Polymorphism Data. Molecular Biology and Evolution. 28:673–683.

V Sirén, J., Hanage, W.P. and Corander, J. 2012. Inference on Population Histories by Approximating Infinite Alleles Diffusion. Submitted.

Author's contributions to articles I-V

I The method was jointly developed by all authors. JS was fully responsible for implementing and testing the method. JS took part in writing the article, while JC had the main responsibility.

II All authors contributed equally to the development of methodology and data analysis. JS took part in writing the article, while JC and JT had the main responsibility.

III The methods were jointly developed by all authors. JS implemented the methods and analyzed simulated data. JS took part in writing the article, while JC had the main responsibility.

IV,V JS contributed the main part in all aspects of the articles.

1 Introduction

Statistics and population genetics share a lot of common history. This is perhaps most clearly highlighted in the work of R.A. Fisher (1890-1962), who not only made important contributions to these fields, but revolutionized them almost single-handedly in 1920's and 30's (Edwards, 2003; Green, 2003; Healy, 2003). He introduced and developed many widely used concepts in statistics, such as likelihood, sufficiency, randomization, experimental design and analysis of variance, but these form only a fraction of his contributions to statistics (Fienberg, 1992). The scope of Fisher's work is summarized by Savage (1976), who wrote that "It would be more economical to list the few statistical topics in which he displayed no interest than those in which he did".

The theory and practice of statistics can be divided broadly in to two approaches: frequentist and Bayesian, which differ most substantially in the way in which uncertainty is handled. In the Bayesian approach all uncertainty is described using probability, making it a complete system where inferences are conducted using the laws of probability theory (Bernardo and Smith, 1994; Robert, 2007). In the frequentist statistics, for which Fisher was one of the leading figures in the 20th century, probability is utilized only on some aspects of uncertainty. The main distinction between these two approaches comes from the interpretation of probability and has been a cause of controversy in the philosophy of statistics (Good, 1959; Cox, 1978; Efron, 1986; Lindley, 2000). This debate has been mostly left to the past and it has become evident that successful inferences can be obtained using many different methods (Efron, 2005; Kass, 2011).

The foundations of mathematical population genetics were laid in 1920's and 30's by Fisher, J.B.S. Haldane and Sewall Wright (Ewens, 2004). Their objective was to formulate an evolutionary theory based on Mendelian laws, which were considered to be incompatible with Darwinian evolution in the early 20th century. A reconciliation between these was achieved by Fisher (1918), whose derivations showed that Mendelian segregation results in maintenance of genetic variation.

The modern methods for inference in population genetics are still based on the early theory developed by Fisher, Wright and Haldane, but an important shift in the direction of thinking has taken place. The classical theory was prospective, in the sense that it considers what happens to genetic variability in the future. The modern theory is in contrast retrospective and asks questions about the history of variation. Also, perhaps due to lack of computational techniques, the early derivations considered either limiting behavior with time or population size to get stationary distributions, or variability after a few generations governed by Mendel's laws. The availability of huge quantities of molecular genetic data has created the need to provide answers to questions not considered by Fisher and others, and consequently, to develop the theory to new directions.

An important aspect of many fields of science involving large numbers of quantitative data is that of modeling. For example in physics models are able to describe the processes in nature almost exactly. In population genetics such models are practically impossible to design as the biological processes behind the phenomena are too complex and unpredictable. Nevertheless it possible to come up with simplified models, which capture many of the important aspects of genetic variation and allow conclusions to be made. This is similar to the general practice of statistics, where models are known to be wrong in advance, but they still facilitate inferences about uncertain quantities based on observed data (Gelman and Shalizi, 2012).

This work describes model-based solutions to two problems in population genetics: discovery of population structure and reconstruction of the history of populations. The former, which is the topic of articles (I-III), requests a representation of genetic variation among individuals of the same species. The latter calls for identification of the processes that have lead to this genetic variation and is discussed in articles (IV-V).

This summary part of the thesis reviews the basis of the methods presented in the articles (I-V) and complements the discussion by placing them in a wider context. The actual details of the statistical models are left to the articles and described only when they facilitate deeper understanding of the models. The structure of the summary is as follows. The Bayesian approach to statistics is presented in Section 2. We also highlight some potential inaccuracies and sources of bias arising in model-based inference. In Section 3, computational issues associated with Bayesian methods are discussed and several popular algorithms are described. Partition-based models of articles (I-III) for inferring population structure are introduced in Section 4. They are presented in the broader context of clustering and classification, which are major fields of study in statistics and machine learning. In Section 5, we describe methods to reconstruct the history of several populations in form of a tree, acknowledging connections to phylogenetic and coalescent theories. Section 5 derives also approximations used in articles (IV,V) to models of mathematical population genetics. Finally, limitations and possible extensions of the models are discussed in Section 6.

2 Bayesian approach to statistics

In this section an overview of the Bayesian approach to statistics is given. Perhaps more appropriate would be to state this as a Bayesian approach, because a large number of different methods have been proposed under the label Bayesian (Good, 1971). The focus of this section is on statistical inference: what can we learn about some unknown quantity θ from observed data X. This leaves out many important fields of statistics such as design of experiments (Dean and Voss, 1999).

Throughout the discussion we assume that the potential inference is carried out to gain knowledge about some aspect of reality. If the results of the inferences were to be used for making decisions, then, as noted by Fisher (1955), a larger variety of methods could be considered. Breiman (2001) argues that formal statistical models, which are the topic of this work, may be found too restrictive is such situations. Nevertheless, subjective Bayesian methods provide a unified framework for combining information from multiple sources, and they have been found valuable in many applied decision problems (Goldstein, 2006).

2.1 Bayesian inference

The Bayesian approach to statistics is characterized by the usage of probability to describe all uncertainty (Bernardo and Smith, 1994; Jaynes, 2003; Gelman et al., 2004; Robert, 2007). The name Bayesian comes from the use of Bayes' theorem to update the probability of hypothesis H after observing data X

$$p(H \mid X) = \frac{p(X \mid H)p(H)}{p(X)}. \tag{2.1}$$

Here $p(H)$ is the probability of H before observing X, $p(X \mid H)$ is the probability of observing X given that H is true and

$$p(X) = p(X \mid H)p(H) + p(X \mid \neg H)p(\neg H)$$

is the marginal probability of X. When considered as a function of H, $p(X \mid H)$ is also the *likelihood* of H. The terms $p(X \mid H)$ and $p(H \mid X)$ are known as *prior* and *posterior* probabilities, respectively, reflecting the update in information after observing the data X. The hypothesis H may correspond to almost anything from a major scientific theory to a statement that the result of a coin toss is heads. In a typical statistical application different hypotheses H_i correspond to the possible values that a parameter in a statistical model can take.

Inferential methods based on Bayes' theorem were introduced already in the late 18th century by an English amateur mathematician Thomas Bayes and the prominent French scientist Pierre Simon Laplace, but they remained in the marginal until the latter part of the 20th century (Fienberg, 1992, 2006). This was mainly due to two reasons. First, the use of probability to describe uncertainty associated with non-random events was controversial. Second, as will be discussed later, the lack of computational methods and resources rendered Bayesian inference applicable only for elementary problems.

The controversy about Bayesian inference did not come from the use of Bayes' theorem, which is a result of elementary probability calculus, but from the specification of prior probability $p(H)$. In frequentist statistics, probability of an event A is taken to be the proportion of cases in an infinite population satisfying A (Good, 1959). Without imagining an infinite population of alternative realities one can not specify the probability of a hypothesis H, which either is or is not true with the actual state being unknown.

In Bayesian inference the probability of a hypothesis H is understood as a degree of belief and Bayes' theorem provides a way of updating it after observing data X. This subjective interpretation of probability and consequently the whole Bayesian approach can be derived from the need to quantify uncertainty in a coherent way (Jaynes, 2003). The coherence here refers to the requirement that the levels of uncertainty associated with different events should not be contradictory. Lindley (2000) provides motivation for the use of probability to describe uncertainty in an informal manner. For more rigorous presentation, the books by Bernardo and Smith (1994) and Robert (2007) derive the rules of Bayesian inference from decision theory.

It is often difficult to associate an exact probability to a hypothesis H. This may be due to lack of any meaningful prior information about H or the difficulty of elicitating expert's knowledge in probabilistic form (Garthwaite, Kadane and O'Hagan, 2005). There has been a lot of research on how to construct prior probabilities and distributions which contain minimum amount information about the question at hand starting already with the work of Bayes and Laplace. Many different rules for deriving such prior distributions have been proposed, which may result in same or different distributions depending on the problem (Kass and Wasserman, 1996). The use of non-informative priors, known as objective Bayes, is regarded by many to be inferior to full subjective analysis, but is widely used as a standard recipe for Bayesian inference (see the articles by Berger (2006) and Goldstein (2006), and the discussion following them).

The biggest advantage and at the same time the greatest disadvantage of the Bayesian approach is that it is a closed system for inference. When conducting Bayesian analysis one has to specify all possible hypotheses $\mathscr{H} = \{H_1, H_2, ...\}$ and assign prior probabilities to them. This ensures that the resulting inferences are coherent. However, it is not always possible to specify completely the set $\mathscr{H}$, as something unexpected might be seen. Also, as discussed above, the assignment of probabilities $p(H_i)$ is far from trivial. These difficulties have led some scientist to suggest that "the Bayesian theory is a theory of how to remain perfect but it does not explain how to become good" (Senn, 2011). While this exaggerates the problem of specifying hypotheses and prior probabilities, it has a grain of truth in that the results from a Bayesian inference are often sensitive to the prior information.

Jaynes (2003) tried to circumvent this problem by keeping a small probability, say 10^{-6}, that some alternative hypothesis H_A not included in $\mathscr{H}$ is true. However, if one does not include H_A and $p(X|H_A)$ in the inferences, the results are not coherent anymore and might be very far from correct (Fitelson and Thomason, 2008).

The discussion so far has been about a theoretical and idealistic way of conducting statistical inference. In practice, there is a large number of obstacles preventing a full Bayesian approach, and the statistician has to usually consider simplified hypotheses, which are known to be false, but which might nevertheless capture some aspects of reality. Along with problems of specifying prior probabilities, computational difficulties associated with the need to evaluate $p(X|H)$ and $p(H|X)$ result in approximate degrees of belief $\hat{p}(H|X)$.

Statistical pragmatism suggested by Kass (2011) offers a more realistic description of the actual practice of statistical inference. It calls for clear separation between the real world and the theoretical world, where the statistical models live in. The statistical models used can give meaningful results about the real world phenomena only if a strong connection can be established between these two worlds. With this view, a great advantage of the Bayesian approach is in the unified framework, which specifies how inferences should be made in the theoretical world. The implications of the different assumptions can be easily assessed by considering different prior distributions and likelihood functions.

This thesis is about developing ways to construct statistical models, which are families of joint distributions for the data X and some unknown parameters θ characterizing the model, and techniques that facilitate computation under these models. Whether, and to what to degree, a specific model corresponds to reality is out of the scope of this work and should be assessed by the scientist willing to use the methods developed here. We note that our motivation for model construction is slightly different from the one used in model construction for data analysis (Gelman et al., 2004). In data analysis, the objective is to construct a model which effectively extracts some information from the data and the model should not be treated as a fixed entity. Instead, the model should be modified or expanded, if it is found to represent poorly some aspect of the data (Gelman and Shalizi, 2012).

While all of the models developed in this work are motivated using a Bayesian approach, they could as well be used in a more traditional statistics framework. The methods for inferring the population structure described in articles (I) and (II) could be even viewed as maximum likelihood methods. Also, the approximations to the Wright-Fisher model could be utilized without prior distributions for the unknown parameters.

2.2 Model validation

The central topic of this work is the development of statistical models for making inferences on problems in population genetics. When building such models one is obliged to make compromises between an accurate representation of reality, identifiability of the model and computational efficiency. The resulting model represents only an approximation to the underlying biological processes and consequently the results of the inferences based on the model are biased. However, as the famous quote by George E. P. Box states, "all models are wrong, but some are useful" (Box, 1979). Thus it is important to recognize the assumptions and simplifications behind each model and to judge whether the model is a reasonable approximation for the problem at hand. This is also reflected in the statistical pragmatism described earlier, which makes a clear distinction between the theoretical and real worlds (Kass, 2011).

It should be noted that the statistical model is not the sole source of inaccuracy in this context. Even if one is able to formulate a probability model that is an accurate description of reality, statistical inferences under the model would be complicated by the computational difficulties discussed in the next section.

In the models developed in this work we can distinguish three levels of approximations, each of which may cause some bias in the inferences. We describe these using the model developed in (IV) as an example. First, the mathematical model serves as an idealistic and simplified description of the processes happening in reality. In (IV), we assume that the individuals are sampled from several populations which are related according to a tree. The genotypes of the individuals are assumed to follow simple Wright-Fisher model with conditionally independent loci given the tree. Second, statistical models are used as approximations to the mathematical models. In (IV), we use Beta-distributions to approximate the infinite population limit of the Wright-Fisher model to make computations feasible. The statistical model also includes the prior distributions for the model parameters and these do not have any biological interpretation. Third, actual inferences under the statistical model need to be carried out using approximative computational methods. These methods may introduce a bias, as is the case with Laplace's approximation and the AMIS algorithm in used in (IV), and this bias needs to be quantified.

The first level of approximation, the mathematical model, is the most crucial part to the inferences. The relevance of a specific mathematical model to the scientific problem at hand should be assessed by the scientist wishing to use the model and it is thus mostly outside the topic of this work.

One possibility to quantify how well the inferences reflect reality is to perform posterior predictive checks as advocated by Gelman, Meng and Stern (1996). The idea is to first generate replicate data sets X^{rep} from the posterior predictive distribution

$$p(X^{rep}|X) = \int_{\Theta} p(X^{rep} \mid \theta) p(\theta \mid X) d\theta.$$

The simulation of replicate data sets can be incorporated in a Monte Carlo algorithm in a straightforward manner. Then, some aspects of the replicated data sets can be compared with observed data X either visually or using some test criteria. If the observed data X appears atypical in comparison to the replicated data sets, this is seen as an indication that the model does not adequately represent the variation in the data X. What action one should take based on these comparisons is left to subjective consideration and no general guidelines can be given. As an example of posterior predictive checks, we compared the distributions of pairwise F_{ST} (Balding, 2003) values with the human data and simulated data in (IV).

The other two levels of approximation, the statistical model and the computational methods, can usually be more easily quantified. For example, computations under the mathematical model may be possible in some cases and these could be compared to those done under the statistical model. Analysis of simulated data sets, generated from either the mathematical or statistical model, provide a way of evaluating the results when the true values of model parameters are known. Multiple different computational techniques may be utilized to quantify their efficiency and bias. For all of the models developed in the articles (I-V) we have utilized these methods of validation.

3 Computation for Bayesian inference

The biggest obstacle that kept Bayesian methods outside mainstream statistics before 1980's was not a philosophical one, but computational. Many inferential problems in the Bayesian framework require maximization and integration of complicated functions. Analytical solutions to these exist in several cases, but in general one has to resort to numerical methods, which were not widely available before the arrival of desktop computers.

To exemplify the typical computational issues arising in Bayesian statistics, consider a situation where we seek to compute the posterior expectation $E(h(\theta) \mid X)$ of some random quantity θ after a transformation h. The function h might be the identity function, a projection of some component of θ or any other function for which the expectation exists and is finite. To compute this expectation one has to evaluate the integral

$$E(h(\theta) \mid X) = \int_{\Theta} h(\theta)p(\theta \mid X)d\theta, \tag{3.1}$$

where $p(\theta \mid X)$ is the posterior distribution of θ after observing data X and Θ is its support.

There does not exist any single method which would work for all integration problems of the type (3.1), but the different methods have their own strengths and weaknesses. When choosing a method to evaluate a specific expectation, these differences should be acknowledged. The numerical approximation methods can be broadly divided into two categories: deterministic methods and Monte Carlo methods (Evans and Swartz, 2000). We provide a short overviews of these and present in detail some particular methods which are used in the articles.

3.1 Deterministic methods

Deterministic approximation methods for evaluating integrals include a wide variety of different approaches. Quadrature based are among the oldest techniques of approximate integration. They approximate the integral as a weighted sum

$$E(h(\theta) \mid X) \approx \sum_{i=1}^{N} w_i h(\theta_i)p(\theta_i \mid X), \tag{3.2}$$

where the points θ_i and their weights w_i are based on a specific rule. As an example consider that $\Theta = [0, 1]$, $\theta_i = (i + 1/2)/N$ and $w_i = 1/N$. As N increases, the approximation (3.2) based on this rule will converge to the true value of the expectation (3.1), although it might be computationally very inefficient. Typically the rule is chosen so that it will produce exact estimates for some class of functions, which depends on n. The quadrature methods are generally very efficient in low dimensions, but as the dimension increases, their accuracy quickly decreases.

Another possibility to evaluate integrals is obtained by approximating the integrand $h(\theta)p(\theta \mid X)$ with some function $\phi(\theta)$. The approximation is assumed to be exact when $\lambda = \lambda_0$, where λ is a parameter characterizing the integrand. Typically, $\lambda_0 = \infty$ or $\lambda_0 = 0$, and the methods are therefore called asymptotic approximations. In the estimation of posterior expectations the sample size of the data X can often be used as the parameter λ. The asymptotic approximations provide a computationally very fast way to evaluate integrals, but their drawback is that the accuracy can not be easily evaluated.

Laplace's approximation Perhaps the most widely used asymptotic method is Laplace's approximation, which was used (IV) to marginalize out allele frequencies. We only consider here this approximation for marginal posterior densities, which has proven popular in Bayesian statistics (Tierney and Kadane, 1986; Rue, Martino and Chopin, 2009). We assume that we can evaluate the joint posterior distribution $p(\theta, \delta \mid X)$ and wish to compute the marginal posterior distribution $p(\delta \mid X)$. Bayes' theorem (2.1) implies that the marginal can be computed as

$$p(\delta \mid X) = \frac{p(\theta, \delta \mid X)}{p(\theta \mid \delta, X)}, \tag{3.3}$$

which is valid for all values of θ in the support of the full conditional distribution $p(\theta \mid \delta, X)$. The problem with direct use of equation (3.3) is that the full conditional distribution is rarely available analytically. Laplace's approximation is obtained by replacing $p(\theta \mid \delta, X)$ in (3.3) with a Gaussian distribution with parameters $\hat{\theta}$ and $\hat{\Sigma}$, where $\hat{\theta}$ is the mode of $p(\theta \mid \delta, X)$ and $\hat{\Sigma}$ is the minus inverse of the Hessian of $log(p(\theta \mid \delta, X))$ evaluated at $\hat{\theta}$. The motivation for this comes from the central limit theorem, which ensures under certain regularity conditions that as the sample size of the data goes to infinity, the full conditional distribution will converge to a Gaussian distribution.

3.2 Monte Carlo methods

Monte Carlo (MC) methods are a class of methods which are based on simulating random variables (Robert and Casella, 2004). For evaluating integrals such as (3.1), the idea behind them is simple. First, simulate N variables $\theta_1, \ldots, \theta_N$ from the posterior distribution $p(\cdot \mid Z)$. Then, approximate the integral as

$$E(h(\theta) \mid X) \approx \frac{1}{N}\sum_{i=1}^{N} h(\theta_i), \tag{3.4}$$

which is unbiased as it is easily seen.

While (3.4) would in theory provide an easy way of estimating the expectation (3.1), in practice simulation from $p(\cdot \mid X)$ is rarely directly possible. Importance sampling is an alternative approach where the variables are simulated from another distribution q known as a *proposal* distribution. The only requirement for the proposal is that the support of the integrand $h(\theta)p(\theta \mid X)$ is contained in the support of q. Given a sample $\theta_1, \ldots, \theta_N$ from q, the IS estimate of the expectation is

$$E(h(\theta) \mid X) \approx \frac{1}{N}\sum_{i=1}^{N} w_i h(\theta_i), \tag{3.5}$$

where $w_i = \frac{p(\theta_i|X)}{q(\theta_i)}$ is the weight of the ith sample. The expectation of each term in the sum is

$$E(w_i h(\theta_i)) = \int_\Theta \frac{h(\theta_i)p(\theta_i \mid X)}{q(\theta_i)} q(\theta_i) d\theta_i = \int_\Theta h(\theta)p(\theta \mid X)d\theta, \tag{3.6}$$

which indicates that the estimate (3.5) will converge to the true value of the expectation (3.1) as the number of sample N increases.

While the use of almost any proposal distribution will guarantee that estimate will be unbiased in the limit, the choice of q is crucial for the performance of the importance sampler. If q is a poor approximation of the posterior $p(\cdot \mid X)$, then most weights w_i will be close to zero and have only negligible influence on the estimator of the integral and the algorithm will be very inefficient. On the other hand, if the tails of $p(\cdot \mid X)$ are heavier than those of q, the ratio $p(\theta \mid X)/q(\theta)$ is not bounded in the support of q and the estimator may have infinite variance.

Adaptive multiple importance sampling Here we describe a recently introduced algorithm called Adaptive multiple importance sampler (AMIS, Cornuet et al., 2012), which is used in article (IV) for integration over branch lengths of a population tree. To understand the idea behind the AMIS algorithm, we briefly discuss the deterministic multiple mixture sampling described in Owen and Zhou (2000), which utilizes a collection of proposal distributions q_i, $i = 1, \ldots, N$ instead of a single choice. The distributions q_i may be the same for several i or the may all be distinct. For each $i = 1, \ldots, N$, the sample θ_i is generated from distribution q_i and the weight w_i is calculated as if the sample was generated from the mixture

$$q(\theta) = \frac{1}{N} \sum_{i=1}^{N} q_i(\theta). \tag{3.7}$$

In practice, one usually uses much a smaller number d of distinct proposal distributions enabling faster evaluation of (3.7). The estimator (3.5) based on the sample $\theta_1, \ldots, \theta_N$ with weights computed using (3.7) is unbiased if the whole sample is used. This is seen by

$$\begin{aligned} E\left(\frac{1}{N}\sum_{i=1}^{N}\frac{p(\theta_i \mid X)}{q(\theta_i)}h(\theta_i)\right) &= \frac{1}{N}\sum_{i=1}^{N}\int_{\Theta}\frac{p(\theta_i \mid X)}{q(\theta_i)}h(\theta_i)q_i(\theta_i)d\theta_i \\ &= \frac{1}{N}\sum_{i=1}^{N}\int_{\Theta}\frac{p(\theta \mid X)}{\frac{1}{N}\sum_{j=1}^{N}q_j(\theta)}h(\theta)q_i(\theta)d\theta \\ &= \int_{\Theta}\frac{p(\theta \mid X)h(\theta)\sum_{i=1}^{N}q_i(\theta)}{\sum_{j=1}^{N}q_j(\theta)}d\theta \\ &= \int_{\Theta}p(\theta \mid X)h(\theta)d\theta. \end{aligned}$$

The AMIS algorithm is an adaptive version of the deterministic multiple mixture. Instead of predefined set of proposal distributions, it learns them based on previously generated samples. The AMIS algorithm proceeds iteratively in $T + 1$ steps. First, at step 0, N_0 samples $\theta_{1,0}, \ldots, \theta_{N_0,0}$ are generated from a proposal distribution q_0. The weight $w_{i,0}$ of each sample i is computed as in regular importance sampling. Then at step t, $1 \leq t \leq T$, a proposal distribution q_t is chosen based on samples and weights from previous steps. The choice of q_t can in principle be arbitrary, but usually is some form of parametric distribution $q(\cdot \mid \mu)$ and the parameter μ is estimated from the previous samples. N_t samples $\theta_{1,t}, \ldots, \theta_{N_t,t}$ are generated using this proposal q_t. The weight $w_{i,k}$ of every sample generated at this and the previous steps is calculated as

$$w_{i,k} = \frac{p(\theta_{i,k} \mid X)}{\left(\sum_{j=0}^{k} N_j\right)^{-1} \sum_{j=0}^{k} N_j q_j(\theta_{i,k})}, \tag{3.8}$$

for $0 \leq k \leq t$, $1 \leq i \leq N_k$. Thus at the step t the weights associated with the samples are the same as if they were generated using the deterministic multiple mixture, with the N_0 first samples from q_0, the next N_1 from q_1 and so on. Difference to the deterministic multiple mixture comes from the fact that the proposal distribution q_t depends on the samples generated on previous steps. As a result, the AMIS estimator is biased

$$E\left(\frac{1}{N}\sum_{t=0}^{T}\sum_{i=1}^{N_t} w_{i,t} h(\theta_{i,t})\right) \neq E(h(\theta \mid X), \tag{3.9}$$

where $N = \sum_{t=0}^{T} N_t$. The degree of bias can be controlled by generating a large proportion of the samples from the initial proposal distribution q_0. In practice, values of $N_0 = \frac{1}{2}N$ have been used in the article (IV) and by Cornuet et al. (2012).

3.3 Markov chain Monte Carlo

The construction of a proposal distribution q, or even a family of such distributions as with the AMIS algorithm, is not feasible in many practical problems. For example, this may be due to the high dimensionality of the parameter space Θ. Markov chain Monte Carlo algorithms (MCMC, Robert and Casella, 2004) provide a way of sampling from complex distributions without requiring a global approximation to it. Instead, many of the MCMC algorithms approximate the distribution only locally. While the history of Markov chain Monte Carlo methods can be traced back to the early 1950's, they remained mostly absent from the statistical literature for a long time. In the last two decades MCMC methods have gained a huge popularity in statistics following the groundbreaking paper by Gelfand and Smith (1990), who demonstrated the potential of MCMC in a wide variety of situations (Robert and Casella, 2011). Their almost universal applicability and the ease of implementation have made MCMC the first choice for computing intractable integrals in Bayesian statistics.

MCMC algorithms operate by simulating a Markov chain $(\theta_1, \theta_2, \ldots)$, whose stationary distribution is the posterior $p(\cdot \mid X)$. After generating N samples from the Markov chain, the expectation (3.1) can be estimated with formula (3.4). Under some specific criteria, which are satisfied by the standard algorithms, the estimate converges to (3.1) as N increases to infinity. In practice, the N_0 first samples are usually discarded as burn-in and the estimator

$$E(h(\theta) \mid X) \approx \frac{1}{N - N_0} \sum_{i=N_0+1}^{N} h(\theta_i) \tag{3.10}$$

is used to reduce the dependence on the possibly poorly chosen starting value.

We describe briefly the Metropolis-Hastings (MH) algorithm, which along with the Gibbs sampling are the two most popular MCMC methods. For using an MH algorithm to simulate values from the posterior distribution $p(\theta \mid X)$, one has to define two components: an initial value θ_0 and a proposal distribution $q\left(\cdot \mid \theta^*\right)$. The MH algorithm then proceeds as follows:

For $t = 1, 2, \ldots$

- Sample θ^* from the proposal $q\left(\cdot \mid \theta_{t-1}\right)$.
- Compute
$$\alpha_t = \min\left(1, \frac{p(\theta^* \mid X) q(\theta_{t-1} \mid \theta^*)}{p(\theta_{t-1} \mid X) q(\theta^* \mid \theta_{t-1})}\right).$$
- Set
$$\theta_t = \begin{cases} \theta^* & \text{with probability } \alpha_t \\ \theta_{t-1} & \text{with probability } (1 - \alpha_t). \end{cases}$$

The proposal distribution $q\left(\cdot \mid \theta^*\right)$ has a similar role in determining the performance of the algorithm as the proposal q has in importance sampling. While the choice of a proposal is easier for MH, because of the need for only a local approximation, MCMC algorithms are often implemented for complex and high-dimensional problems. A typical choice is a symmetric Gaussian proposal distribution with a covariance matrix Σ. It has been shown that the optimal choice of Σ should be proportional to the covariance structure of the posterior distribution $p(\theta \mid X)$ under specific conditions (Roberts and Rosenthal, 2001). However, the covariance structure of the posterior is usually unknown and estimating it constitutes a similar problem as evaluating (3.1).

Haario, Saksman and Tamminen (2001) introduced the Adaptive Metropolis (AM) algorithm, which adaptively learns the covariance structure during iterations. The resulting process is not a Markov chain anymore, as the distribution of θ_t depends on the whole history $\theta_0, \ldots, \theta_{t-1}$, but it belongs to the class of adaptive MCMC methods (Haario, Saksman and Tamminen, 2001; Andrieu and Thoms, 2008; Roberts and Rosenthal, 2009). The convergence properties of such methods have been under intensive study for the last decade, and the algorithms have proved to be effective in situations where standard MCMC algorithms fail without careful tuning. We implement AM in article (V) to facilitate inference of population genetic parameters.

4 Population structure, clustering and classification

A central concept in population genetics and subsequently in many other fields of biology is that of a population. Much of the early work by Wright, Fisher and Haldane in the 1920's considered the evolution of genetic patterns in a population, but no generally agreed definition of a population exists even today. Waples and Gaggiotti (2006) list 18 distinct definitions of a biological population in ecology, evolution and statistics. The vagueness about the concept of a population reflects the fact that it is an artificial construction, which is used as a simplified representation of complex biological processes.

In this work we follow Waples and Gaggiotti (2006) and define population as:

> A group of individuals of the same species living in close enough proximity that any member of the group can potentially mate with any other member.

They termed this definition as the evolutionary paradigm and it is compatible with the main body of population genetics research. It also offers a clear baseline for the inference of population structure which is the topic of this section.

4.1 Inferring population structure

The need to make inference about the unknown population structure comes up in many biological studies. For example, in conservation biology the viability of an endangered species depends on whether it forms a single panmictic population or is fragmented into several isolated populations (Hedrick, 2001; Pearse and Crandall, 2004). In genetic association studies the failure to take population structure into account may result in severely biased results (Marchini et al., 2004). For studying the human evolutionary history the genetic structure of present-day populations may offer important insights about the past (Rosenberg et al., 2002).

The inference about population structure is usually conducted using genetic markers, which are locations (*loci*) in the genome where variation exists between individuals. Different variants of the same marker are called *alleles*. Examples of genetic markers include single nucleotide polymorphisms (SNP, Morin et al., 2004), where a single nucleotide site has more than one nucleotide in population, and microsatellites (Ellegren, 2004), which are tandem repetitions of short sequences with typically only a few nucleotides. For SNPs the number of alleles present is usually two, whereas microsatellites may have dozens of distinct alleles.

A single genetic marker rarely contains enough information about the population structure and several marker loci are needed. The number of loci used may be almost anything ranging from a few to several hundreds of thousands, as is the case with human SNPs (Li et al., 2008). Many loci have the potential to complicate the inference by creating dependence between the markers. This can be avoided by selecting markers which are sufficiently far apart in the genome, so that recombination breaks links between them.

In diploid organisms each individual carries two copies of the same marker gene, one from each parent. Many genotyping techniques can not identify the phase of the alleles, but return instead only the genotype of the individual. For example, consider that there are two types of alleles at a locus: A and a. Then the genotypes which can be observed are AA, Aa and aa.

We now look at the frequencies of the different genotypes in a population of infinite size. Throughout this work we use the term frequency to denote the proportion of an allele or a genotype in a population, instead of the total number. The genotype frequencies of a selectively neutral locus are completely determined by the allele frequencies, and consequently the basis of identifying population structure lies in allele frequencies. To see the reason for this we have to look at what Mendel's laws imply about genotype frequencies (Ewens, 2004). Consider again a biallelic locus with alleles A and a. Suppose that the genotypes AA, Aa and aa are present in a population with frequencies X, $2Y$ and Z, respectively. According to the Mendel's law the genotype frequencies of the next generation are now

$$\begin{aligned} X' &= X^2 + \frac{1}{2}(4XY) + \frac{1}{4}\left(2Y\right)^2 = (X+Y)^2 \\ 2Y' &= \frac{1}{2}(4XY) + \frac{1}{2}(2Y)^2 + 2XZ + \frac{1}{2}(4YZ) = 2(X+Y)(Y+Z) \\ Z' &= \frac{1}{4}\left(2Y\right)^2 + \frac{1}{2}(4YZ) + Z^2 + = (Y+Z)^2. \end{aligned}$$

By denoting the frequency of the allele A with $p = (X+Y)$ these can be written as

$$X' = p^2, 2Y' = 2p(1-p) \text{ and } Z' = (1-p)^2.$$

This result is known as the Hardy-Weinberg law, and it states that in an infinite population with random mating, allele frequencies completely define genotype frequencies in a selectively neutral locus. A population satisfying this is referred to be in Hardy-Weinberg equilibrium.

However, if there is further substructure in a population, then the HW-equilibrium will usually not hold. As an example, consider a population which consists of two subpopulations that are both in HW-equilibrium with the frequencies of allele A being p and $p+a$, respectively, such that $0 \le p \le 1$ and $-p \le a \le 1-p$. Assume that the population is infinite in size and a proportion $0 < q < 1$ of the individuals belong to the first subpopulation. The frequency of A in the whole population is now given by $p' = p + (1-q)a$. Then for the whole population to be in Hardy-Weinberg equilibrium we need that

$$\begin{aligned} 2p'(1-p') &= q2p(1-p) + (1-q)2(p+a)(1-p-a) \Leftrightarrow \\ (p+(1-q)a)(1-p-(1-q)a) &= qp(1-p) + (1-q)(p+a)(1-p-a) \Leftrightarrow \\ p(1-p) - p(1-q)a + (1-q)a(1-p-(1-q)a) &= p(1-p) + (1-q)(a(1-p-a) - pa) \Leftrightarrow \\ (1-q)a(1-2p-(1-q)a) &= (1-q)a(1-2p-a) \Leftrightarrow \\ (1-q)qa^2 &= 0 \Leftrightarrow \\ a &= 0. \end{aligned}$$

In other words, the two subpopulations need to have exactly the same frequency of A for the whole population to be in HW-equilibrium. On the other hand, it is easily seen that if we take a random sample of a population in Hardy-Weinberg equilibrium, then the genotypes of the sample will also be in HW-equilibrium. Thus the inference about population structure can be framed as finding maximal groups from a sample, such that in each group HW-equilibrium is satisfied.

There are alternative approaches to learning population structure that try to infer continuous differences between individuals, instead of assigning individuals to discrete populations. For example methods based on principal component analysis have been introduced and proven successful in the analysis of large SNP data sets (Patterson, Price and Reich, 2006). There is evidence that at least in humans genetic variation is of more continuous form, which would indicate that such methods would be preferable over models which assume discrete populations (Novembre et al., 2008). We do not pursue here the debate whether genetic variation is discrete or continuous. We only note that we do not assume that discrete populations exist in most cases, but they can serve as a reasonable approximation. Also, detecting continuous differences often requires a lot of genetic data, and while this is currently available for humans, it is more rare for many other species.

The problem of finding discrete population structure has been under intensive study for over a decade and many programs have been introduced to conduct the inference, such as STRUCTURE (Pritchard, Stephens and Donnelly, 2000; Falush, Stephens and Pritchard, 2003), Partition (Dawson and Belkhir, 2001), BAPS (Bayesian analysis of population structure; Corander, Waldmann and Sillanpää, 2003; Corander et al., 2004, II), Geneland (Guillot et al., 2005; Guillot, 2008) and TESS (Chen et al., 2007). The models behind these vary in details, but they follow more or less the same logic. Outside the field of population genetics, the inference about discrete population structure can be seen as clustering.

4.2 Predictive clustering

In this section we derive results concerning the predictive approach to clustering, which is the task of classifying observations into groups without additional knowledge of the groups (Jain, Murty and Flynn, 1999). This is motivated by the problem of inferring population structure from genetic data described in the previous section, but applications of the clustering framework are not limited to population genetics. In fact, the need to group observations without background information occurs in many different fields of science and clustering is one of the central topics in machine learning and statistics.

Consider that we have a total of N observations and associated with each observation i, $i = 1, \dots, N$, we have a vector x_i of some measurements. Based on these measurement vectors x_i we wish to group the observations to classes $s_1, \dots, s_K$, for some $K \in \{1, \dots, N\}$, so that each observation i belongs to exactly one class s_c, $1 \leq c \leq K$. In population genetic context the observations $1, \dots, N$ are individuals and the vector x_i contains the allele types observed for individual i over several marker gene loci. With diploid organisms there are two measurement vectors associated with each individual, but the models considered below easily extend to this case. In fact, they allow a varying number of measurement vectors associated with each individual.

Perhaps the most common methods used for clustering are those based on some distance measure between the measurement vectors. These methods proceed by computing the distance $d(x_i, x_j)$ between each pair of measurement vectors and collecting them to a matrix D. The observations are then grouped hierarchically according to some criterion based on the distances until they all belong to the same group or class. A clustering is then obtained by stopping the process when K groups are remaining.

We concentrate here on Bayesian probabilistic clustering methods, which assume that the measurement vectors associated with observations belonging to a single cluster were generated by a common cluster-specific probability distribution. Such clustering methods are comprised of two equally important ingredients: the data generating model associated with each cluster and the prior distribution for the clusterings. The importance of the latter will be discussed in the next subsection.

A clustering of observations is represented by a partition of the set $[N] = \{1, \ldots, N\}$, which is a collection $\{s_1, \ldots, s_K\}$ of subsets s_c of $[N]$ such that $\bigcup_{c=1}^{K} s_c = [N]$ and $s_c \bigcap s_{c'} = \emptyset$ for all $c \neq c'$. In other words, each element of $[N]$ belongs to exactly one set s_c. The labels c do not carry any information and the subsets s_c are identified only in terms of their content.

The conditional probability of the whole data $\mathbf{x} = (x_1, ..., x_N)$ given the partition S is assumed to be of the form

$$p(\mathbf{x}|S) = \prod_{c=1}^{K} p(\mathbf{x}_{s_c}), \tag{4.1}$$

where $\mathbf{x}_{s_c}$ is the collection of measurement vectors for observations in s_c. Statistical inference about the unknown partition S is based on the posterior distribution

$$p(S|\mathbf{x}) = \frac{p(\mathbf{x}|S)p(S)}{p(\mathbf{x})}, \tag{4.2}$$

which is obtained after specifying the prior distribution $p(S)$ for the partitions. The marginal probability of the data is given by the sum

$$p(\mathbf{x}) = \sum_{S \in \mathscr{S}} p(\mathbf{x}|S)p(S),$$

where $\mathscr{S}$ denotes the space of all possible partitions.

While the posterior distribution (4.2) represents all the information about the unknown partition S, the complicated structure and the size of the partition space $\mathscr{S}$ makes it necessary to summarize it in some way. This is often done by choosing the partition $\hat{S}$, which maximizes the posterior (4.2). In theory the search for $\hat{S}$ necessitates the exploration of the whole space $\mathscr{S}$ and provides no additional computational advantage, but in practice $\hat{S}$ is approximated using a heuristic algorithm as is done in the most recent versions of BAPS (II). The actual posterior probability of $\hat{S}$ might be very low especially in high dimensional problems and, consequently, it might not represent the underlying structure adequately. This can be acknowledged by using decision theory to derive optimal estimates of the partition under a specific loss function (Robert, 2007; Corander, Gyllenberg and Koski, 2009).

In this work we assume that each measurement vector x_i consists of L discrete features $x_{i,j}$, which are assumed to be independent in each cluster. This corresponds to unlinked marker genes in the context of the inference of population structure. We assume that $x_{i,j}$ can take r_j different values, which occur at frequencies $\theta_{j,c} = (\theta_{1,j,c}, \ldots, \theta_{r_j,j,c})$ in cluster s_c. Furthermore, by assuming that the observations in cluster s_c constitute a random sample from the cluster, the conditional cluster-specific probability for $\mathbf{x}_{j,s_c}$, the measurements of feature j in s_c, has the product form

$$p(\mathbf{x}_{j,s_c}|\theta_{j,c}) = \prod_{l=1}^{r_j} \theta_{l,j,c}^{n_{l,j,c}}, \tag{4.3}$$

where $n_{l,j,c}$ is the number of times value l was measured for feature j in cluster s_c. A prior distribution for the parameters $\theta_{l,j,c}$ needs to be defined to obtain a marginal distribution for $\mathbf{x}_{j,s_c}$. A natural choice is a symmetric Dirichlet distribution with parameters $(\alpha_j, \ldots, \alpha_j)$ which is conjugate to (4.3) and facilitates an analytic marginalization over the frequencies. Common values used for α_j in the absence of auxiliary information include r_j^{-1}, 1/2 and 1. With a Dirichlet prior the marginal distribution is now given by

$$p(\mathbf{x}_{j,s_c}) = \frac{\Gamma\left(r_j\alpha_j\right)}{\Gamma\left(\sum_{l=1}^{r_j}\left(\alpha_j + n_{l,j,c}\right)\right)} \prod_{l=1}^{r_j} \frac{\Gamma\left(\alpha_j + n_{l,j,c}\right)}{\Gamma\left(\alpha_j\right)}, \tag{4.4}$$

Corander, Gyllenberg and Koski (2007) show that this model and the choice of Dirichlet prior can be motivated in the biological context by assuming that the values r_j are known. If r_j is unknown, the marginal distribution (4.4) is replaced by Ewens sampling formula under a particular assumption about exchangeability (Ewens, 2004).

4.3 Prior distribution for the partition

The choice of prior distribution is crucial for satisfactory performance of the above clustering method. As with any other application of Bayesian inference, the prior distribution should be chosen on the basis of some auxiliary information. However, this information is not always available and in such cases a weakly informative prior distribution might be considered attractive. The size and complicated structure of the space of possible partitions of the set $[N]$ makes it difficult to recognize the effects of different prior distributions to the results. The number of different partitions of a set of size n is given by the nth Bell number

$$B_n = e^{-1} \sum_{m=1}^{\infty} \frac{m^n}{m!}$$

(Rota, 1964; Stanley, 2012). Bell numbers also satisfy recurrence relation

$$B_{n+1} = \sum_{m=1}^{n} \binom{n}{m} B_m,$$

which can be used to compute numerically B_n for small values of n.

We do not try to describe here different strategies for choosing the prior distribution and evaluate the strength of various approaches, but refer to Quintana (2006) for discussion of some particular choices. Instead, we look at two distributions, which at first sight may both look non-informative, and show how they may have a dramatic impact on the inference. First, we consider the uniform distribution on $\mathscr{S}$, the space of all partition of the set $[N]$,

$$p_1(S) = \frac{1}{B_N}. \tag{4.5}$$

With this prior, the mode of the posterior distribution (4.2) coincides with the maximum likelihood estimate $\hat{S}$, which is obtained by maximizing the conditional probability of whole data (4.1) as a function of S. The motivation to use uniform prior may thus come from the desire to find the clustering S which best describes data and "let the data speak for itself".

On the other hand, if we wish to compute probabilities such as

$$p(i, j \in s_c, \text{ for some } c \mid \mathbf{x}) = \sum_{S; i,j \in s_c} p(S \mid \mathbf{x}) \tag{4.6}$$

or

$$p(|S| = K \mid \mathbf{x}) = \sum_{S; |S|=K} p(S \mid \mathbf{x}), \tag{4.7}$$

then the prior has a considerable impact on the results. Under the uniform prior (4.5) the prior probability corresponding to (4.6) is

$$p_1(i, j \in s_c, \text{ for some } c) = \frac{B_{N-1}}{B_N}.$$

Figure 1 shows how this probability decays to zero as N increases. Similarly, the prior probability corresponding to (4.7) is given by

$$p_1(|S| = K) = \frac{S(N, K)}{B_N}, \tag{4.8}$$

where $S(n, k)$ denotes the Stirling number of the second kind and equals the number of ways in which a set of size n can be partitioned into k non-empty subsets (Stanley, 2012). Values of $S(n, k)$ can be computed using the recurrence

$$S(n, k) = kS(n-1, k) + S(n-1, k-1).$$

As shown in in Figure 1 the uniform prior (4.5) favors strongly values of K which are approximately $N/2$. These results imply that while the posterior mode usually a good estimate of the true clustering, the uncertainty associated with it might not be adequately captured by the posterior distribution under (4.5).

The behavior of the uniform prior (4.5) on the distribution of the number of clusters (4.8) could suggest the uniform distribution on the number of clusters as an alternative possibility

$$p_2(S) = \frac{1}{N} \frac{1}{S(N, |S|)}.$$

However, this distribution contains implicitly strong preferences about the partitions, which can be seen by looking at the prior odds in favor of partition S_1 over S_2

$$\frac{P_2(S_1)}{P_2(S_2)} = \frac{S(N, |S_2|)}{S(N, |S_1|)}$$

Figure 1 plots the logarithm of the above odds for 50 observations, which clearly indicates that partitions with K close to 1 or N are favored over others.

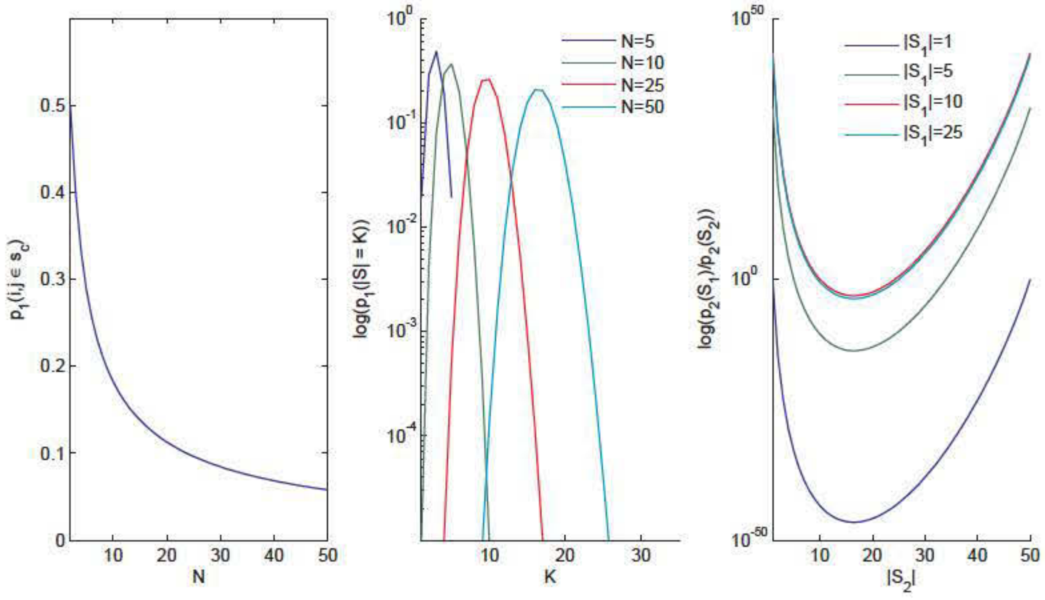

Figure 1: **Partition probabilities with uniform distributions on $\mathscr{S}$ and K.** Left: Probability that two observations belong to the same cluster under p_1. Center: Logarithm of probability distribution of the number of clusters under p_1. Right: Logarithm of the odds for S_1 over S_2 with $N = 50$ observations under p_2.

Prior utilizing spatial information The previous examples were given to highlight what kinds of assumptions are hidden behind two simple prior distribution, which could at first sight look like non-informative choices. Similar to any other forms of Bayesian inference, if other relevant information is available for the observations, then this can and should be incorporated in the prior distribution.

For example in population genetic studies, the sampling locations of the individuals may be available. It could be expected that individuals in close proximity of each other are genetically more similar than individuals further away from each other. This similarity could be caused by a continuous landscape of genetic variation, or geographic barriers that prevent gene flow between adjacent areas (Wright, 1943; Guillot et al., 2009). Thus when the genetic information is sparse, the sampling locations of the individuals can be used to facilitate the inference about population structure.

Several models for detecting discrete population structure with the assistance of coordinate information have been introduced in last few years (Guillot et al., 2005; Francois, Ancelet and Guillot, 2006; Chen et al., 2007; Guillot, 2008, I). All of the methods are based on constructing Voronoi tessellations over the spatial domain (Kreveld et al., 1997). Voronoi tessellation is a division of the plane into polygons based on a set of points A, so that each point y in the plane belongs to the polygon associated with the point in A closest to y. Differences in the models stem from the choice of set A to construct the tessellation. Guillot et al. (2005) assumed that A was random and distributed according to a homogeneous Poisson process, whereas the model proposed by Francois, Ancelet and Guillot (2006) used the sampling coordinates of individuals as A.

We introduced in article (I) an extension to the basic clustering method of BAPS, where a Voronoi tessellation is constructed based on the sampling coordinates of the individuals. The tessellation induces a graph $G = (V, E)$ known as Delaunay triangulation. G is defined by letting the set of vertices V to be the individuals $[N]$ and including an edge (i, j) in the graph if the polygons corresponding to individuals i and j in the tessellation are neighbors. A prior distribution on the space of partitions $\mathscr{S}$ is then formulated based on G. To facilitate fast computation of the prior, the model considers only graphs G which are triangulated (Lauritzen, 1996). When the Delaunay triangulation is not a triangulated graph, edges are added to it to produce a triangulated version G^* (Heggernes, 2006). A graphical probability model associated with a triangulated graph G has the property that the joint probability of some variable $y^{[N]} = (y_1, \ldots, y_N)$ over the vertices can be computed as

$$p(y^{[N]}) = \frac{\prod_{c\in} p(y^c)}{\prod_{s\in\mathscr{P}} p(y^s)},$$

where $\mathscr{C}$ and $\mathscr{P}$ are sets consisting of specific subsets of V known as cliques and separators, respectively, and y^d denotes the variables associated with the vertices in $d \subset V$. The prior on partitions is now constructed by letting y_i denote the cluster to which individual i is assigned and defining probabilities $p(y^d)$ to favor subsets with small number of clusters.

An advantage of this specification over the earlier models is that the posterior probability of a partition S (4.2) can be analytically computed up to a normalizing constant, avoiding the need to implement a Markov chain Monte Carlo algorithm as in Guillot et al. (2005) and Francois, Ancelet and Guillot (2006). The performance of the different methods has been the topic of several reviews (Frantz et al., 2009; Francois and Durand, 2010; Safner et al., 2011). The model proposed in article (I) has been found generally competitive to the other methods, when the assumptions behind it are not violated. When the underlying genetic structure has a continuous form, the methods might occasionally create artificial clusters.

Constrained spaces of partitions Prior information on the partition may also be available in terms of constrains, so that instead of the whole space of partitions $\mathscr{S}$ the support of the prior distribution is a subspace $\mathscr{S}' \subset \mathscr{S}$. In the inference of population structure of diploid organisms, there are usually two measurement vectors associated with each individual. The partition of N individuals into populations can now also be seen as a problem of clustering of $2N$ measurement vectors under the constraint that the measurements associated with a single individual should belong to same population. Similarly, if we have knowledge that some individuals belong to the same population this restricts the space of possible partitions.

Another form of constraint which is often utilized is restriction on the maximum number of clusters M. It might be deemed impossible, or at least highly unlikely, that N observations would represent K close to N distinct clusters. Such a constraint is used in the BAPS program for the uniform prior (4.5) and it usually results in faster computation. If M is chosen big enough then this restriction should not affect the accuracy of the results, as only a negligible amount of posterior mass would be assigned to partitions with $K > M$. The effect of M can always be evaluated by analyzing the data with multiple choices of M.

Sometimes it is preferable to constraint the partition to have exactly $K = M$ clusters. This might be the case, if the clusters are not assumed to be any real entities, but represent only an approximation to the unknown structure in the data. Then varying the value of M and for each searching the partition could reveal more about the underlying structure than a single search in the unconstrained space $\mathscr{S}$.

In (II) we introduced a possibility to BAPS for analysis with fixed number of populations. While mathematically this constraint changes prior distribution in a trivial way, it makes the search in the partition space more complicated. The stochastic greedy search algorithm implemented in BAPS takes advantage of the natural neighborhood structure in the partition space, which is broken by the constraint $K = M$. For example, consider that the underlying structure in the data is best represented with four clusters A_1, A_2, A_3, A_4, where A_1 and A_2 are similar to each other and correspondingly A_3 and A_4 are also similar. Suppose that we wish to infer the best partition with three clusters. Possible candidates for this would be $S_1 = \{\{A_1, A_2\}, A_3, A_4\}$ and $S_2 = \{A_1, A_2, \{A_3, A_4\}\}$, but the search algorithm can traverse between these two only by considering intermediate partitions $\{\{A_1, A_2\}, \{A_3, A_4\}\}$ or $\{A_1, A_2, A_3, A_4\}$, which have 2 and 4 clusters, respectively. We solved this issue by introducing complicated moves where the algorithm can temporarily visit in partitions with $M - 1$ or $M + 1$ clusters.

4.4 Predictive classification

A more extreme version of constraint is obtained in the case that the correct clustering is known for a subset of the observations and the remaining observations need to be assigned to these or other classes. This task is known as *classification* in the machine learning literature (Bishop, 2007). Instead of considering constrained partitions of $[N]$, it is helpful to denote the observations with known origin with set $[M]$ and let $[N]$ denote the remaining observations. The set $[M]$ and the measurement vectors associated with it are often called the training data, and $[N]$ with corresponding measurements the test data. The known classification structure of $[M]$ classification is denoted by T and the unknown structure of $[N]$ by S.

Classification can be divided into two different types: *supervised classification* and *semi-supervised classification* (Chapelle, Schlkopf and Zien, 2006). In supervised classification the classes present in T are assumed to represent an exhaustive set of possible classes and the observations in $[N]$ are classified to these. In semi-supervised classification the observations with unknown origin may be classified to the classes in T or some other previously unknown classes.

Traditionally classification is done by forming a classification rule based on the training data and then assigning the observations in $[N]$ one at a time to the classes. This is often a computationally efficient way of doing supervised classification when the amount of training data is large enough. However, it causes information to be lost, which could result in sub-optimal classification accuracy, when the amount of training data is small.

In article (III) we show that this loss of information can be avoided by considering the joint classifications of the test data. The classification is now based on the full posterior of S similar to (4.2)

$$p(S|\mathbf{x}, \mathbf{z}, T) = \frac{p(\mathbf{x}|S, \mathbf{z}, T)p(S \mid T)}{p(\mathbf{x} \mid \mathbf{z}, T)}, \tag{4.9}$$

where $\mathbf{x}$ and $\mathbf{z}$ are the measurement vectors associated with observations in $[N]$ and $[M]$, respectively. This is in contrast with marginal classification, where each observation $i \in [N]$ is classified according to the distribution

$$p(i \in t_c|x_i, \mathbf{z}, T) = \frac{p(x_i|i \in t_c, \mathbf{z}, T)p(i \in t_c \mid T)}{p(x_i \mid \mathbf{z}, T)}, \tag{4.10}$$

where t_c denotes any of the eligible classes. We also show that as the amount of training data increases, the classifications based on (4.10) and (4.9) will converge to each other for discrete measurement vectors. This result is intuitive, as the test data contains no additional information about the class-specific probability distributions when the amount of training data goes to infinity.

5 Modeling the history of populations

In the previous section methods for recovering population structure from genetic data were introduced. These methods are descriptive as they attempt to infer structure in the data, but do not offer any insight on how this structure has been generated. In this section we formulate models for inferring the evolutionary history of populations as a tree. We assume that at some point in history there has been a single ancestral population, and that the current populations have emerged after a series of splits from this ancestral population. The order and timing of these splits is described by a rooted bifurcating tree T, with leaves representing the observed populations and the root being the common ancestral population. Throughout this discussion we assume that the observed populations are known and we have sampled individuals from them.

We only consider model-based methods for reconstructing the evolutionary history of populations. There exists a vast literature on methods based on distances and other statistics computed for population pairs (e.g. Felsenstein, 2004, Chapter 11). A widely used measure for differentiation between populations is the F_{ST} correlation coefficient defined by Wright (Wright, 1943, 1949; Balding, 2003). F_{ST} can be used to estimate the divergence time of several populations from a common ancestor, as it increases almost linearly with the scaled time t/N, where t is the number of generations since the split and N is the population size (Nicholson et al., 2002).

In the following discussion we do not make clear a distinction between a population tree and a species tree, but use these terms interchangeably. Our motivation stems from the need to reconstruct the history of several closely related populations from same species, but the methods could also be applied to data from multiple species. In most of the models described here, the populations are assumed to evolve in complete isolation of each other after splits. Thus it could be argued that the difference between species and populations reflects only the time after split events. With sufficiently long time after the split, the groups could possibly become distinct species, whereas if the time spans are short, the groups would still represent a single species. However, it should be noted that at the species level mutation is usually expected to be the dominant process describing the evolution of genetic material. With more closely related populations the role of mutation depends on the genetic marker under study, but genetic drift plays also an important part.

5.1 Phylogenetic inference

The models developed here belong to the broad class of phylogenetic methods, which are used to analyze and represent differences between biological units such as species or populations (Felsenstein, 2004). Phylogenetic methods are perhaps most widely used in the analysis of differences between species from gene sequences. These are based on identifying same genes on multiple closely related species, quantifying the differences between the sequences and building a phylogenetic tree that gives a plausible evolutionary history to these differences. The methods used to infer the tree vary greatly, but they share the assumption that the differences in gene sequences between species are result of mutation events.

The main difficulty when interpreting a phylogenetic tree obtained from analysis of single gene is that it describes the evolutionary history of this single gene, not of the entire species (Pamilo and Nei, 1988; Degnan and Rosenberg, 2009; Edwards, 2009). These trees, called gene trees, may be very far from the species tree especially if the species are very closely related, and gene trees computed from two different genes have completely different topologies. Edwards (2009) argues that the species or population tree, which describes the evolutionary history of the species or populations under study, should be the central focus of phylogenetic studies instead of the gene tree.

The analysis of multiple genes simultaneously has the potential to improve the accuracy of the estimated species tree if the information from the different genes is properly combined. However, concatenation of the gene sequences, which has been a common strategy for multiple gene phylogenetics, has been shown to lead to inconsistent estimates of the species phylogeny (Kubatko and Degnan, 2007).

One important factor creating heterogeneity among gene trees is deep coalescence caused by genetic drift. Deep coalescence occurs when the ancestral lineages of two genes coexist the whole length of a branch in the species tree. It is most commonly associated with short branch lengths in the species tree, but it has the potential to cause bias in the estimates of the branch lengths even when the topology of the gene tree agrees with that of the species tree. This variation in gene trees can be taken into account by using coalescent theory to model the distribution of gene trees given a species tree (Liu et al., 2009). The basic coalescent process traces the ancestries of several genes from a population backwards in time until a common ancestor is found (Hein, Schierup and Wiuf, 2005). The coalescent provides an alternative sample based representation of the Wright-Fisher model of genetic drift, which will be discussed in the next subsection. The basic coalescent extends to samples from multiple populations by restricting the coalescent events which can occur at a given time point (Degnan and Rosenberg, 2009).

The main difficulty when using coalescent in species tree inference is that one has to consider all possible gene trees compatible with the species tree. In practice, this is usually done by MCMC simulation on the combined space of all gene trees and species trees, which is computationally very expensive when the number of individuals under study is large (Rannala and Yang, 2003; Wilson, Weale and Balding, 2003; Liu et al., 2008; Heled and Drummond, 2010).

Recently, there has been interest in developing methods which avoid the need to sample gene trees. Nielsen et al. (1998) showed how to compute the likelihood of a population tree for biallelic loci without mutation. The method is computationally expensive and suitable only for data sets with small number of populations and individuals. RoyChoudhury, Felsenstein and Thompson (2008) introduced a pruning algorithm for computing the likelihood of a tree based on dynamic programming that is computationally much more efficient. While in principle it is applicable to any multiallelic locus without mutation, the computation of likelihood has complexity $O(n^r)$, with r alleles and n individuals. The pruning algorithm has been generalized to biallelic loci with mutation by Bryant et al. (2012), who also introduced an MCMC implementation for conducting inference on the population tree.

In multiallelelic situation with mutation the computation of the coalescent likelihood is practically intractable. Inference in such cases is possible by using Approximate Bayesian Computation (ABC) methods, which do not require the evaluation of the likelihood (Beaumont, Zhang and Balding, 2002; Beaumont, 2010; Marin et al., 2011). Instead, these methods proceed by simulating replicate data sets with different parameter values and comparing some aspects of the replicates to the observed data. Parameter values which produce replicate data sets similar to the observed data are then considered to have high posterior probability. ABC methods were first developed for problems in population genetics and this application field has been central to the development of new methods. Currently there are multiple software packages implementing ABC methods, which can be used to infer the population or species tree (Cornuet et al., 2008; Lopes, Balding and Beaumont, 2009; Wegmann et al., 2010).

5.2 Modeling the change in allele frequencies

An alternative approach to the multispecies coalescent for modeling the population tree is obtained by considering population level properties such as allele frequencies. Genotypes of the individuals are in this approach used only to infer the allele frequencies in the observed populations. Such methods have been proposed and applied since the early days of computational phylogenetics, but have been somewhat overshadowed by the phylogenetic methods based on molecular evolution (Edwards and Cavalli-Sforza, 1964; Cavalli-Sforza and Edwards, 1967; Felsenstein, 1973; Thompson, 1975; Felsenstein, 1981). The change in the allele frequencies over time is modeled using Brownian motion, which is an infinite dimensional generalization of the Gaussian distribution (Mörters and Peres, 2010).

The biological motivation for using the Brownian motion for modeling the change of allele frequencies is obtained by considering the Wright-Fisher model first introduced by R.A. Fisher and further developed by Sewall Wright in the 1920's and 30's (Fisher, 1922, 1930; Wright, 1931; Ewens, 2004). The term Wright-Fisher model might refer to several different models of varying complexity that describe the evolution of gene frequencies in a population including multiple alleles, mutation and selection. We first describe the simple Wright-Fisher model with two selectively neutral alleles and no mutation, which forms the mathematical foundation behind most of the allele frequency based methods. The multispecies coalescence methods discussed in the previous subsection also assume that the alleles in the populations follow Wright-Fisher model.

Consider a randomly mating population of N individuals and assume that the population evolves in discrete, non-overlapping generations. Note that many textbooks in population genetics define the Wright-Fisher model with population size $2N$, as it is used in many applications to diploid organisms for which the number of genes is twice the number of individuals. Our choice of population size N is consistent with the description of the models in articles (IV) and (V). We follow a single locus for which there exists two different types of alleles A and a in the population. At any time point t there are X_t alleles of type A and $N - X_t$ of type a, $0 \leq X_t \leq N$. The individuals for generation $t+1$ are obtained as a random sample with replacement from the previous generation t. This implies that conditional on X_t the number of A alleles in generation $t+1$ has binomial distribution

$$X_{t+1} \mid X_t \sim Binomial(N, X_t/N). \tag{5.1}$$

From the above equation it is clear that if X_t takes value 0 or N at any generation t it will stay there, a situation which is known as fixation.

Equation (5.1) provides in principle the way of computing the conditional distribution of X_t given X_0 for any non-negative integer t , but is computationally tractable only for small values of t. However, if the population size N is not too small, the binomial distribution (5.1) may be accurately approximated with a Gaussian distribution. First, consider a scaling of the process where the time is divided by the population size $\tau = t/N$ and instead of the actual number of A alleles, we consider their relative frequency $\theta_\tau = X_\tau/N$. In the transformed scale we get

$$\theta_{\tau+1/N} \mid \theta_\tau \approx N\left(\theta_\tau, \frac{1}{N}\theta_\tau(1-\theta_\tau)\right).$$

Notice that in the above formula the population size appears only in the step size and we may expect the behavior of two Wright-Fisher processes with different but large enough population sizes to be similar in the transformed scale. This might motivate us to consider the conditional distribution of $\theta_{\tau+\epsilon}$ given θ_τ for arbitrary $\epsilon \geq 0$, which would be given by the diffusion approximation

$$\theta_{\tau+\epsilon} \mid \theta_\tau \approx N\left(\theta_\tau, \epsilon\theta_\tau(1-\theta_\tau)\right). \tag{5.2}$$

The approximation given by Equation (5.2) is good if ϵ small and θ_τ is not close to the boundaries. If ϵ is large, then the Gaussian distribution is a poor approximation since the frequencies are constrained to the interval $[0, 1]$, but the Gaussian distribution has the whole real line as support. For example consider that $\theta_\tau = 0.5$. Then the probability mass of the approximation (5.2) which lies outside the interval $[0, 1]$ is 0.025 if $\epsilon = 0.2$, 0.16 if $\epsilon = 0.5$ and 0.32 if $\epsilon = 1$. Another problem associated with the direct use of the Gaussian approximation is that the variance is dependent on the frequency θ_τ. This has the effect that we would get a different result if we first computed the conditional distributions $\theta_{\tau+\epsilon/2} \mid \theta_\tau$ and $\theta_{\tau+\epsilon} \mid \theta_{\tau+\epsilon/2}$ and then marginalizing out $\theta_{\tau+\epsilon/2}$, instead of using (5.2).

To reduce the inhomogeneity of the variance, different transformations of the frequencies have been considered. For example, the transformation $\sin^{-1}\theta$ stabilizes the variance for all but extreme frequencies and the resulting variate has variance approximately $\epsilon/4$ over time period of ϵ (Thompson, 1975; Felsenstein, 1981). Another approach to improve the approximation was suggested by Nicholson et al. (2002), who used the Gaussian approximation with atoms placed on the boundaries to reflect fixation. Their objective was to compute divergence times of multiple populations from a single ancestral population. This can be viewed as learning the branch lengths of a star tree, which is a multifurcating phylogenetic tree where all the populations have simultaneously diverged from a common ancestral population.

An alternative approximation known as Balding-Nichols model is obtained by using a Beta distribution with same mean and variance

$$\theta_{\tau+\epsilon} \mid \theta_\tau \approx Beta\left(\theta_\tau, \epsilon\theta_\tau(1-\theta_\tau)\right), \tag{5.3}$$

where $\phi = (1-\epsilon)/\epsilon$ (Balding and Nichols, 1995, 1997; Rannala and Hartigan, 1996). It was originally developed in the context of equilibrium under migration and drift in island populations. The model and its multidimensional extension is widely used in learning population structure to introduce correlation in the allele frequencies between population (Falush, Stephens and Pritchard, 2003; Guillot, 2008; Gaggiotti and Foll, 2010). Similar to Nicholson et al. (2002) a star tree with unknown branch lengths is assumed in these methods to describe the evolutionary history of the populations.

It should be noted that the ϵ in the approximations (5.2) and (5.3) does not linearly correspond to s/N for any number of generations of s. Instead, it is related to the F_{ST} coefficient described earlier, see Nicholson et al. (2002) for details. If we would like to estimate the number of generations s that would correspond to ϵ, we would need to use the formula

$$s = N(1 - e^{-\epsilon}),$$

which is nearly linear for small values of ϵ.

In article (IV), we use the approximation (5.3) to model the change in allele frequencies in the history of populations described by the tree T. This possibility was mentioned by Zhang (2008), but deemed computationally too expensive. We model the allele frequencies in each node c of the tree conditional on the frequency in its parent $pa(c)$ with (5.3) using the length of the branch between c and $pa(c)$ as the parameter ϵ. An important advantage of the Balding Nichols model is that the allele frequencies corresponding to the observed populations can be analytically integrated out, similarly as in the marginal probability of a locus (4.4) used for learning the population structure described in the previous section. The inference about T is conducted in a full Bayesian framework in which the unknown tree T and the allele frequencies in the root node are given prior distributions. To manage the inference in the complex model, we utilize Laplace's approximation (3.3) for the allele frequencies, the AMIS algorithm (3.9) for the branch lengths of the tree and a greedy search algorithm in the space of tree topologies.

5.3 Generalizations

The discussion of the previous subsection considered the simple Wright-Fisher model with two selectively neutral alleles and no mutation. This model is suitable for analyzing SNP data as was done in article (IV), but is inadequate for many other genetic markers with multiple alleles or mutation. We now extend the basic model and consider approaches to approximate the extended models. It turns out that the approximations will be increasingly difficult to obtain as more complexity is added to the model and their computational complexity increases accordingly. However, it should be noted that exact coalescent methods in the spirit of those developed by RoyChoudhury, Felsenstein and Thompson (2008) and Bryant et al. (2012) are practically intractable for loci with multiple alleles.

Mutation First we introduce mutation to the biallelic Wright-Fisher model. The individuals of generation $t+1$ are obtained as a random sample from the previous generation t, but in addition to that a proportion u of the A alleles mutate to a and proportion v of the a alleles mutate to A. The conditional distribution of X_{t+1} given X_t is still binomial, but with a different parameter

$$X_{t+1} \mid X_t \sim Binomial(N, \psi_t), \tag{5.4}$$

where $\psi_t = N^{-1}((1-u)X_t + v(N - X_t))$. If both u and v are positive, then neither of the two alleles can go to fixation, but instead X_t will fluctuate around $Nu/(u+v)$. The distribution of the X_t given X_0 converges to a stationary distribution as t goes to infinity. The stationary distribution for the relative frequency θ_τ is approximately given by the Beta distribution

$$Beta(Nv, Nu), \tag{5.5}$$

with the approximation being exact in the infinite population limit (Wright, 1931).

For inferring the population history of several populations, we would like to compute the conditional distribution of $\theta_{\tau+\epsilon}$ given θ_τ for values ϵ which are too small for the process to reach stationarity. While we could approximate the distribution in the same spirit as in (5.2) by basing the approximation on the one generation distribution, it would be less accurate as mutation alters the process in such way that as functions of ϵ the mean and variance of $\theta_{\tau+\epsilon}$ are not closely linear to the one generation values.

We can, however, compute the mean and variance functions explicitly. As a slight abuse of notation, we first consider θ_t, the frequency of A in generation t, and then move to the continuous scale with θ_τ. We also drop the condition on θ_0 from the following formulas to simplify the notation. The expectation of θ_t given θ_0 can be computed using the law of iterated expectations

$$\begin{aligned} E(\theta_t) &= E(E(\theta_t \mid \theta_{t-1})) \\ &= E\left((1-u)\theta_{t-1} + v(1-\theta_{t-1})\right) \\ &= (1-u-v)E(\theta_{t-1}) + v \\ &= (1-u-v)^t\theta_0 + \sum_{i=0}^{t-1}(1-u-v)^i v. \end{aligned}$$

Now with large N and denoting $m_1 = uN$ and $m_2 = vN$

$$(1-u-v)^t = \left(1 - \frac{m_1+m_2}{N}\right)^{\tau N} \approx e^{-(m_1+m_2)\tau}.$$

The sum term can be approximated as an integral

$$\begin{aligned} \sum_{i=0}^{t-1}(1-u-v)^i v &\approx m_2 \int_0^\tau e^{-(m_1+m_2)\gamma} d\gamma \\ &= \frac{m_2}{m_1+m_2}\left(1 - e^{-(m_1+m_2)\tau}\right), \end{aligned}$$

giving

$$E(\theta_\tau) \approx e^{-(m_1+m_2)\tau}\theta_0 + \frac{m_2}{m_1+m_2}\left(1 - e^{-(m_1+m_2)\tau}\right). \tag{5.6}$$

Now we turn to the variance function, which we compute similarly using the law of total variance. We make the simplifying approximation that the variance of θ_{t+1} given θ_t is given by $N^{-1}\theta_t(1-\theta_t)$ instead of $N^{-1}\psi_t(1-\psi_t)$ as implied by Equation (5.4). This does not affect the results in the infinite population limit, but makes the derivations much more readable. The variance can be calculated as

$$\begin{aligned}
Var(\theta_t) &= E(Var(\theta_t \mid \theta_{t-1})) + Var(E(\theta_t \mid \theta_{t-1})) \\
&= \frac{1}{N} E(\theta_{t-1}(1-\theta_{t-1})) + Var((1-u-v)\theta_{t-1} + v) \\
&= \frac{1}{N}(E(\theta_{t-1}) - Var(\theta_{t-1}) - E(\theta_{t-1})^2) + (1-u-v)^2 Var(\theta_{t-1}) \\
&= \frac{1}{N} E(\theta_{t-1})(1-E(\theta_{t-1})) + \left((1-u-v)^2 - \frac{1}{N}\right) Var(\theta_{t-1}) \\
&= \frac{1}{N}\sum_{i=0}^{t-1}\left((1-u-v)^2 - \frac{1}{N}\right)^{t-1-i} E(\theta_{t-1})(1-E(\theta_{t-1})). \qquad (5.7)
\end{aligned}$$

With $\gamma = i/N$ the first term inside the sum can be approximated as

$$\begin{aligned}
\left((1-u-v)^2 - \frac{1}{N}\right)^{t-1-i} &= \left(1 - 2(u+v) + (u+v)^2 - \frac{1}{N}\right)^{t-1-i} \\
&= \left(1 - \frac{2(m_1+m_2)+1}{N} - \frac{(m_1+m_2)^2}{N^2}\right)^{t-1-i} \\
&\approx e^{-(2(m_1+m_2)+1)(\tau-\gamma)}.
\end{aligned}$$

We can now approximate the sum (5.7) as an integral

$$\begin{aligned}
Var(\theta_\tau) &\approx \int_0^\tau e^{-(2(m_1+m_2)+1)(\tau-\gamma)} E(\theta_\gamma)(1-E(\theta_\gamma))d\gamma \\
&= e^{-(2(m_1+m_2)+1)\tau}\int_0^\tau e^{(2(m_1+m_2)+1)\gamma} E(\theta_\gamma)(1-E(\theta_\gamma))d\gamma \qquad (5.8)
\end{aligned}$$

Before solving this integral, we rearrange

$$\begin{aligned}
E(\theta_\gamma)(1-E(\theta_\gamma)) = &\left(e^{-(m_1+m_2)\gamma}\theta_0 + \frac{m_2}{m_1+m_2}\left(1-e^{-(m_1+m_2)\gamma}\right)\right) \times \\
&\left(1 - e^{-(m_1+m_2)\gamma}\theta_0 - \frac{m_2}{m_1+m_2}\left(1-e^{-(m_1+m_2)\gamma}\right)\right) \\
=&e^{-(m_1+m_2)\gamma}\theta_0 + \frac{m_2}{m_1+m_2}\left(1-e^{-(m_1+m_2)\gamma}\right) - e^{-2(m_1+m_2)\gamma}\theta_0^2 \\
&- 2e^{-(m_1+m_2)\gamma}\theta_0 \frac{m_2}{m_1+m_2}\left(1-e^{-(m_1+m_2)\gamma}\right) \\
&- \left(\frac{m_2}{m_1+m_2}\right)^2\left(1-e^{-(m_1+m_2)\gamma}\right)^2 \\
=&\left(\left(\theta_0 + \frac{m_2}{m_1+m_2}\right)^2 - \frac{m_1 m_2}{(m_1+m_2)^2}\right)e^{-(m_1+m_2)} \\
&- \left(\theta_0 - \frac{m_2}{m_1+m_2}\right)^2 e^{-2(m_1+m_2)\gamma} + \frac{m_1 m_2}{(m_1+m_2)^2}.
\end{aligned}$$

Since for any real a, b

$$\int_0^b e^{ax}dx = a^{-1}(e^{ab}-1),$$

the variance (5.8) becomes

$$
\begin{aligned}
&Var(\theta_\tau) \approx \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5.9)\\
&e^{-(2(m_1+m_2)+1)\tau}\left(\left(\left(\theta_0+\frac{m_2}{m_1+m_2}\right)^2-\frac{m_1m_2}{(m_1+m_2)^2}\right)\frac{1}{m_1+m_2+1}\left(e^{(m_1+m_2+1)\tau}-1\right)\right.\\
&\left.-\left(\theta_0-\frac{m_2}{m_1+m_2}\right)^2(e^\tau-1)+\frac{m_1m_2}{(m_1+m_2)^2}\frac{1}{2(m_1+m_2)+1}\left(e^{(2(m_1+m_2)+1)\tau}-1\right)\right).
\end{aligned}
$$

The conditional distribution of θ_τ given θ_0 can now be now approximated using a Gaussian or Beta distribution with expectation and variance given by Equations (5.6) and (5.9), respectively. The parameters of the Beta distribution should be

$$
\begin{aligned}
&\left(\frac{E(\theta_\tau)(1-E(\theta_\tau))}{Var(\theta_\tau)}-1\right)E(\theta_\tau) \text{ and}\\
&\left(\frac{E(\theta_\tau)(1-E(\theta_\tau))}{Var(\theta_\tau)}-1\right)E((1-\theta_\tau)).
\end{aligned}
$$

If the scaled mutation rates m_1 and m_2 are big enough and neither dominates the other, the process should stay clearly away from the boundaries. It could be expected that a Beta distribution would be a better approximation to the process with mutation, since the stationary distribution (5.5) is also a Beta distribution, although we have not studied this.

Multiple alleles Next we consider an extension of the basic model for a locus with r alleles $a_1, \dots, a_r$. At generation t there are $X_{t,i}$ copies of allele of type a_i, $i = 1, ..., r$. Denote by $\theta_{t,i} = X_{t,i}/N$ the relative frequency of type a_i alleles at generation t. If mutation is not present the Wright-Fisher model generalizes straightforward to multiple alleles. The individuals of generation $t+1$ are still obtained as a random sample with replacement from the previous generation t, and the allele counts $X_{i,t}$, $i = 1, \dots r$, have multinomial instead of binomial distributions. Correspondingly, the Beta distribution of the Balding-Nichols model (5.3) is replaced by a Dirichlet distribution for the allele frequencies $\theta_{i,\tau}$.

Computation with multiple alleles poses a much more difficult problem compared to the biallelic situation. Specifically, Laplace's approximation which is used to marginalize out the allele frequencies in (IV) , has a much poorer performance. To see this, one should note that Laplace's approximation is based on approximating the full conditional distribution of the frequencies with a Gaussian distribution. With tens or possibly hundreds of frequency parameters constrained to sum up to 1, their full conditional distribution is bound to be far from a Gaussian.

However, the main limitation of the multiple allele model without mutation is that such loci often have mutation rates which can not be ignored in the same way as in the biallelic case. For example, microsatellites are known to have much higher mutation rates than for example SNPs (Ellegren, 2004). Similarly, gene sequences such as those obtained from Multi Locus Sequence Typing (MLST, Maiden et al., 1998) and analyzed in article (V) are known to mutate frequently.

General case We now develop results for a general Wright-Fisher model with multiple alleles and mutation. Let $u_{i,j}$ be the proportion of type j alleles that mutate to type i, with $u_{i,i}$ indicating the proportion of alleles of type i that do not mutate. This implies that $\sum_{i=1}^{t} u_{i,j} = 1$. The allele frequencies of generation $t+1$ have now distribution

$$\theta_{t+1} \mid \theta_t \sim Multinomial(N, U\theta_t), \tag{5.10}$$

where θ_s is the vector of allele frequencies $\theta_{s,i}$ at generation s and U is a matrix with elements $u_{i,j}$. As an example, the biallelic model with mutation (5.4) has the mutation matrix

$$U = \begin{bmatrix} 1-u & v \\ u & 1-v \end{bmatrix}. \tag{5.11}$$

The expectation and variance for θ_t given θ_0 could in principle be calculated similarly as in the biallelic case. The expectation would be following the same arguments as earlier

$$\begin{aligned} E(\theta_t) &= E(E(\theta_t \mid \theta_{t-1})) \\ &= E(U\theta_{t-1}) = U^t\theta_0. \end{aligned}$$

Computation of this expectation would require evaluation of the tth power of the matrix U, for which analytical formulas do not exist in the general case. A possible solution for some cases could be obtained from an eigenvalue decomposition of the matrix U, if this was available. Then it would only be necessary to raise scalars to the required power, instead of the whole matrix.

In article (V) we developed theory for Wright-Fisher model with a truncated version of the infinite alleles model (Kimura and Crow, 1964). In the infinite alleles model every mutation creates a new allele type, and the truncated version is obtained by following only $r-1$ distinct allele types and letting the rth allele type correspond to other alleles. The model can be represented in the above form with the elements of the mutation matrix U given by

$$u_{i,j} = \begin{cases} 1-u & \text{if } i=j, 1 \le i \le r \\ 1 & \text{if } i=j=r \\ u & \text{if } i=r \text{ and } 1 \le j \le r \\ 0 & \text{otherwise.} \end{cases} \tag{5.12}$$

We developed a two-stage Beta-Dirichlet approximation, which had the same first two moments as the Wright-Fisher model. The Adaptive Metropolis algorithm was implemented to facilitate the inference with the model. The potential of the model was illustrated by analyzing the history of world-wide populations of *Streptococcus pneumoniae* using MLST data (Enright and Spratt, 1998).

6 Discussion

Several statistical models and methods have been developed in this work mainly to problems in population genetics. The proposed methods have been shown to produce accurate inferences under a wide variety of conditions. Each method is a result of multiple compromises between accuracy and computational efficiency and has been designed with a specific problem in mind. Consequently, the performance of the method might be inferior, when applied to another problem. Perhaps the clearest example of this is the joint classifier described in (III), which outperforms the traditionally used marginal classifier with sparse training data, but otherwise increases the computational cost substantially.

Another example is the use of sampling locations of individuals, which may be helpful when inferring the population structure from genetic data with only few markers as demonstrated in (I), but only adds complexity to the computations when analyzing huge data sets such as the human SNP databases (Li et al., 2008). Similarly, the population allele frequency based approach developed in (IV) and (V) is well suited for the of the bacterial MLST databases, with large number of samples, but only a few loci. With other data sets, where the number of individuals is small, coalescent based approaches such as those suggested by Bryant et al. (2012) might be more sensible.

While the methods were developed for particular problems in population genetics, they can be used with slight modifications in a wide variety of problems. The clustering and classification methods discussed in Section 4 are applicable to many different fields of science. As an example of this, we have successfully applied them to a problem of crime linking in serial homicide from behavioral patterns (Salo et al., 2012).

The class of stochastic models, which includes the Wright-Fisher model used for inferring population history, is used in ecology, linguistics and physics outside population genetics (Blythe and McKane, 2007). The approximations developed in this work could possibly be utilized for some problems in these fields. In population genetics, the approximations could be used for more complicated evolutionary histories of populations than those, which can be described by a tree, similarly as in Cornuet et al. (2008). Additionally, selection could be incorporated to the Wright-Fisher model and, at least for biallelic loci, approximated using similar techniques as with mutation.

An important aspect of this work has been software development. When developing complex statistical methods for applied problems, their usefulness depends whether they may be easily used by other scientists. Developing dedicated and easy to use software implementing the methods provides researches in the applied fields an opportunity to utilize these methods in a straightforward manner.

The methods described in articles (I) and (II) and partially those in (III) are implemented in the computer program BAPS, which is freely available on the Internet [1]. The methods developed in articles (IV) and (V) are currently being implemented and a free software package will be made available in the near future.

[1] http://www.helsinki.fi/bsg/software/BAPS/

References

Andrieu C, Thoms J. 2008. A tutorial on adaptive MCMC. Statistics and Computing. 18:343–373.

Balding DJ. 2003. Likelihood-based inference for genetic correlation coefficients. Theoretical Population Biology. 63:221 – 230.

Balding DJ, Nichols RA. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica. 96:3 – 12.

Balding DJ, Nichols RA. 1997. Significant genetic correlations among caucasians at forensic DNA loci. Heredity. 78:583 – 589.

Beaumont MA. 2010. Approximate Bayesian computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics. 41:379–406.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics. 162:2025–2035.

Berger J. 2006. The case for objective Bayesian analysis. Bayesian Analysis. 1:385–402.

Bernardo J, Smith A. 1994. Bayesian theory. John Wiley & Sons Inc.

Bishop C. 2007. Pattern Recognition and Machine Learning. Springer, New York.

Blythe RA, McKane AJ. 2007. Stochastic models of evolution in genetics, ecology and linguistics. Journal of Statistical Mechanics: Theory and Experiment. P07018.

Box GEP. 1979. Robustness in the strategy of scientific model building. In: Launer R, Wilkinson G, editors, Robustness in Statistics, Academic Press, pp. 201–236.

Breiman L. 2001. Statistical modeling: The two cultures. Statistical Science. 16:199–215.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg N, RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. Molecular Biology and Evolution. First published March 14, 2012, doi: 10.1093/molbev/mss086.

Cavalli-Sforza L, Edwards A. 1967. Phylogenetic analysis: models and estimation procedures. Am J Hum Genet. 19:233–257.

Chapelle O, Schlkopf B, Zien A. 2006. Introduction to semi-supervised learning. In: Chapelle O, Schlkopf B, Zien A, editors, Semi-supervised learning, MIT Press.

Chen C, Durand E, Forbes F, Francois O. 2007. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Molecular Ecology Notes. 7:747–756.

Corander J, Gyllenberg M, Koski T. 2007. Random partition models and exchangeability for bayesian identification of population structure. Bulletin of Mathematical Biology. 69:797–815.

Corander J, Gyllenberg M, Koski T. 2009. Bayesian unsupervised classification framework based on stochastic partitions of data and a parallel search strategy. Advances in Data Analysis and Classification. 3:3–24.

Corander J, Waldmann P, Marttinen P, Sillanpää MJ. 2004. BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics. 20:2363–2369.

Corander J, Waldmann P, Sillanpää MJ. 2003. Bayesian analysis of genetic differentiation between populations. Genetics. 163:367–374.

Cornuet JM, Marin JM, Mira A, Robert C. 2012. Adaptive multiple importance sampling. Scandinavian Journal of Statistics. First published February 15, 2012, doi: 10.1111/j.1467-9469.2011.00756.x.

Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM, Balding DJ, Guillemaud T, Estoup A. 2008. Inferring population history with DIY ABC: a user-friendly approach to approximate bayesian computation. Bioinformatics. 24:2713–2719.

Cox DR. 1978. Foundations of statistical inference: The case for eclecticism. Australian Journal of Statistics. 20:43–59.

Dawson KJ, Belkhir K. 2001. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genetical Research. 78:59–77.

Dean A, Voss D. 1999. Design and analysis of experiments. Springer Verlag.

Degnan J, Rosenberg N. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution. 24:332–340.

Edwards A, Cavalli-Sforza L. 1964. Reconstruction of evolutionary trees. In: Heywood V, McNeill J, editors, Phenetic and Phylogenetic Classification, Systematics Association Publ. No. 6, London, pp. 67–76.

Edwards AWF. 2003. R. A. Fisher: Twice professor of genetics: London and Cambridge, or 'a fairly well-known geneticist'. Journal of the Royal Statistical Society. Series D (The Statistician). 52:311–318.

Edwards S. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

Efron B. 1986. Why isn't everyone a Bayesian? The American Statistician. 40:1–5.

Efron B. 2005. Bayesians, frequentists, and scientists. Journal of the American Statistical Association. 100:1–5.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics. 5:435–445.

Enright MC, Spratt BG. 1998. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology. 144:3049–3060.

Evans M, Swartz T. 2000. Approximating integrals via Monte Carlo and deterministic methods. New York: Oxford University Press.

Ewens W. 2004. Mathematical population genetics: theoretical introduction. New York: Springer Verlag, second edition.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics. 164:1567–1587.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet. 25:471.

Felsenstein J. 1981. Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. Evolution. 35:1229–1242.

Felsenstein J. 2004. Inferring Phylogenies. Sunderland, MA: Sinauer Associates.

Fienberg S. 2006. When did Bayesian inference become Bayesian. Bayesian analysis. 1:1–40.

Fienberg SE. 1992. A brief history of statistics in three and one-half chapters: A review essay. Statistical Science. 7:208–225.

Fisher R. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh. 52:399–433.

Fisher R. 1922. On the dominance ratio. Proceedings of the Royal Society of Edinburgh. 42:321–341.

Fisher R. 1930. The genetical theory of natural selection. Clarendon Press.

Fisher R. 1955. Statistical methods and scientific induction. Journal of the Royal Statistical Society. Series B (Methodological). 17:69–78.

Fitelson B, Thomason N. 2008. Bayesians sometimes cannot ignore even very implausible theories (even ones that have not yet been thought of). The Australasian Journal of Logic. 6:25–36.

Francois O, Ancelet S, Guillot G. 2006. Bayesian clustering using hidden markov random fields in spatial population genetics. Genetics. 174:805–816.

Francois O, Durand E. 2010. Spatially explicit Bayesian clustering models in population genetics. Molecular Ecology Resources. 10:773–784.

Frantz AC, Cellina S, Krier A, Schley L, Burke T. 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? Journal of Applied Ecology. 46:493–505.

Gaggiotti OE, Foll M. 2010. Quantifying population structure using the F-model. Molecular Ecology Resources. 10:821–830.

Garthwaite PH, Kadane JB, O'Hagan A. 2005. Statistical methods for eliciting probability distributions. Journal of the American Statistical Association. 100:680–701.

Gelfand AE, Smith AFM. 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association. 85:398–409.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. Bayesian Data Analysis. Boca Raton: Chapman and Hall/CRC, second edition.

Gelman A, Meng X, Stern H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica. 6:733–759.

Gelman A, Shalizi CR. 2012. Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology. First published February 24, 2012, doi: 10.1111/j.2044-8317.2011.02037.x.

Goldstein M. 2006. Subjective Bayesian analysis: principles and practice. Bayesian Analysis. 1:403–420.

Good IJ. 1959. Kinds of probability. Science. 129:443–447.

Good IJ. 1971. 46656 varieties of Bayesians. The American Statistician. 25:56–63.

Green PJ. 2003. Notes on the life and work of R. A. Fisher. Journal of the Royal Statistical Society. Series D (The Statistician). 52:299–301.

Guillot G. 2008. Inference of structure in subdivided populations at low levels of genetic differentiation–the correlated allele frequencies model revisited. Bioinformatics. 24:2222–2228.

Guillot G, Estoup A, Mortier F, Cosson JF. 2005. A spatial statistical model for landscape genetics. Genetics. 170:1261–1280.

Guillot G, Leblois R, Coulon A, Frantz AC. 2009. Statistical methods in spatial genetics. Molecular Ecology. 18:4734–4756.

Haario H, Saksman E, Tamminen J. 2001. An adaptive Metropolis algorithm. Bernoulli. 7:223–242.

Healy MJR. 2003. R. A. Fisher the statistician. Journal of the Royal Statistical Society. Series D (The Statistician). 52:303–310.

Hedrick PW. 2001. Conservation genetics: where are we now? Trends in Ecology & Evolution. 16:629 – 636.

Heggernes P. 2006. Minimal triangulations of graphs: A survey. Discrete Mathematics. 306:297 – 317.

Hein J, Schierup MH, Wiuf C. 2005. Gene Genealogies, Variation and Evolution. New York: Oxford University Press.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution. 27:570–580.

Jain AK, Murty MN, Flynn PJ. 1999. Data clustering: a review. ACM Computing Surveys. 31:264–323.

Jaynes E. 2003. Probability theory: the logic of science. Cambridge University Press.

Kass RE. 2011. Statistical inference: The big picture. Statist. Sci. 26:1–9.

Kass RE, Wasserman L. 1996. The selection of prior distributions by formal rules. Journal of the American Statistical Association. 91:1343–1370.

Kimura M, Crow J. 1964. The number of alleles that can be maintained in a finite population. Genetics. 49:725–738.

Kreveld M, Overmars M, Schwarzkopf O, Berg M, Schwartskopf O. 1997. Computational geometry: Algorithms and applications. Springer Verlag.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology. 56:17–24.

Lauritzen S. 1996. Graphical models. Oxford University Press, Oxford.

Li J, Absher D, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 319:1100–1104.

Lindley DV. 2000. The philosophy of statistics. Journal of the Royal Statistical Society. Series D (The Statistician). 49:293–337.

Liu L, Pearl DK, Brumfield RT, Edwards SV. 2008. Estimating species trees using multiple-allele dna sequence data. Evolution. 62:2080–2091.

Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV. 2009. Coalescent methods for estimating phylogenetic trees. Molecular Phylogenetics and Evolution. 53:320 – 328.

Lopes JS, Balding D, Beaumont MA. 2009. PopABC: a program to infer historical demographic parameters. Bioinformatics. 25:2747–2749.

Maiden MCJ, Bygraves JA, Feil E, et al. (13 co-authors). 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences. 95:3140–3145.

Marchini J, Cardon L, Phillips M, Donnelly P. 2004. The effects of human population structure on large genetic association studies. Nature genetics. 36:512–517.

Marin JM, Pudlo P, Robert C, Ryder R. 2011. Approximate Bayesian computational methods. Statistics and Computing. First published October 21, 2011, doi: 10.1007/s11222-011-9288-2.

Morin PA, Luikart G, Wayne RK, the SNP workshop group. 2004. SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution. 19:208 – 216.

Mörters P, Peres Y. 2010. Brownian Motion. Cambridge University Press.

Nicholson G, Smith AV, Jónsson F, Gústafsson O, Stefánsson K, Donnelly P. 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society. Series B. 64:695–715.

Nielsen R, Mountain J, Huelsenbeck J, Slatkin M. 1998. Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. Evolution. 2:669–677.

Novembre J, Johnson T, Bryc K, et al. (11 co-authors). 2008. Genes mirror geography within Europe. Nature. 456:98–101.

Owen A, Zhou Y. 2000. Safe and effective importance sampling. Journal of t. 95:135–143.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. Molecular Biology and Evolution. 5:568.

Patterson N, Price A, Reich D. 2006. Population structure and eigenanalysis. PLoS Genetics. 2:e190.

Pearse DE, Crandall KA. 2004. Beyond FST: Analysis of population genetic data for conservation. Conservation Genetics. 5:585–602.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Quintana FA. 2006. A predictive view of Bayesian clustering. Journal of Statistical Planning and Inference. 136:2407–2429.

Rannala B, Hartigan JA. 1996. Estimating gene flow in island populations. Genetical Research. 67:147–158.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164:1645–1656.

Robert C. 2007. The Bayesian choice: from decision-theoretic foundations to computational implementation. New York: Springer Verlag, second edition.

Robert C, Casella G. 2004. Monte Carlo statistical methods. New York: Springer, second edition.

Robert C, Casella G. 2011. A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. Statistical Science. 26:102–115.

Roberts G, Rosenthal J. 2001. Optimal scaling for various Metropolis-Hastings algorithms. Statistical Science. 16:351–367.

Roberts GO, Rosenthal JS. 2009. Examples of adaptive MCMC. Journal of Computational and Graphical Statistics. 18:349–367.

Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, Zhivotovsky L, Feldmann M. 2002. Genetic structure of human populations. Science. 298:2381–2385.

Rota GC. 1964. The number of partitions of a set. The American Mathematical Monthly. 71:498–504.

RoyChoudhury A, Felsenstein J, Thompson EA. 2008. A two-stage pruning algorithm for likelihood computation for a population tree. Genetics. 180:1095–1105.

Rue H, Martino S, Chopin N. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society. Series B. 71:319–392.

Safner T, Miller MP, McRae BH, Fortin MJ, Manel S. 2011. Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. International Journal of Molecular Sciences. 12:865–889.

Salo B, Sirén J, Corander J, Zappala A, Bosco D, Mokros A, Santtila P. 2012. Using Bayes' theorem in behaviourial crime linking of serial homicide. Legal and Criminological Psychology. First published February 3, 2012, doi: 10.1111/j.2044-8333.2011.02043.x.

Savage LJ. 1976. On rereading R. A. Fisher. The Annals of Statistics. 4:441–500.

Senn S. 2011. You may believe you are a Bayesian but you are probably wrong. Rationality, Markets and Morals. 2:48–66.

Stanley RP. 2012. Enumerative Combinatorics, Volume 1. Cambridge University Press, second edition.

Thompson E. 1975. Human evolutionary trees. Cambridge: Cambridge University Press.

Tierney L, Kadane J. 1986. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association. 81:82–86.

Waples RS, Gaggiotti O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular Ecology. 15:1419–1439.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L. 2010. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics. 11:116.

Wilson IJ, Weale ME, Balding DJ. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. Journal of the Royal Statistical Society: Series A. 166:155–201.

Wright S. 1931. Evolution in Mendelian populations. Genetics. 16:97–159.

Wright S. 1943. Isolation by distance. Genetics. 28:114.

Wright S. 1949. The genetical structure of populations. Annals of Human Genetics. 15:323–354.

Zhang Y. 2008. Tree-guided Bayesian inference of population structures. Bioinformatics. 24:965–971.