Bachelor Thesis

Using Approximate Bayesian Computation to Infer the Number of Populations from SNP Genotype Data

Supervisors: Manfred Opper,
Olivier Françios,
Michael Blum

Fabian Bergmann, 372918

Pages: 34

Submission Date: April 15, 2019

Contents

1 Introduction

A central objective in population genetics is to eveluate population structure. Genetical differences of member individuals are analysed to detect systematic genetic similarities and dissamilarites that could indicate the presence of various subpopulations. The biological definition of a population generally follows the lines of: a population is a "group of organisms of the same species living within a sufficiently restricted geographical area so that any member can potentially mate with any other member of the oppisite sex". Subsequently however, as only a theoretical setting is assumed without any further context, like geography, social hierarchy etc., a population will describe a group of individuals that are associated with one another because their genetic information is sufficiently homogenous to be discriminated as a group from other individuals of the species. An adpted notion of a subpopulation will not be necessary since the described understanding of a population overlaps with that of a subpopulation (each detectable subpopulation would be a new population), also no further hierarchical structure is present in the used theoretical framework that would make further granulation necessary.
An allele is a variant of a gene found at a specific location on a chromosome (locus), and consequently alleles are responsible for the appearance of genetic variation in a species. An allele frequency describes the probability $p(a|k)$ that an individual from population k has the allele a. Since individuals in a population possess similar genotypes, so some alleles are encountered more frequently in these individuals than others and what therefore sets them apart, the allele frequencies sufficiently summarise the population.
The ampleness of information usually found in a genom poses a challenge dimensionality wise if all of it were to be captured. For tasks that solely involve the analysis of genetic variation between individuals and groups of individuals it suffices to rely on the evaluation of a limited amount of genes. Mostly genese are chosen that are known to be subject to genetic variation, these particular genes are called markers. **introduce snp markers now?**
The usual approach to investigate population structure follows the construction of a statistical model according to biological hypotheses which is then evaluated with respect to how well it fits collected genotype data. The goodness of fit could be evaluated by some statistical means such as hypothesis testing, or by comparison with other models that are based on different biological hypotheses. Many models thereby employ hyperparameters that demand the specifiaction of their values. The hyperparameter values have to be chosen carefully, since for example by merely picking the parameter values that maximise the likelihood of the model could lead to unrealistic values (especially if model assumptions are faulty). In addition, the comparison of models becomes at best tricky if they both propose different values for the same hyperparameters. A tool that estimates realistic hyperparameter values model independent would therfore be of great facilitation to the inference and evaluation process.
One particular hyperparameter that is frequently required in population genetics is the specifiaction of how many populations are present in the to be analysed data, however until now no automated solution exists that satisfyingly and reliably predicts a realistic value. Throughout the following pages an approximate bayesian computation method is presented that intends to deliver a reasonable and fairly neutral estimate for the number of populations in given genotype data.
It is fundamentally based on the insight that the behaviour of the spectrum of a covariance matrix is closely linked to the clustering structure found in the respective data. In summary, for k distinguishable (linearly independnt) clusters $k-1$ eigenvalues can be observed that are significantly larger than the rest. This knowledge is consequently exploited to construct a summary statistics, necessary for a lower dimensional computation, solely based on the eigenvalues

of the data covariance matrix.
In the past methods have been attempted that rely on the same insight, however these mostly use heuristic approaches that intend to identify the jump transition from the lowest large significant eigenvalue to the smaller ones. The results are at best mixed, particularly for data that is noisy and where data structure is not easily detectable. In contrast, the subsequently proposed method attempts to generalise the connection between eigenvalues of the covariance matrix and number of populations by employing a supervised learning approach with boosted decision trees.
As for any supervised learning method, a sufficient amount of labeled data will be required to train the model. Constructing an extensive training data set with real genotype data is highly intricate and resource demanding, furthermore, even if an ample amount of consistent genotype data were available, labeling it with the ground truth would be impossible. Of course no method exists that would deliver an unambigous estimate for the number of populations, mostly for the reason that for many real world data sets no unambigous answer can be made. Real genotype data often exhibits a more complex structure. Populations as already indicated are subject to a hierarchical population structure, such that the number of populations detected will depend on the assumptions about what is a population.
The biological definition of a population generally follows the lines of: a population is a "group of organisms of the same species living within a sufficiently restricted geographical area so that any member can potentially mate with any other member of the oppisite sex". However, transfering this definition into a theoretical setting without any further real world context, like geography, social hierarchy, etc., is cumbersome if not even impossible for the porpuses of generalistion. Therefore, subsequently a population will describe a substantially large enough group of individuals from the same species that are associated with one another because their genetic information is sufficiently homogenous for them to be discriminated as a whole from other specie individuals.
As elicited, the inference of the number of populations is always succumbed by some assumption about the nature of a population. Nevertheless the model assumptions, learned in a supervised learning approach through the training data, is kept as universal and neutral as possible. The training data, which is supposed to resemble real SNP genotype data samples, is synthetically generated by a generative model, which can simulate a variety of population structure scenarios based on well accepted biological assumptions. Apart from that, the training data is generated as assumptionless as possible.
At the core of the generative model lies the idea that a new population splits of from an existing population if for a substantial amount of time a group of individuals of the existing population only interbreed with one another. In a real world context a reason for the split off could for example be that the group decides to settle at different geographic location and thus isolates itself. The general observed genotype of a population is always changed by a force called random genetic drift. If two groups of individuals from the same species do not exchange genetic information by mating, both groups exhibit a different genetic drift such that their genotypes become dissimilar enough to classify them as two seperate populations. **Furthermore, the use of summary statistics decreases model dependency and improves generality???**
assumptions that are learned by the supervised learning model are kept as universal and neutral as possible to guarentee the transferbility of its predictions. **add to discussion, what definition of a population did the model gather** In this sense an assumption free inference of the number of populations is anyways impossible.
Fundamental to the method is the insight that Thereby Usually for this purpose the validity of a certain model is evaluated by analysing the underlying genotype data of the organisms.

Models are often susceptible to the realistic choice of hyperparameters, such that for example the likelihood of a model with a certain hyper parameter greater is than for a much more realistc hyper parameter. Since for many models the declaration of the number of populations in the given data is crucial to infer the remaining parameters, a successful and stable technique for determining this parameter would lead to a significant facilitation for the inference process.
The eigenvalues are therefore a fruitful choice as summary statistics, in order to decrease the dimensionality in the calculation. The likelihood function for this problem is elusive, therefore a gradient boosting tequnique with decision trees wil be employed to bypass its explicit computation.
Gradient Boosting is a supervised machine learning method, thus it requires a significant amount of data for training and testing. Real world genotype data however is too extensiv to acquire in such a magnitude as is necessary, therefore an artificial data set is simulated and used. The generation of the artificial data follows commonly accepted theories for the simulation of population structure.

1.1 Problem

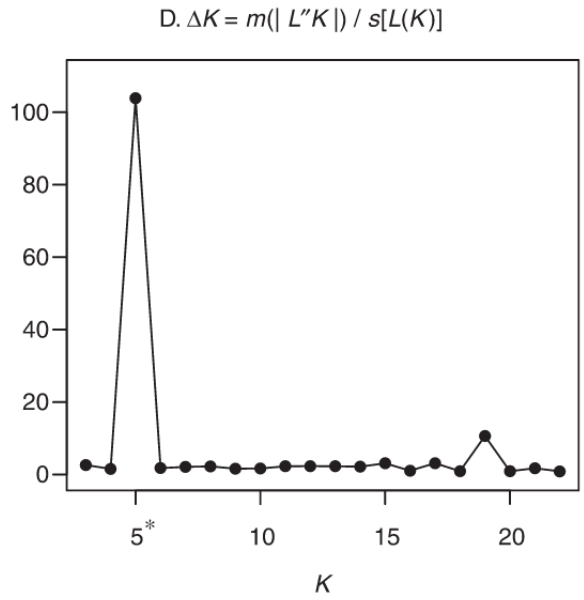
An allele is a variant of a gene found at a specific location on a chromosome (locus). Its existence originates from a random mutation in a gene. Alleles found at the same locus are generally considered to be interchangeable which allows for the random passing on of an allele from the parents to their offspring. For example, in a diploid species (as humans are) an individual possesses pairs of homologous chromosomes, meaning they possess two alleles (not necessarily different) of the same gene. Only one of the two alleles is randomly passed on to the child, the other is received by the child from the corresponding mating partner. Many other factors such as non-random mating, natural selection, linkage disequilibrium and so forth furtherly influence the successful spreading of an allele. As a result, alleles are responsible for the appearance of genetic variation in a species.
Since individuals in a population possess similar genotypes, so some alleles are encountered more frequently in these individuals than others and what therefore sets them apart, the allele frequencies sufficiently summarise the population. An allele frequency describes the probability $p(a|k)$ that an individual from population k has the allele a.
The ampleness of information usually found in a genom poses a challenge dimensionality wise if all of it were to be captured. For tasks that solely involve the analysis of genetic variation between individuals and groups of individuals it suffices to rely on the evaluation of a limited amount of genes. Mostly genese are chosen that are known to be subject to genetic variation, which are the main interest. These particular genes are called markers. Chosen as representation for the variation in the genotypes for the model are single-nucleotide-polymorphism (SNP) markers, which are the smallest possible variation for only a single base nucleotide is exchanged by another. The law of large numbers explains that only a sufficient finite amount of markers need to be determined to recognise the population structure.
With the established biological framework, the task can be formulated more specific: Given the SNP genotype data of individuals from the same species, organised in a matrix X where each row corresponds to an individual, the task is to infer how many populations K are present in the data X. If the markers sufficiently capture the genetic variation of the sampled individuals, it is expected for members of the same population to cluster together in the feature space spanned by by the parametarised alleles of the selected SNP markers. A cluster is a set of data points that are similar enough (by some means of evaluation) to be grouped together, which coenciedes with the given definition for a population. So for the K populations K different clusters should be observable in the feature space. Each cluster corresponding to a specific

population, where the position of the cluster should be determined by the allele frequencies of the respective population.
A population is a "group of organisms of the same species living within a sufficiently restricted geographical area so that any member can potentially mate with any other member of the oppisite sex" (Hartl, Clark, and Clark 1997). Especially due to genetic drift, the change of the allele frequencies in a population that occurs because of finite random sampling from the available gene pool, populations are destinguishable in their genetic information, although they might have split from a single population a reasonable amount of generations ago. For further information the reader is referred to (Hartl, Clark, and Clark 1997). Therefore, if a sufficient amount of genetic information is used to span the feature space, usually in the form of genetic variations found at genetic markers, individuals should cluster together with other individuals of the same population as their genetic data is more homogenous **add more reasoning???, law of large numbers???**. A cluster is a set of data points that are similar enough (by some means of evaluation) to be grouped together. Therefore, should the The number of populations K should accord to K clusters found in the feature space, so the problem simplifies to identifying the number of clusters found in the data matrix X. **give a definition for clusters???, measure of genetic distance???**

1.2 Approaches currently used

Clustering problems have been well invstigated and many methods for solving these problems have been proposed. Predominantly the methods can be apportioned among two groups.
One group compares the similarities between datapoints by the means of a distance measure. Structure is attempted to be recognized by evaluating the distance of every datapoint to every other datapoint. Examples of distance based clustering methods are centroid based clustering, hierarchical clustering, etc.
The other group contains methods that are based on statistical model assumptions. Every datapoint is considered to be a random draw of a probability distribution. The parameters of the distributions are inferred via common statistical methods, such as maximum likelihood or baysian methods.
For a long time model based approaches were favored and used for determining desired parameter values. A widely applied (Rosenberg, Pritchard, et al. 2002; Harter et al. 2004; Rosenberg, Burke, et al. 2001) implementation of a model based method is the software package STRUCTURE first developed by (Pritchard, Stephens, and Donnelly 2000; Falush, Stephens, and Pritchard 2003). In short, at core lies the modelling of K unobserved populations by assigning them specific allele frequencies, where it is assumed that the genotype of an individual is drawn according to its population allele frequencies. Furtherly the whole is embedded into a bayesian framework, where priors are set over the population membership of each individual and the allele frequencies of each population, in order to include scenario specific conditions like geography into the model and thus increasing model flexibility. Given some genotype data, the parameters of the model are inferred via a Markov Chain Monte Carlo (MCMC) algorithm with Gibbs sampling.
why not use mcmc for k? For the matter of determining the number of populations K present in the data the posterior $P(K|X)$ distribution is approximated for various K, where X is the given data. As only the maximum posterior of the selected K values is to be determined following reduced relation is received

$$P(K|X) \propto P(X|K)P(K)$$



(For further information on Bayes' theorem see section about Approximate Bayesian Computation)
The likelihood $P(X|K)$ is estimated by making a gaussian assumption about the distribution of the deviance D of the parameters model parameters conditioned on the data X. The assumption allows for determining the gaussian distribution by only calculating the mean and variance of D with the MCMC samples of the stationary distribution. **further information appendix** However, even Pritchard, Stephens, and Donnelly (2000) themselves describe this as a "dubious" assumption rendering the method as only a supportive argument for picking K.
Nonetheless STRUCTURE is widely used for inferring the number of populations, however in a different way than originally intended. Most often the decision of K follows a heuristic based on the likelihood $P(X|K)$, which is calculated with the implemented monte carlo algorithm in STRUCTURE. The heuristic most commonly applied, investigated and proposed by Evanno, Regnaut, and Goudet 2005, relies on the search for the largest jump in the second derivative of a composed loglikelihood function. The likelihood function $L(K)$ is calculated by averaging the loglikelihood of each MCMC step and further subtracting half of the variance off that mean **(a penalisation for instable models)**. The second derivative is defined as $L''(K|X) = L'(K+1) - L'(K)$ with of couse $L'(K) = L(K+1) - L(K)$. Several runs are completed and therefore also several results for $L(K)$ received. The selection criteria then is $\Delta K = \frac{m(L''(K))}{s(L(K))}$, where m is the mean and s the standard deviation.
The method offers reasonable results, even for hierarchichal clusters (the population structure in the highest hierarchy layer is the expected result), however it was conceived only as an adhoc solution strongly reliant on the results outputed by the STRUCTURE algorithm. Furthermore, there are indications that the method is biased towards lower Ks if not several exhaustive sampling with the MCMC is completed, which tends to be at least computationally very time comsuming as the dimension of genotype data has rapidly expanded with the introduction of more modern genom sequencing equipment.

A more recent approach involves the insight that cluster structure is also resembled in a structured form in the spectrum of the covariance matrix from the respective data. The connection

between the spectrum of a matrix and its clustering structure has been subject to research for a fair amount of time. It was first discovered in graph theory Donath and Hoffman 1973 Fiedler 1973 and later introduced into machine learning (Shi and Malik 2000; Meila and Shi 2001; Ng, Jordan, and Weiss 2002) for further information see Von Luxburg 2007. In general the relevant insight states that: suppose K clusters can be observed in the data matrix X (w.l.o.g. X is a square matrix), then the first $k-1$ eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{k-1}$ are significantly larger than the remaining eigenvalues, also the corresponding eigenvectors. The number of clusters can thus be inferred by examining the eigenvalues of the data.
the In the context of inferring the number of populations by exploiting the appearence of population structure in the eigenvalues was firstly applied by Patterson, Price, and Reich 2006. They used insights from random matrix theory, which state that (under various assumptions, like the covariance matrix has properties of a random Wishart matrix) the significant $K-1$ eigenvalues are approximately Tracy-Widom distributed. The Tracy-Widom distribution describes the distribution of the largest eigenvalue for varoius random gaussian matrix ensembles, like the Gaussian orthogonal ensemble ($\beta = 1$) to which the Wishart-matrix belongs. A statistical test is constructed where the eigenvalues are checked with a chosen p-value if there is substantial evidence for the eigenvalue to be Tracy-Widom distributed. **Performancewise, the method tends to overestimate the number of populations need source!!!.**

Further research has been made with random matrix theory (RMT) to concretise the behaviour of the first $k-1$ eigenevalues, including, under certain assumptions, a mathematicaly quantifieable threshhold that destinguishes significantly larger eigenvalues from lower ones. K. Bryc, W. Bryc, and Silverstein 2013, indicating further that it is only necessary to detect the jump point in the eigenvalues to infer the number of populations. Therefore, instead of relying on ad hoc heuristics or attempting to formulate xpecific mathematical methods with multiple underlying assumptions which in turn have trouble with real data, the use of a machine learning technique that generalises the connection between the number of populations present in genotype data and the correlating structure observed in the eigenvalues.

A natural approach for solving problems involving clustering, would be to use well established and commonly used methods, like maximising the likelihood of an expactation maximisation in combination with a model quality estimator such as the Bayesian Information Criterion to counter overfitting. In general bayesian approaches, sucha as maximising the likelihood with regard to a theoretical model of the population structure is always a possible approach, however makes the estimation of the number of populations highly dependable on the a-priori assumptions of the model for what determines mathematically a new population Falush, Stephens, and Pritchard 2003. Fitting a hierarchical tree model with bayesian methods has also been attempted Corander et al. 2004. Different assumptions from different models could lead to different estimations, which would undermine their comparability. Nevertheless, a maximum-likelihood approach was implemented in the software STRUCTURE Pritchard, Stephens, and Donnelly 2000 Falush, Stephens, and Pritchard 2003 and widely applied Rosenberg, Pritchard, et al. 2002 Harter et al. 2004 Rosenberg, Burke, et al. 2001. Furthermore, in some cases, especially for data that involves a high number of populations, a very distinctive maximum is not obtained, for the maximum-likelihood function tends to be smoother as higher values are examined **more explanation???**. Some approaches add further heuristics, such as also taking the second order rate of change of the likelihood function into consideration Evanno, Regnaut, and Goudet 2005, which however appears more like mending the performance of an approach that was solely conceived as preliminary remedy Pritchard, Stephens, and Donnelly 2000.

A more recent approach involves the insight that a cluster structure is also resembled in a structured form in the spectrum of the respective data matrix. The connection between the spectrum and a matrix was first discovered in graph theory Donath and Hoffman 1973 Fiedler 1973 and later introduced into machine learning Shi and Malik 2000 Meila and Shi 2001 Ng, Jordan, and Weiss 2002, for further information see Von Luxburg 2007. In general the relevant insight states that: suppose K clusters can be observed in the data matrix X (w.l.o.g. X is a square matrix), then the first $k-1$ eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{k-1}$ are significantly larger than the remaining eigenvalues, also the corresponding eigenvectors **span a subspace that approximates a simplex with the clusters as vertices???**. The number of clusters can thus be inferred by examining the eigenvalues of the data, firstly detected and applied in the context of population genetics by Patterson, Price, and Reich 2006.
examples
Approaches have been made with random matrix theory (RMT) to concretise the behaviour of the first $k-1$ eigenevalues Patterson, Price, and Reich 2006, including a mathematical threshhold that destinguishes significantly larger eigenvalues from lower ones K. Bryc, W. Bryc, and Silverstein 2013. Utilising insights from random matrix theory furtherly remains its beginnings, so far no well performing method based on RMT has been developed. **cite ???**

2 Biological Background

ADD MORE !!!

2.0.1 Key words

- **Chromosome:** A DNA molecule that encodes genetic information.
- **Gene:** A DNA (or RNA) sequence that specifies the structure of a particular functional molecule.
- **Locus:** A particular position on the chromosome, like the position of a specific gene.
- **Allele:** A variant form of a given gene. Different alleles can lead to distinct phenotypic traits.

3 Modelling

The subsequently presented model resembles the hierarchical model implemented in STRUCTURE. A first layer, referred to followingly as the F-layer, originates from Falush, Stephens, and Pritchard 2003 to introduce the possibility of adjusting the correlation of allele frequencies between populations. The F-layer is responsible for determining the allele frequencies of the populations. The second layer, from now on referred to as the admixture layer, stems from Pritchard, Stephens, and Donnelly 2000 and allows for individuals to have admixed genotypes. In other words, the frequencies of several populations are partially contributing to the pool from which the genotype of an admixed individual is sampled from. Thus, the admixture layer samples the genotypes of the individuals, whereby controling the proportions every population contributes to the genotype of an individual.
The genetic information to evaluate individuals is gathered from specific locations in the genom (loci) that can exhibit known genetic variants (alleles). These specific locations are called markers allow for an easier handling of the demensionality and complexity of a genom for certain

tasks, such as classefication. To detect properly the magnitude of difference between the genotypes of two individuals from the same species, a sufficient amount of loci that carry genetic variation have to be used. For modelling SNP variations will be used. The biological meaning of the loci generated by the following model for the task of inferring the number of populations is irrelevant, solely necessary is the fact that at each loci different alleles should exist. Each population expresses the alleles at different frequencies, such that, by the law of large numbers, the more loci are simulated the more seperable the populations should be from another and individuals easierly assignable to a population. From a modelling view, it is the different allele frequencies that define a population.

add snp is biallelic Alleles captured from SNP markers are biallelic, for the SNP markers locate a mutation in a single nucleotide base pair. That a base pair is effected by several mutations and these mutations assert themselves in the population is, because of the low chances of single nucleotide being affected by a mutation, very unlikely and therefore the possibilities of more than two alleles at an SNP marker is neglected. This allows for a notation simplification of the allele frequencies. Let subsequently p_{kl} determine the probability that an individual from population k has a mutation at locus l. Of course the existince of mutant variants is always relative to some reference genome from which over sufficiently long amount of time mutations formed and were able to gather some share in the allele frequencies of a population.

The allele frequencies of a population k at a locus l for an allele a will be denoted as $p_k(l_a)$. For simplification purposes each loci in the model can be interpreted as having only two alleles. The model assumes a point in time where **no different alleles existed??? or alleles where considered the same???** and overtime mutations introduced new alleles that could asser themselves. However the model does not distinguish between the new alleles, only if an individual carries a mutant variant or not (furthermore ploidity is also ignored). The notation therefore also simplifies for the modell to $p_k(l)$ (probability of having a mutant allele at locus l if from population k).

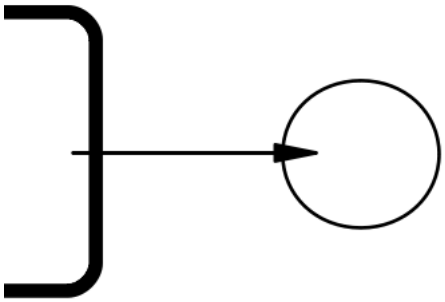
3.1 Further Model assumptions

The model assumes several properties:

- Hardy-Weinberg Equilibrium: Relevant to the modelling is that, wiht a Hardy-Weinberg Equilibrium the allele frequencies within a population do not change, the populations are "stationary".
- Linkage equilibrium: Loci are independent from one another.
- **No hierarchical clusters present, like families in a population???**
- **further assumptions ???**

3.2 F-Layer

further informations, reader can consult textbooks The motivation for introducing correlation between allele frequencies of populations is that poplutions often originate from a common ancestral population, so each population stems from the same starting allele frequencies. Reasons for the emergence of new populations are numerous, ranging from geographical divides to social structure impacting the genetic makeup, the component common to all is that a group



of individuals excludes themselves sufficiently from mating with the rest of the population and thus practices inbreeding. Schmetically in population genetics these scenarios of fragmentation are often abstracted as a set of K secluded islands to which some individuals of an ancestral population A migrated and interbreeding on each island occurs.
The formation of populations on the islands with distinct allele frequencies is due to a phenomenon called random genetic drift. From a simplified view point, excluding all the biological idiosyncrasies and only looking at the alleles at a single locus, random genetic drift can be seen as an urn model, where the alleles at the locus make up the balls that are drawn from the urn. The proportions of the balls should represent the allele frequencies of the population. The alleles that the offspring receives are sampled from the urn (with replacement). However, since the amount of individuals in the offspring is finite, the offspring will most likely not possess exactly the same allele frequencies as the parents did. This causes the allele frequencies of a population to vary until one allele achieves fixation. Fixation occurs when by chance only one specific allele is sampled from the urn, thus eliminating all its competing alleles.

Repeated over time and expanded to the continous case, genetic drift can be formulated mathematically as a stochastic process on the change of allele frequencies, where the total dominance of an allele is an absorbing state (**crow1970introduction**). Therefore, after a sufficient amount of time a decrease in genetic diversity will be observed in inbreeding populations as more alleles tend to fixation. Wright's F-statistic is a widely used measure for evaluating the divergence of allele frequencies of populations from the mean. It consists of a single value F_{ST} and is calculated as $F_{ST} = Var(p_a)/\overline{p_a}(1-\overline{p_a})$ **more explanation, fixation???**, where $Var(p_a)$ is the variance of the allele frequencies for allele a across the evaluated populations and $\overline{p_a}$ the mean. Once all alleles are fixed in (considering an infinite amount of populatoins), it is expected for a proportionate amount according to $\overline{p_a}$ to have fixed the allele a. The variance would consequently conclude to $\overline{p_a}(1-\overline{p_a})$ making the equation equal to one and indicating total fixation. If the variance is near zero all populations would still most likely have not diverged far from the overall starting allele frequency.

The model will employ an anlogous adoption of the F-statistic. The ancestral allele frequency p_{Al} acts as the mean from which all other frequencies originate. Instead of a single parameter F_{st}, a hyperparameter F_k for each population k is introduced that controls the magnitude of its divergence from the ancestral population.

The alleles frequencies for a population k are sampled from a beta distribution, with following parametrisation:

$$p_{kl} = beta(p_{Al}\frac{1-F_k}{F_k}, (1-p_{Al})\frac{1-F_k}{F_k}) \tag{1}$$

The choice of using a beta distribution is motivated by the fact that it is the stationary distribution that is received for the stochastic process if a counter force working against allele fixation is added, such as accounting for a steady stream of migrants from the ancestral population or further mutations that sustain an allele. **parameters derived by setting them equal to the moments, set variance and mean equal??? mention maximum entropy (balding2003likelihood)**
It is assumed that the loci are independant from one another, as a consequence factors like linkage disequilibrium can not be respected, however it allows for the independent sampling at each locus with the same F_k value.
Proceeding, the allele frequencies for a population k are joined together to a vector p_k and then merged with all other K populations to a matrix $\mathbf{F} = [p_1 p_2 \dots p_K]^T$ of size $K \times L$, where L is the number of loci. Each column of $\mathbf{F}$ gives the allele frequency for each population at a specific locus l.

The Distinct populations within a species form mainly due to random genetic drift when there is a fragmentation of the general population into subgroups present. Genetic random drift can be interpreted as a stochastic process where the allele frequencies of a gene change randomly over time. The first modelling atempts applied a markov chain **cite**, where a transition models the change of allele frequencies from one generation to the next. Further refinements have been added over time, such as converting the simplification of discrete populations to the continous case and adding biological idiosyncrasies **cite**. The random change in the stochastic process stems from the fact that a population has a finite amount of members, whose genetic makeup was "drawn" from a pool with probabilites according to the allele frequencies of the population at the time of the conception of the members, therefore it is likely that the allele frequencies in a population are not passed on exactly. Considering the law of large numbers inversely, the smaller a population is the faster genetic drift will make an impact. The allele frequencies change until one allele is able to assert to fixation, meaning no other variants are substantially left and only the one allele is passed on.

All K populations in the model emerged from a single ancestral population A. This is founded on the biological background that for various reasons such as migration, geography, climate, sub popualations form within a single population and as mating between subpopulations decreases or even ceases the populations are effected by distinct random genetic drifts, as a consequence their genetic makeups diverge from another until they can be considered distinct populations. Random mutations and natural selection, if environments are substantially different, in addition support the divergence. **Nonetheless, the fixation of most genes is mostly due to genetic drift???.** As certain alleles becom more dominant in populations the genetic variation declines and more individuals become homozygous (so less are heterozygous). Using this insight, a divergence statistics that is widely used measures the decrease in heterozygosity between the original population and the newly emerged subpopulation.
Let H_S denote the heterozygosity of the original population and H_T the heterozygosity of the subpopulation. Then

$$F_{ST} = \frac{H_S - H_T}{H_S}$$

gives a percentage by how much heterozygosity decreased.

what is the connection to the F-model??? how to get to $\frac{1-F^k}{F^k}$
This divergence measure, or F-statistic, is used analogously to parametarise the genetic drift from the ancestral population, like following citefalush2003inference:

$$Dir(p^A_{l_1}\frac{1-F^k}{F^k}, p^A_{l_2}\frac{1-F^k}{F^k}, \ldots, p^A_{l_{a_l}}\frac{1-F^k}{F^k}) \tag{2}$$

Where $p^A_{l_i}$ is the allele frequency from the ancestral population of the allele i, from a_l alleles, at locus l and F^k is the drift value of population k. At a very low F^k (like 0.05) the fraction $\frac{1-F^k}{F^k}$ is considerably above 1, thus the probability mass is mostly concentrated around ancestral allele frequencies. As the value of F^k increases:

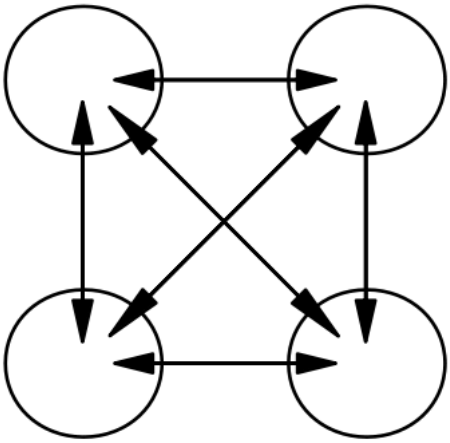
- for a moderate value (like 0.3) the probability mass spreads out further, such that more loci will have different allele frequencies and a divergence to the ancestral population is obervable (for a sufficient amount of loci).
- for a high value (like 0.5) the probability mass concentrates itself at the vertices of the $k-1$ simplex. This corresponds to a high chance of allele fixation. The fixation will be more likely for the allele that was dominant, mathematically because of the multiplication with $p^A_{l_i}$ in the ancestral population, to begin with.

A population is defined by its allele probabilites, whereby the members of the population are approximately homogenous. The allele probalities are a categorical distributions over the possible alleles at each locus, so how often a genetic variant appears in a population. So at principal, to simulate a new population, its allele probabilities have to be determined. The simulated data the modell produces is supposed to resemble SNP data, where no distinction is made between the different variants of a mutation, but solely if a mutation at ll exists compared to an undetermined hypothetical origin population.
The modell has a hierarchical structure starting with allele probabilities that are derived from an unseen ancestral population. Let $p_A(l)$ denote the probability, sampled from a uniform distribution, of an individual from the ancestral population A having a variation at locus l. Subsequently, following the genetic drift modelling from (1) by Falush, Stephens, and Pritchard 2003, for each population k an F-value F^k is chosen to introduce the magnitude of genetic drift from the ancestral population. These are used to derive the mutation probabilities of population k at locus l:

$$p_k(l) = beta(p_A(l)\frac{1-F^k}{F^k}, (1-p_A(l))\frac{1-F^k}{F^k}) \tag{3}$$

The dirichlet distribution from (1) degenerates to a beta distribtion as effectively only two alleles are considered for each locus.

Proceeding, the probability values are joined to a vector p_k and then merged with all other K populations to a matrix $\mathbf{F} = [p_1 p_2 \ldots p_K]^T$ of size $K \times L$, where L is the number of loci, such that each column of $\mathbf{F}$ gives the probabilities of each population for a specific locus l.



3.3 Admixture Layer

Apart from populations drifting from another apart, individuals of populations also migrate and mate with other individuals of other populations, which results in individuals having a genotpye that exhibits the shared ancestary of different populations. The admixture layer introduces the prospect of modelling diverse ancestary according to the admixture model presented in Pritchard, Stephens, and Donnelly 2000, whereby the flexibility is achieved by sampling the proportions of admixture from a dirichlet distribution. Its hyperparameters allow for the fine-tuning of the probabilites over the potential admixing proportions of the populations. Section 3.6 provides a more thorough look of the modelling with the dirichlet distribution.
Before translating the admixture proportions to the genotypes of individuals, another matrix $\mathbf{Q}$ is created that contains the admixture proportions of every individual. Mathematically, for an individual i the mixture weights q_i are sampled from a dirichlet distribution with K, according to the number of populations, influencing hyperparameters $\alpha 1, \alpha 2, \ldots, \alpha_K$. A "non admixed" individual j also receives admixture proportions with the speciality that the only non-zero value is a one at the kth position (with a lenient notation), indicating that the individual belongs to population k, so $q_j = [0_1, \ldots, 1_k, \ldots, 0_K]^T$. All N individuals are combined to a mixing matrix $\mathbf{Q} = [q_1, q_2, \ldots q_M]^T$ of dimensions $N \times K$.

3.4 Combining the Layers

The admixture proportions of an individual act upon each locus by weighing the allele frequency of each population at the locus according to the respective proportions and subsequently summing them. Since this accords to a linear operation, the matrix multiplication with both established matrices $\mathbf{P} = \mathbf{QF}$ yields a matrix $\mathbf{P}$ of dimension $N \times L$ that holds the allele frequencies at each locus for all individual from which their genotypes are sampled. In a final step each entry of $\mathbf{P}$ is used to sample twice from a bernoulli distribution (binomial - two tries), as it is assumed that individuals are diploidic. Each individual consequently has at each locus a value of $0, 1, 2$, according to how many mutations in total their both of their chromosomes exhibit.

Further assumptions that the model expresses include random mating and thus the absence of further lower level population structure within the populations and that the alleles captured by

the markers are neutral, meaning that they have no impact on the fitness of an individual and are therefore (with the allele independency assumption) exempt from natural selection.

3.5 Summary

In summary the generation of a new population setting for which the number of populations K is known proceeds as following:

1. Sample the ancestral allele frequencies $p_A(l) \sim Uniform(0,1)$
2. Determine the F-values F^k
3. For each of the K populations:
   (i) Sample $p^k(l) \sim beta(p_l^A \frac{1-F^k}{F^k}, (1-p_l^A)\frac{1-F^k}{F^k})$
   (ii) Combine allele probabilities into matrix $\mathbf{F}$
4. For each individual i:
   (i) Choose admixture coefficients $q_i \sim Dir(\alpha_1, \ldots, \alpha_k)$
   (ii) Combine admixture coefficients into matrix $\mathbf{Q}$
5. Calculate admixture $\mathbf{P} = \mathbf{QF}$
6. Convert each value p of $\mathbf{P}$ by sampling $bernoulli(p)$

3.6 Looking at Admixture

The admixture of an individual is determined by the Dirichlet distribution. The dirichlet distribution is parameterised by K hyperparameters $\alpha_1, \alpha_2, \ldots, \alpha_k$. K corresponds to the desired dimension of the output. The probability density function

$$f(x_1, \ldots, x_K, \alpha_1, \ldots, \alpha_K) = \frac{1}{\Gamma(\sum} \cdots$$

where $\sum_{i=1}^{K} x_i = 1$ and all $x_1 \geq 0$. So the Dirichlet distribution defines a probability density on the $K-1$-simplex and is therefore a natural choice for sampling admixture coefficients.
Of particular interest for the described modell are the hyperparameters, also called concentration parameters, as they control the mode and the variance around it. For values $a_i \geq 1$ the distribution has a single mode, whose coordinates at the maximum x is given by Bishop 2006:

$$x_i = \frac{\alpha_i - 1}{\sum_{k=1}^{K} a_k - K}$$

The mode moves therefore more towards those directions, simplex vertices that have a relatively higher valued corresponding hyperparameter compared to the other hyperparameters. In addition, the variance σ, given by

$$\sigma_i = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

where $\alpha_0 = \sum_{i=1}^{K} \alpha_i$, reaveals that higher values of hyperparameters leads to a decrease of the variance, meaning a higher concentration around the mode.
These two properties can be exploited to control the probability of sampling certain admixture coefficients. Furthermore, by sampling from the same dirichlet distribution one is able to simulate various population scenarios, such as a detached admixed cluster, which would correspond to a mode with high concentration parameters, or a population that experienced migration originating from another population, which would coincide with a degenerated dirichlet distribution that only has two nonzero, concentration values for the two involved population, which is in the end a beta distribution.

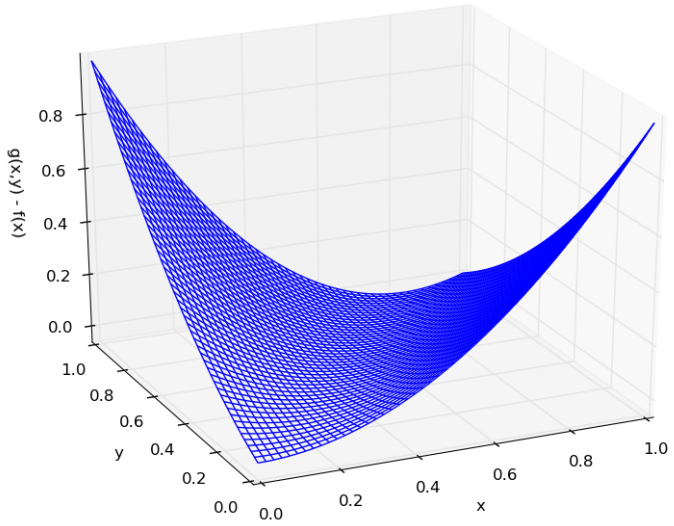
3.7 Analysis

The population centroids given by the population allele frequencies (they are the probabilities used for the bernoulli sampling and thus the mean) construct the vertices of a simplex in which the individuals approximately lie. Outliers are solely due to the natural variance created by sampling at the end from a bernoulli distribution. The probabilities the genetic information from each individual is sampled from, nonetheless always combine to a vector that lies within the simplex, for the probabilities are through the admixture coefficients a linear combination of the population allele frequencies or, in other words, of the simplex vertices. From another perspective, the matrix $\mathbf{F}$ that holds the centroids of the populations as rows, then the matrix maps every vector s from the support of an L dimensional dirichlet distribution accordingly on to the simplex spanned by the centroids (so $\mathbf{F}s$). The matrix $\mathbf{F}$ linearly transforms the L-simplex to the desired population simplex.
As presented the allele frequencies, from which the genotype of an individual is sampled, are constructed by a linear combination of the population allele frequencies, whereby the coefficients in the $K-1$ simplex lie. The $K-1$ simplex is defined by its vertices given by the K unitvectors. The sampling space from which the allele frequencies for a locus l are determined corresponds to a simplex that moved the $K-1$ simplex vertices to the allele frequencies of the populations. From another perspective one can interpret the allele frequencies as constructing a change of basis transformation through the diagonal matrix $\mathbf{D} = diag(p_{1l}, \ldots, p_{Kl})$ that squishes the unitvectors to the allele frequencies. Therefore, the allele frequencies from which the genotype is sampled could be determined by sampling a vector v from the $K-1$ simplex and transforming it into the allele frequency simplex through $\mathbf{D}v$. The transformation $\mathbf{FQ}$ given above generalises this idea for all L loci.

For further assessment a mean of quantifying the genetic dissimilarity between two individuals is necessary. As a measure of genetic distance between two individuals i and j a natural choice is to use a normalised manhatten distance because the possible genotype values $0, 1, 2$ are discrete and already reflect dissimilarity appropriately. More concretely, let N be the number of loci used as genetic markers, then $\{0,1,2\}^N \subseteq \mathbb{R}^N$ is the set containing all possible values for the genetic information of an individual. The measure of genetic distance is

$$D = \frac{1}{2N} \sum_{n=1}^{N} |l_n^i - l_n^j|$$

where l_n^i and l_n^j are the values of individual i and j respectively at locus l_n. The normalisation keeps the measure invariant to the number of loci used, as recovering more genetic information should not increse the genetic distance per se, as well es calculating out the diploidity. The



measure ranges from 0, as two individuals are genetically similar, to 1, meaning genetic disimilarity.
Suppose two individuals i and j are generated by the described modell, so sampling from a bernoulli distrebution for each loci l with the respective allele frequencies $p_i(l)$ and $p_j(l)$. The expected genetic difference of both individuals then is:

$$\begin{aligned}\mathbb{E}[D] &= \frac{1}{2N}\sum_{n=1}^{N}\mathbb{E}[|l_n^i - l_n^j|] \\ &= \frac{1}{2N}\sum_{n=1}^{N}\sum_{(x,y)\in\{(j,i),(i,j)\}} p_x(l)(1-p_x)(1-p_y(l))^2 \\ &+p_x(l)^2 p_y(l)(1-p_x(l)) \\ &+2p_x(l)^2(1-p_y(l))^2\end{aligned}$$

For individuals sampled from the same allele frequencies the expectation equates to:

$$\begin{aligned}&\frac{1}{N}\sum_{n=1}^{N} p(l)(1-p(l))(p(l)+(1-p(l))^2 \\ &= \frac{1}{N}\sum_{n=1}^{N} p(l)(1-p(l))\end{aligned}$$

This lends further justification to the choice of the distance measure for diploid individuals. It corresponds to the expected genetic variance found in a population without any further underlying substructure, as also used in the calculation of the F_{ST} value.

Let the expected distance of two individuals with arbitrary allele frequencies at a locus l be given by the function $g(p_i, p_j)$ and for two individuals sampled from the same allele frequencies $f(p_i)$. Figure **insert tag** illustrates that it is expected that the genetic distance for individuals sampled from different allele frequencies possess a greater genetic dissimilarity, therefore it is expected for individuals from the same population to cluster together and the generative

model. Furthermore, it should be expected that populations that have not diverged much from the ancestral population are more difficult to distinguish.

Apart from establishing that individuals from different clusters are distinguishable, another clustering quality that would be advantageous to aassess is the clustering density. As presented, individuals are sampled from probabilities located in the allele frequency simplex, however the resulting genotype could very much lie outside of the simplex due to the necessary descretization and thus the consequently induced variance of the binomial sampling. By determining the variance of two individuals from the same population and therefore sampled from the same allele frequencies a better impression of the cluster (population) density can be obtained. Also the expected severity of individuals lying outside the simplex can be assessed.
But before calculating the variance, for simplicity reasons the expectation of the genetic difference squared is calculated:

$$\begin{aligned}
\mathbb{E}[D^2|i,j\in k] &= \frac{1}{4N^2}\mathbb{E}[(\sum_{n=1}^{N}|l_n^i - l_n^j|)^2] \\
&= \frac{1}{4N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}[|l_n^i - l_n^j||l_m^i - l_m^j|] \\
&= \frac{1}{4N^2}(\sum_{n=1}^{N}\mathbb{E}[|l_n^i - l_n^j|^2] + \sum_{n=1}^{N}\sum_{\substack{m=1\\m\neq n}}^{N}\mathbb{E}[|l_n^i - l_n^j||l_m^i - l_m^j|]) \\
&= \frac{1}{4N^2}(\sum_{n=1}^{N}2p(l_n)^4 - 2p(l_n)^3 + p(l_n)^2 + p(l_n) + \sum_{n=1}^{N}\sum_{\substack{m=1\\m\neq n}}^{N}4p(l_n)(1-p(l_n)p(l_m)(1-p(l_m)))
\end{aligned}$$

and since

$$\mathbb{E}[D|i,j\in k]^2 = \frac{4}{4N^2}(\sum_{n=1}^{N}\sum_{m=1}^{N}p(l_n)(1-p(l_n)p(l_m)(1-p(l_m)))$$

The variance simplifies to

$$\begin{aligned}
Var(D) &= \mathbb{E}[D^2] - \mathbb{E}[D]^2 \\
&= \frac{1}{4N^2}(\sum_{n=1}^{N} -p(l_n)^4 + 4p(l_n)^3 - 3p(l_n)^2 + p(l_n))
\end{aligned}$$

This reveals that the variance decreases significantly of the order $\mathcal{O}(N^{-2})$ with the number of loci. Therefore, less severe outliers and better observable clusters are to be expected the more loci are simulated. Furthermore, the terms in the summation possess a single maximum at 0.5, so a higher variance exists the further alles are from fixation.

3.8 Relationship to LDA

The presented modell is strongly related to a modell more commonly known under the name latent dirichlet allocation (LDA) Blei, Ng, and Jordan 2003. LDA uses in the setting of natural language processing (NLP) an ensemble of words that are probabilistically associated with

certain topics, in order to determine which topics are exhibited by analysed documents, that preferably possess some associated words. Since not all existing words are associated with topics but only a selection, the selected words can be perceived as reasonable indicators of the topic context, as are the used genetic markers to determine population affiliation. A one hot encoding of wether a word is present in a document or not is the respective equivalent of wether an individual possess a gene variant at a genetic maker or not. The encoding of the selected words or respectively of the genetic markers span the feature space in which the topics/populations lie. The topics/populations then span a simplex in which the documents/individuals are mapped according to the admixture.

4 Theory

The prior described model will serve as generative model for constructing a training data set that will be used in a supervised learning scheme. Since the dimensionality of SNP genotype data poses as a challange, approximate bayesian computation (ABC) approaches will be employed. The core approximation is the construction of a summary statistics composed of the ordered eigenvalues of the coveriance matrix from the genotype data. The chosen supervised learning technique, gradient bosting with decision trees, furtherly attepmts to generalise the connection between the patterns found in the ordered eigenvalues to the numbers of populations in the respective original genotype data.

4.1 Approximate Bayesian Computation

At heart of approximate bayesian computation (ABC) lies the inference of a desired parameter value θ for given data D by relating the conditional probability of that data given the parameter value $P(D|\theta)$ to the symmetric counter part, the conditional probability of the parameter given the data $P(\theta|D)$. This is done by exploiting Bayes' rule:

$$P(\theta|D) = \frac{p(D|\theta)P(\theta)}{P(D)}$$

Where $P(D|\theta)$ is often called the likelihood, $P(\theta|D)$ the posterior, $P(\theta)$ the prior and $P(D)$ the evidence which is used in Bayes' rule solely for nomalisation porpuses.
For many problems the problem space is intractable or dimensionally too large to compute the likelihood. ABC intends to circumvent these problems.

4.1.1 Rejection Algorithm

The rejection algorithm is naturally derived from a certain perspective on conditional probability. Let A, B be a probability spaces and $P(x \subseteq A | y \subseteq B)$ the to be determined conditional probability. By infinitely sampling an event from A and from B and furtherly always recording if the produced events correspond to x and y, then the desired conditional probability is computable by taking the past unsuccessful sampling iterations into consideration. Algorithmically this can be expressed as:

Algorithm 1 Conditional Probability $P(x \subseteq A | y \subseteq B)$

```
1: i ← 1
2: for ∞ do
3:     repeat
4:         x_i ← event sampled from A
5:         i ← i + 1
6:     until sample y from B
7:     output x_i
```

And the computation amounts to:

$$\begin{aligned} P(x \subseteq A | y \subseteq B) &= \sum_{i=1}^{\infty} P((x = x_i) \cap y)(1 - P(y))^{i-1} \\ &= P(x \cap y) \sum_{i=1}^{\infty} (1 - P(y))^{i-1} \\ &= \frac{P(x \cap y)}{P(y)} \end{aligned}$$

The rejection algorithm employs a basic approach for finding the posterior distribution of the desired parameter θ for specific data D. Given a known prior distribution of θ, the algorithm samples values $\hat{\theta}$ from the prior and then inputs $\hat{\theta}$ into an appropriate model to simulate some data $\hat{D}$. If the simulated data lies with in a margin of error $\epsilon \geq 0$ from data D for a chosen metric ρ , so $\rho(D, \hat{D}) \leq \epsilon$, then the sampled prior value $\hat{\theta}$ is accepted by adding it to the final sample of parameter values for θ. The final sample should approximate the desired posterior. For further infromation and possible refinements such as using linear or non linear regression to counter a low acceptance rate or using Sequential Monte Carlo - ABC to sample from areas with higher posterior density the reader is referred to Csilléry et al. 2010.
If the task is for example to compare different models concerning a specific data set, For tasks that only require a good point estimate of θ that fits well to the data D, like the maximum a posteriori (MAP), the rejection algorithm might be too elaborate. Also, the algorithm demands to be run multiple times which could be computationally costly. Furthermore, the explicit construction of a model that simulates data introduces assumptions into the computation, as this is a further approximation it is potentially problematic for the decency of the approximation. **use instead implicit models like GAN???**

4.1.2 In Context of Supervised Learning

A supervised learning method attepmts to find a general connection between any given input data D and its desired output θ by training a malleable model. The model is instructed to infer the general connection by adapting itself in such a way that it minimises the emperical risk (for some chosen loss function) when solving a finite training set of size N, which is a data set with "presolved" values $((D_1, \theta_1), \dots (D_N, \theta_N))$. From another perspective, a supervised learning algorithm attempts to forge a model in such a way that it perfects the approximation of the desired mapping from the input space described by D into the output space, which is desirably the correct value for θ. With a uniformly distributed training dataset a supervised learining model will approximate the maximum likelihood (maximise $P(D|\theta)$ w.r.t. θ). By changing the

proportions of θ in the training data, a prior $P(\theta)$ can be implicitly set and the trained model will consequently attempt to approximate the maximum aposteriori value (MAP) (maximise $P(D|\theta)P(\theta)$ w.r.t. θ).
The task of choosing K can be viewed as a model selection problem. A rejection algorithm variant usually constitutes a reasonable choice for choosing a model for a particular data set, by for example using posterior ratios. Since the rejection algorithm produces an approximate posterior distribution a broader choice for the selection criterion exists. However, the parameter space is, even with taking summary statistics into account, because of the large dimensionality of snp genotype data seems rather daunting to tame computationally. Out of this reason one settles with the MAP value as sufficient selection criterium for the exchange of the posterior density calculation being waived. The supervised learning technique is subject to the same assumptions through the generative model and uses the same summary statisticsjust like a rejection algorithm.

4.1.3 Summary Statistics

Large dimensionality of a data set can undermine the practability of an ABC-method. By summarising the data one attempts to reduce dimensionality, while still sustaining a good approximation of the posterior. So if $S(D)$ is a summary statistics of some data D then the acceptance criterion for the rejection algorithm converts to $\rho(S(D), S(\hat{D})) \leq \epsilon$, whereby $P(\theta|D) \approx P(\theta|S(D))$ holds sufficiently. The use of summary statistics is not confined to the rejection algorithm, rather it is a general tool that allows for a tradeoff between reduction of dimensionality and the goodness of the approximation, since each summarisation usually forfeits some of the principal information. If no information is lost, so $P(\theta|D) = P(\theta|S(D))$ applies, then the summary statistics is called sufficient. **However, only exponential families have finite sufficient summary statistics, for they are maximum entropy distributions???**. A good informative choice of summary statistics is highly task and data set dependent Matthew A Nunes and Balding 2010. An overview of common heuristics and algorithms for choosing summary statistics can be found in Blum et al. 2013.
To infer the number of populations K expressed in a given dataset X the conditional probability $P(K|X)$ with respect to K is maximised. Since the large dimensionalities of the used datasets pose substantial computational difficulties, the datasets are summarised in an effective manner, such that the approximation $P(K|X) \approx P(K|sum(X))$ is sufficient for the intended inference. Bayes' theorem then yields

$$P(K|sum(X)) = \frac{P(sum(X)|K)P(K)}{P(sum(X))}$$

4.2 Choosing the Summary Statistics

The choice of adequate summary statistics is essential to obtain significant results. Large dimensional data often times demands it to be summarised, so the intended methods a reasonably applicable. In doing so, the manner summary is of great importance because each summarisation usually forfeits some of the principal information. So one is confronted with the problem of how to effectively manage the trade off between the practicability the method and the loss of information that could endager desired results.
The entropy of a distribution measures the existing uncertainty about which event appears if

one samples from the distribution. It is defined mathematically for a given continous probability mass function $P(X)$ as

$$h(x) = -\int_{supp(P)} P(x)\log(P(x))dx$$

The principal of maximum entropy states that given some prior information about the underlying probability distribution, such as already drawn samples or a constraining property, the maximum entropy distribution that incorporates the prior information is the best distribution to respect the remaining uncertainty Jaynes 1957. In other words, the maximum entropy distribution is the best distribution to fit the already obtained information if no further assumptions are to be added.
For a given mean μ and covariance Σ the multivariate continuous distribution that maximises the entropy is the multivariate Gaussian, for a proof the reader is referred to Cover and Thomas 2012. The entropy of the multivariate Gaussian is derived as following:

$$\begin{aligned}
h(x) &= -\int_{-\infty}^{\infty} N(x|\mu,\Sigma)\ln(N(x|\mu,\Sigma))dx \\
&= E[\ln(N(x|\mu,\Sigma)] \\
&= E[\ln(det(2\pi\Sigma)^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)})] \\
&= \frac{1}{2}\ln(det(2\pi\Sigma)) + \frac{1}{2}E[(x-\mu)^T\Sigma^{-1}(x-\mu)] \\
&= \frac{1}{2}\ln(det(2\pi\Sigma)) + \frac{1}{2}E[trace(\Sigma^{-1}(x-\mu)^T(x-\mu))] \\
&= \frac{1}{2}\ln(det(2\pi\Sigma)) + \frac{1}{2}E[trace(I)] \\
&= \frac{1}{2}\ln(det(2\pi e\Sigma))
\end{aligned}$$

The only non-constant factor influencing the entropy of a multivariate gaussian is the determinant of the respective covaraince matrix. Since any real symmetric matrix is diagonalisable, $det(\Sigma)$ breaks down to $det(\Sigma) = det(\mathbf{Q}^{-1})\cdot det(\mathbf{\Lambda})\cdot det(\mathbf{Q}) = \prod_{i=1}\lambda_i$, thus revealing that actually the eigenvalues of the covariance matrix are responsible for the magnitude of the entropy. In conclusion, by summarising the data by its covariance matrix, one implicitly approximates the data as being guassian and secondly it suffices for the summary to only take the eigenvalues into consideration.

4.2.1 PCA

Principle component analysis (PCA) is a statistical method that performs a basis transformation on a given data set, such that no linear correlations are anymore present in the data. Since the direction of a linear correlation corresponds to the direction of the highest variance in a concerning subspace, a new axis must be aligned according that particular direction. This construction of the axes is done by requiring in an iterative manner each new axis to align with the direction that captures the most variance in the data, which however has not been captured by previous axes.
The variance and the magnitude of variety found in a data set form a duality. Decreasing the variance in a data set by projecting it into a subspace decreases the variety, thus it endagers distinguishability between data points, which is the information. The amount of sustained

variance after projecting the data into a subspace can therefore act as an indicator for how much information was retained. So by always maximizing the captured variance of a newly added axis to the transformation, which is a subspace of the principal data set, the highest possible amount of information is retained for a projection into a subspace with a particular rank K (under the assumption that variance corresponds to information). The subspace is spanned by the K largest eigenvectors of the empirical covariance matrix, as is subsequently shown.
Let $S = \frac{1}{N} X X^T - \overline{XX}^T$ denote the empirical covariance matrix of the data matrix X. Then the expression $u^T S u$ is the empirical variance of $u^T X$, which is the data X projected on to the vector u.

$$\begin{aligned} u^T S u &= \frac{1}{N} u^T X X^T u - u^T \overline{XX}^T u \\ &= \frac{1}{N} u^T X (u^T X)^T - \overline{u^T X} (\overline{u^T X})^T \\ &= \frac{1}{N} (u^T X)^2 - (\overline{u^T X})^2 \end{aligned}$$

The empirical variance is maximised with the restriction $\parallel u \parallel = 1$ because u is supposed to be part of a new standard basis. By using a lagrange multiplier to add this restriction, the equation $\max_u u^T \Sigma u - \lambda(u^T u - 1)$ is obtained.

$$\begin{aligned} \frac{d}{du}(u^T \Sigma u - \lambda(u^T u - 1)) &= 0 \\ \Sigma u - \lambda u &= 0 \\ \Sigma u &= \lambda u \end{aligned}$$

The solution coincides with the definition of the Eigenvectors, where λ is the eigenvalue of u. Since u should be maximised, the overall solution is the eigenvector belonging to the largest eigenvalue. Solving the eigenproblem on a semi-definite matrix such as the covariance matrix Σ, yields the following factorisation:

$$\Sigma = Q \Lambda Q^T$$

Where Q is an orthogonal matrix that has the eigenvectors of Σ as its columns, and λ a diagonal matrix with the corresponding eigenvalues on its diagonal.

4.2.2 In context to clustering

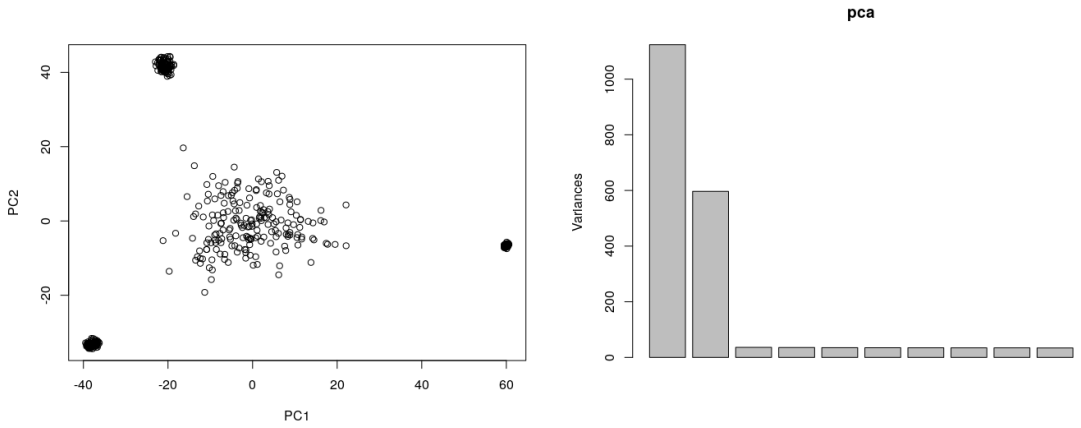
The reason for the clustering structure in data being also expressed in the eigenvalues of the covariance matrix is quite intuitive. Subsequently the term "within variance" will describe the highest possible variance recieved when a single cluster is projected on to a vector and the term "in-between variance" the highest possible variance received when multiple clusters are projected onto a vector.
The within variance of a set of clusters will usually be greater than in-between variance of any of the considered clusters, because more . Of course several clustering assumptions, such as about the spread and the density of individual clusters, are implied in this statement. The assumption made in the previously described generative model that a population k can be summarised by its allele frequencies which determines the centroid in the generative model around the individuals

of a population cluster. The centroid c_k for diploid organisms would be given by a vector where each entry corresponding to a locus l would be the bernouli mean $2p_{kl}$ (or close to it for a finite amount of population members). Also the bernoulli distribution exhibits a single mode around its mean (usually not exactly because of the descretisation). With sufficient distance between the population clusters and sufficient amount of members it is assumable that the in-between variance of a set of population clusters generated by the proposed generative model is significantly greater than the within variance of each population cluster. Furthermore, the bernoulli sampling can be regarded as noise that adds no further structure.
since the eigenvectors align themselves along the biggest possible variance, it is to be expected that their oreintation is greatly dominated by the with-in variance and thus their eigenvalues are significantly larger. The allele frequencies are from which genotypes are sampled stem from a simplex where the poplation allele frequencies compose the vertices. As discussed the simplex can be seen as the squishing of the $(K-1)$-simplex, that implements the constraint of all vectors having their values sum to 1. The constraint inhibts a degree of freedom, such that the simplex can be fully spanned by $K-1$ vectors. So it is to be expected that exactly $K-1$ eigenvectors will be needed to fully capture all the in-between variances of K population clusters. Therefore, the according eigenvalues of the first $K-1$ eigenvectors should be significantly greater than the remaining.
Admixed individuals have minimal impact on the magnitu
Another more intuitive argumentation could look like following. Consider the allele frequencies of a population as the origin. Then $K-1$ linearly independent vectors are needed to be able to reach the allele frequencies of every other population. If less would be needed, then the allele frequencies would not be independent and some population would not be considered a population, but a group of admixed individuals that are sampled from a linear combination from other population frequencies. The $K-1$ linearly independent vectors could consequently be composed of direction vectors pointing to the population allele frequencies from origin. This implies that the space between the allele frequencies is spanned by the direction vectors and thus they capture all the in between variance. The first $K-1$ eigenvectors orient themselves along the greatest variance and are linearly independent, therefore it is to be expected that they approximately span the space spanned by the direction vectors. What remains is the smaller within variance found in the clusters, concluding that the first $K-1$ eigenvalues should be significantly greater than the remaining.

4.3 Examples

A synthetic problem instance generated by the modell, could look like shown in Figure 1, where three populations that span the simplex are observable. The populations have fairly distinct F-values (meaning they drifted away from the ancestral poulation at quite different magnitudes), therefore the clusters are well seperable from one another. Within the simplex is a cluster of several admixed individuals located. They were all sampled from the same dirichlet distribution $Dir(8,8,8)$, with uniform hyperparameters, so they are concentrated around a central mode and all populations participate on average the same amount to the admixture.
Just by looking at the corresponding biggest eigenvalues, it is fairly easy, with the use of the previously insights, to infer the number of populations. The first two eigenvalues are significantly larger than the rest, thus the number of populations should be three.
Figure 2 shows a similar scenario as Figure 1, whereby the only difference consists of a different admixture of the admixed individuals. In this example admixed individuals are sampled from three different dirichlet distributions. In turn one of the hyperparameters is set to zero, thus always on population does not partake in the admixture of an individual. As a consequence the

Figure 1: Projection of three populations and one admixed on to the firs two PCs and corresponding eigenvalues



Three populations with F-values of $0.1, 0.5, 0.9$ and 100 individuals each were sampled. The admixed population derived the proportional weightings for its allele probabilites by sampling for each allele from a $Dir(8, 8, 8)$ distribution. 200 individuals were sampled for the admixed. For this simulation 10000 loci were simulated.

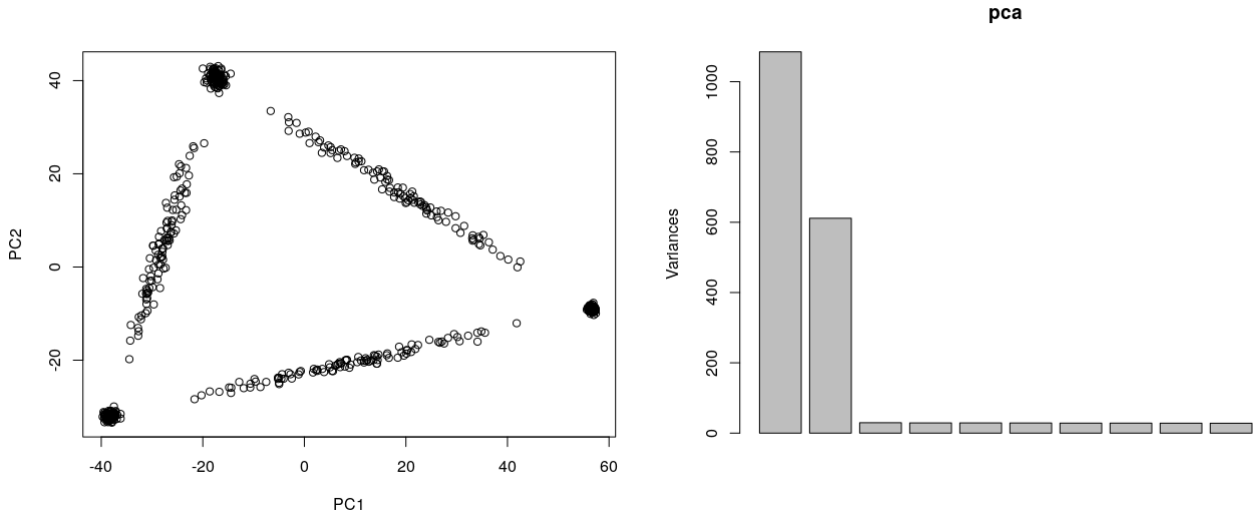
individuals are spread along the edges of the simplex, also because the non-zero hyperparametes model a lower concentration than in Figure 1.
Again the eigenvalues feature two significant large eigenvalues, making the inference of the number of populations using solely the eigenvalues once again a simple task. The natural variance of the descretisation of the allele values through the bernoulli distribution, which would allow individuals to lie outside the simplex, only has a neglectable marginal effect on the eigenvalues.

4.3.1 Difficulties

The past examples only demonstrate fairly simple problem instances, for which even a human recognition capabilities suffice. It is also possible to construct more difficult problem instances. Figure 3 is an example of such. The ploy is to simulate a setting where most of the variance is already captured by less than $k - 1$, hence making it more difficult to recognise the cut-off for significant and insignificant eigenvalues. In figure 3 there are two populations that have similar allele frequencies (F-values 0.05 and 0.1) and one population that has been subject to strong genetic drift and therefore is genetically completely distinct (F-value 0.99). Furthermore, most individuals are distributed over one of the two close populations and the one very far apart. This introduces high a high incentive for the eigenvalues to try to capture the variance of the individuals of the two populations, since a bigger distance accounts for higher variance. Inversesly, because the population that holds a fewer amount of individuals is not very far away from the positioning of the first eigenvector, such that the second eigenvector, which has to orient itself perpendicular to the first, does not capture very much variance. Also the low amount of individuals in the smaller population makes the orientation of the second eigenvector susceptible to the variance of the other populations or to any outliers. An admixed population also resides between the two greater populations, giving the first eigenvalue even more weight.

The graph with the biggest eigenvalues reveals the described dilemma. The first eigenvector

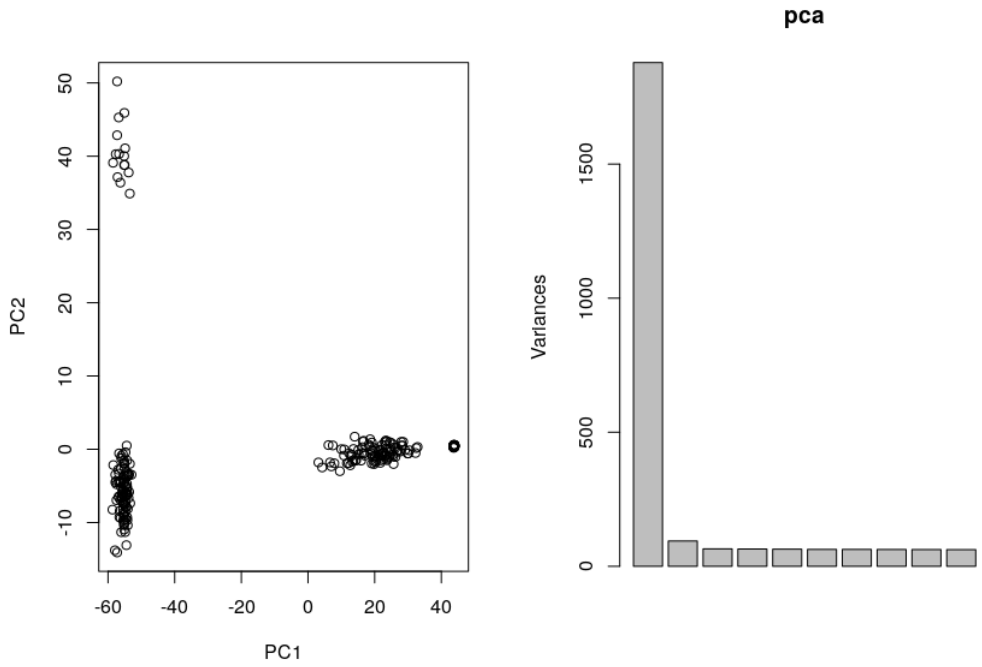
Figure 2: Projection of three populations and one admixed on to the firs two PCs



Three populations with F-values of $0.1, 0.5, 0.9$ and with 100 individuals each were sampled. Between each pair of population clusters lies an admixed population sampled from a dirichlet with 5s and a 0 for not involved populations, which corresponds to a $beta(5, 5)$. Each admixed cluster holds 200 individuals. The simulation used again 10000 loci.

accounts for almost all of the variance between the populations, rendering the other significant eigenvalue almost insubstantial and undiscernible from the other insignificant eigenvalues. In different scenarios the distance between the populations and the distribution of the individuals over the populations could be even more disadvantageous (although the question could more be of the nature to decide what is considerable to be a population, **this is part of the discussion???**). In addition, the situation becomes even more difficult if even more populations are simulated with "extreme" F-values.

Figure 3: Example of a difficult case



Three populations with F-values of $0.05, 0.01, 0.99$. The first population with the smallest F-value has 15 members, while the others have a 100 each. The mixture proportions of the admixed population were sampled from a $Dir(0, 10, 30)$. 10000 loci were simulated.

5 Boosting Decision Trees

Subsequently a concise overview of gradient boosting and decision trees is presented. For further details and idiosyncrasies of the methods the reader should consult more elaborate literature, like Trevor, Robert, and JH 2009.

5.1 Gradient Boosting

Gradient boosting is a supervised learning method for classification and regression, that iteratively adds basis learners to a linear combination to reduce an arbitrary differentiable loss function.

Let χ denote the input space. The task then is to approximate the function $f^*(x)$ that maps an arbitrary input $x \in \chi$ to the desired output $y \in \mathbb{R}$. An ensemble of M different basis learners $g_1, g_2, \ldots, g_M : \chi \to \mathbb{R}$ can be used to generalise over a training set $((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N))$ of N training pairs in a linear fashion as following:

$$f(x) = \sum_{m=1}^{M} \phi_m g_m(x)$$

where $\phi_i \in \mathbb{R}$ are the weights for each basis learner. Linear models for a set of given basis learners can be fitted via konventionell methods such as least squares, lasso, ridge Bishop 2006. $f(x)$ can be used as an approximation for various tasks like regression, classification (by for

example using a threshhold),

The model is limited by the explicit choice of the basis learners. Instead an algorithm that finds the best base learners from a hypothesis space $\mathcal{H}$ would increase the adaptive potential of the model. Thus, for a given differentiable loss function $l(x, f(x))$, an algorithm should find the best M base learners $h_1, h_2, \ldots, h_m \in \mathcal{H} : \chi \to \mathbb{R}$ and corresponding weights such that the empirical risk is minimised:

$$\underset{\phi_i, h_i}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} l(y_i, \sum_{m=1}^{M} \phi_m h_m(x_i)) \tag{4}$$

Solving this task is an optimisation problem usually beyond practicability. A common optimisation technique is too update the parameters through the use of gradient descent. Several problems arise in the case of differentiating over the hypothesis space:

- Firstly, the hypothesis space would need to be parameterised with finite dimensions. In the face of a countless manifold of possible base learners a rather daunting task. One is far better off by confining the possibilities of base learners, by choosing a subset $H \subset \mathcal{H}$ like decision trees or neural networks.
- Many modelling frameworks such as neural networks (e.g. number of hidden layers) and decision trees (e.g. tree depth) have to be parametrised at least partly discretely. Of course all discrete model features could be fixed to a constant, but that would greatly degrade the modeling capabilities.
- Lastly, some base learners do not model a differentiable functions. For e.g. decision trees model functions that possess jump discontinuouties.

Gradient boosting algorithms, firstly developed and described by Freund and Schapire 1997; J. H. Friedman 2001; J. H. Friedman 2002, circumvent the necessity of differentiability by growing the ensemble of base learners iteratively that minimises the empirical risk with respect to so called pseudo residuals.

The backbone of an gradient boosting algorithm is constituted by forward stagewise additive modeling, which works as following: Let $H \subset \mathcal{H}$ be the chosen set of possible base learners.

1. Initialise with constant like $f_0(x) = 0$
2. For each stage $m \in 1, \ldots, M$:
   (i) solve
   $$\underset{\phi_i, h_i}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} l(y_i, f_{m-1}(x_i) + \phi_m h_m(x_i))$$
   (ii) Set $f_m(x) = f_{m-1}(x) + \phi_m h_m(x)$

The optimisation step, although the amount of parameters is reduced compared to (3), only possess closed viable solving techniques for a limited amount of loss functions, like L2-loss or exponential loss. More information is provided in J. Friedman, Hastie, Tibshirani, et al. 2000.

To expand the optimisation step to any arbitrary differentiable loss function, the numerical optimisation via gradient boosting is used.

The optimisation procedure fixates at stage m the current estimates made by $f_{m-1}(x)$ of the training data in a vector $\mathbf{f}_m = [f_{m-1}(x_1), f_{m-1}(x_2), \dots f_{m-1}(x_N)]^T$.

The loss function then can be reformulated as:

$$L(\mathbf{f}) = \sum_{i=1}^{N} l(y_i, \mathbf{f}_i)$$

Then the gradient of the loss function is calculated w.r.t $\mathbf{f}$.

$$\begin{aligned}\hat{h}_m &= \nabla_{\mathbf{f}_m} L(\mathbf{f}_m) \\ &= [\partial_{\mathbf{f}_1} l(y_1, \mathbf{f}_1), \dots, \partial_{\mathbf{f}_N} l(y_N, \mathbf{f}_N)]^T\end{aligned}$$

Like in conventional gradient descent algorithms the model is adjusted in a manner that the empirical risk is minimsed along the direction of steepest descent. The direction of steepest descent corresponds to the negative of the gradient, which is $-\hat{h}_m$. Base learners that output differentiable function approximations can express the gradient via chainrule through the adjustable parameters of the model itself (like the weights in a neural network) and hence these are reciprocally changed to minimise the loss. Other base learners on the other hand have to resort to a different strategy.
The embedding in the forward additive model allows for the negative gradient $-\hat{h}_m$ to be approximated directly by a base learner. The objectiv of the new base learner $h_m(x)$ consequently concludes therein to approximate

$$h_m(x) \approx -\hat{h}_m$$

as well as possible. A possible approach would be to train a base learner on the training data, but where the labels are exchanged by $-\hat{h}_m$.
Intuitively, a base learner $h_m(x)$ should be considered as an approximate step in the direction of steepest descent of the emirical risk. Following this setup, the weight $\phi_m > 0$ can be considered as the corresponding step size to be adjusted to ones taste. As a conclusion the iterative construction of the final linear model with gradient boosting

$$f(x) = \sum_{m=1}^{M} f_{m-1}(x) + \phi_m h_m(x)$$

is a sequence of gradient descent steps towards a minimum of the empirical risk.

5.2 Decision Trees

Typical decision trees are a supervised learning method that solve a regression or classification problem by segmenting the feature space in to distinct regions, whereby all data points lying in the same region are assigned the same value by the tree.

Let $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ be the training data. The input dimension of a point x_i is D.

Suppose a tree T partitions the featurespace $X_1, X_2, \ldots, X_D$ into M regions $R_1, R_2, \ldots, R_M$. The response function outputed by T is given by:

$$f_T(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

where I is an indicator function signaling if the input x is part of a particular region. c_m is the response for a region R_m. The optimal value of c_m depends on a chosen loss function that should be minimsed w.r.t c_m over all the training points assigned to region R_m. As an example, for square loss in a regression setting this would accord to the empirical average of the label values from the training points lying in the region R_m.

A far more challenging optimisation task is finding the optimal tree that partitions the feature space into M regions, for a given loss function l.

$$\underset{R_m, c_m}{\operatorname{argmin}} \sum_{i=1}^{N} l(y_i, \sum_{m=1}^{M} c_m I(x_i \in R_m))$$

The possibilities of partitioning feature space grows exponentially with the number of features, rendering the encounter of an optimal tree in most cases computationally infeasable.

As an alternative, a greedy approximation approach can be used (CART add citation). The greedy algorithm chooses the best dimension X_d and splitting point s that minimises the loss for the two new regions that arise (w.l.o.g. only binary trees will be considered). The two new regions are defined as:

$$R_1(d, s) = \{X | X_d \leq s\} \qquad R_2(d, s) = \{X | X_d > s\}$$

The feature space is divided by a plane that is orthogonal to the axis corresponding to feature X_j. The plane cuts the value of s on that axis.
The optimisation objective thus reduces to the choice of positioning the best partitioning plane orthogonal to an axis:

$$\underset{s,d}{\operatorname{argmin}}[\underset{c_1}{\operatorname{argmin}} \sum_{x_i \in R_1(d,s)} l(y_i, x_i) + \underset{c_2}{\operatorname{argmin}} \sum_{x_j \in R_2(d,s)} l(y_j, x_j)]$$

The optimisation step yields, for a naive implementation that checks for every dimension the splitting value of a plane between every two neighbouring data points, a worstcase runningtime of $\mathcal{O}(D \cdot Nlog(N))$ ($Nlog(N)$ for sorting feature values), which is computationally much more feasable. However between two neighbouring data points x_1 and x_2 there are infinitely many positions to place the splitting plane that evaluate to the same empirical risk. As a convention the splitting value s that lies in the middle of x_1 and x_2, when they are projected on the axis X_d, is chosen by convention. The reasoning is that no further assumptions are to be added through the positioning of the plane and since the splitting of the feature space can be viewed as a bernoulli experiment (data point either lies left or right of plane) the bernoulli distribution with expected value of 0.5 is the maximum entropy bernoulli.

After finding the best split, the procedure is repeated on the newly constructed regions until the desired depth of the tree is reached. Regression and classification trees differ from one another

only through the selection of a different loss function. For regression standard loss functions like square loss would be reasonable choices. Classification trees have revealed better results for loss function that reward node purity, so to which degree does a node hold data points of a single class. Such loss functions are for example the gini index or cross entropy loss for classification. From an information theory perspective, node purity leads to a greater reduction in entropy.

Growing a too big tree T_0 is susceptible to overfitting. A regularisation technique that aids in the construction of a well generalising tree is cost-complexity pruning. The goal of pruning is to find a sub-tree $T \subseteq T_0$, by conflating all hierarchically lower nodes that lead off an internal node into that internal nodes, that minimises:

$$C_\alpha(T) = \sum_{i=1}^{N} l(y_i, f_T(x_i)) + \alpha|T|$$

where $f_T(x)$ is the estimate of Tree T for input x and $|T|$ denotes the number of terminal nodes of T. The tuning parameter $\alpha \geq 0$ punishes larger more complex trees according to its value. It resembles the regularisation term of ridge regression.
The optimal sub-tree corresponding to a particular tuning parameter T_α can be found using weakest link pruning. Weakest link pruning conflates those terminal nodes into an internal node to which the terminal nodes are all adjacent, such that the empirical risk increases minimally. Continuing with this procedure until only the root stub remains gives a sequence of subtrees $T_1, T_2, \ldots, T_n$ in which with a probability of 1 the optimal subtree T_α can be found Breiman et al. 1984. Cross validation can be used to find the optimal value for α.

Decision trees exhibit high variance, meaning that completely different splits occur and thus the output prediction rules change cosiderably when there are minor changes added to the training data. The reason resides inherently with how the splits are chosen in a greedy fashion. For example, two promising features could reduce the loss almost equally much, but just the best of both is considered for the next split. Adjusting the training data slightly by for example adding new data points could possibly change the value of the loss function enough to choose the other feature for a split the hierarchical nature consequently propagates the difference further down the tree. This behaviour reveals that the confinement to the greedy perspective when constructing a tree to some degree neglects the goal of generalisation in return for tractability.

Remedies that address model stability involve the introduction of bias. The bias-variance tradeoff possess an eminent role in machine learning as its a principle that is prominent for many models. In summary, it describes the forfeit of expected accuracy in return for decreasing the variance of an estimated parameter when the training sample is being varied. Creating an ensemble of decision trees is a widely used and fruitful approach. Ensemble approaches for decision trees include bagging, random forests and as well gradient boosting. Several further refinements improve the quality of an ensemble tree models, including randomly masking different features and training data entries for each tree, as this generates differing trees that place their splits differently and therefore deliver more uncorrelated predictions.

6 Results

7 Discussion

References

[BBS13] Katarzyna Bryc, Wlodek Bryc, and Jack W Silverstein. "Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations". In: *Theoretical population biology* 89 (2013), pp. 34–43.

[Bis06] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[Blu+13] Michael GB Blum et al. "A comparative review of dimension reduction methods in approximate Bayesian computation". In: *Statistical Science* 28.2 (2013), pp. 189–208.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[Bre+84] Leo Breiman et al. "Classification and regression trees. Wadsworth Int". In: *Group* 37.15 (1984), pp. 237–251.

[Cor+04] Jukka Corander et al. "BAPS 2: enhanced possibilities for the analysis of genetic population structure". In: *Bioinformatics* 20.15 (2004), pp. 2363–2369.

[Csi+10] Katalin Csilléry et al. "Approximate Bayesian computation (ABC) in practice". In: *Trends in ecology & evolution* 25.7 (2010), pp. 410–418.

[CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[DH73] WE Donath and AJ Hoffman. "Lower bounds for the partitioning of graphs". In: *IBM Journal of Research and Development* 17.5 (1973), pp. 420–425.

[ERG05] Guillaume Evanno, Sebastien Regnaut, and Jérôme Goudet. "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study". In: *Molecular ecology* 14.8 (2005), pp. 2611–2620.

[FHT+00] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)". In: *The annals of statistics* 28.2 (2000), pp. 337–407.

[Fie73] Miroslav Fiedler. "Algebraic connectivity of graphs". In: *Czechoslovak mathematical journal* 23.2 (1973), pp. 298–305.

[Fri01] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[Fri02] Jerome H Friedman. "Stochastic gradient boosting". In: *Computational statistics & data analysis* 38.4 (2002), pp. 367–378.

[FS97] Yoav Freund and Robert E Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

[FSP03] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies". In: *Genetics* 164.4 (2003), pp. 1567–1587.

[Har+04] Abigail V Harter et al. "Origin of extant domesticated sunflowers in eastern North America". In: *Nature* 430.6996 (2004), p. 201.

[HCC97] Daniel L Hartl, Andrew G Clark, and Andrew G Clark. *Principles of population genetics*. Vol. 116. Sinauer associates Sunderland, 1997.

[Jay57] Edwin T Jaynes. "Information theory and statistical mechanics". In: *Physical review* 106.4 (1957), p. 620.

[MS01] Marina Meila and Jianbo Shi. "A random walks view of spectral segmentation". In: (2001).

[NB10] Matthew A Nunes and David J Balding. "On optimal selection of summary statistics for approximate Bayesian computation". In: *Statistical applications in genetics and molecular biology* 9.1 (2010).

[NJW02] Andrew Y Ng, Michael I Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm". In: *Advances in neural information processing systems*. 2002, pp. 849–856.

[PPR06] Nick Patterson, Alkes L Price, and David Reich. "Population structure and eigenanalysis". In: *PLoS genetics* 2.12 (2006), e190.

[PSD00] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. "Inference of population structure using multilocus genotype data". In: *Genetics* 155.2 (2000), pp. 945–959.

[Ros+01] Noah A Rosenberg, Terry Burke, et al. "Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds". In: *Genetics* 159.2 (2001), pp. 699–713.

[Ros+02] Noah A Rosenberg, Jonathan K Pritchard, et al. "Genetic structure of human populations". In: *science* 298.5602 (2002), pp. 2381–2385.

[SM00] Jianbo Shi and Jitendra Malik. "Normalized cuts and image segmentation". In: *Departmental Papers (CIS)* (2000), p. 107.

[TRJ09] Hastie Trevor, Tibshirani Robert, and Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. 2009.

[Von07] Ulrike Von Luxburg. "A tutorial on spectral clustering". In: *Statistics and computing* 17.4 (2007), pp. 395–416.