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: 12

Submission Date: February 19, 2019

1 Introduction

In population genetics theoretical models are constructed according to devised hypotheses to investigate the population structure and its evolution for some organisams at interest. 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. However, so far no automated solution exists that satisfyingly alleviate this concern.

Throughout the following pages an approximate bayesian computation method will be presented that intends to deliver a reasonable estimate for the number of populations in SNP genotype data. It is based on a generalisation of currently used approaches, that attempt to exploit the behaviour of the spectrum of a clustered data matrix. The eigenvalues will moreover also make up the utilised summary statistics. 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

Given the genotype data of individuals, organised in a matrix X where each row corresponds to an individual, the task is to infer how many populations K are present in X. 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" **hartl1997principles**. 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 **hartl1997principles**. 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???**. The number of populations K should accord with 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

A natural approach for problems involving clustering, would be to use a well established method and adapting it to infer the number of clusters, such as by maximising the likelihood of an expactation maximisation in combination with a model quality estimator like the Akaike Information Criterion to avoid overfitting. In general 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. **falush2003inference** 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 **pritchard2000inference falush2003inference** and widely applied **rosenberg2002genetic harter2004origin rosenberg2001empirical**. 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 **evanno2005detecting**, which however appears more like mending the performance of an approach that was solely conceived as preliminary remedy **pritchard2000inference**.

A more "modern" 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 **donath1973lower fiedler1973algebraic** and later introduced into machine learning **shi2000normalized meila2001random ng2002spectral**, for further information see **von2007tutorial**. 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 **patterson2006population**.
examples
Approaches have been made with random matrix theory (RMT) to concretise the behaviour of the first $k-1$ eigenevalues **patterson2006population**, including a mathematical threshhold that destinguishes significantly larger eigenvalues from lower ones **bryc2013separation**. Utilising insights from random matrix theory furtherly remains its beginnings, so far no well performing method based on RMT has been developed. **cite ???**

1.2.1 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.

2 Modelling

2.1 Biological Background

The subsequent admixture model, follows a model proposed by **pritchard2000inference**.
model assumptions

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. **mention linkage disequilibrium and hardy-weinberg equilibrium** 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. Furthermore, ploidity is also neglected, so it is assumed the genetic data is aploidic, since **ploidity adds nothing unique for the means of inference. citation needed** The following modell has a hierarchical structure starting with allele probabilities that are derived from an unseen ancestral population. Let p_l^A 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 by **falush2003inference** established F-model, for each population k an F-value F^k is sampled from a **add distribution**. These are used to derive the mutation probabilities of population k at locus l:

$$p_l^k = beta(p_l^A \frac{1-F^k}{F^k}, (1-p_l^A)\frac{1-F^k}{F^k})$$

The probability values are joined to a vector p^k and then merged with all other K populations to a matrix $\mathbf{F} = [p^1 p^2 \dots p^K]^T$ of size $K \times L$, where L is the number of locus, such that each column of $\mathbf{F}$ gives the probabilities of each population for a specific locus l.

explain reasoning for F-values

To introduce the prospect of mating between individuals from different populations, possibilities of simulating admixed individuals and even whole admixed populations are added according to the admixture model presented in **pritchard2000inference**. An admixed individual is defined by possessing a mixture of genetic data from various populations. The mixture is a weighting modeled by a categorical distribution according to the influence of each population on the individual. For an individual i the weights q_i are sampled from a dirichlet distribution with K influence paramaters $j_1, j_2, \dots, j_K$ each corresponding to a population, thus $q_i \sim dir(j_1, j_2, \dots, j_K)$. A non admixed individual j also receives mixing parameters with the difference that the only non-zero value is a one at position k, indicating that the individual belongs to population k, so $q_j = [0_1, \dots, 1_k, \dots, 0_K]^T$. All M individuals are combined to a mixing matrix $\mathbf{Q} = [q_1, q_2, \dots q_M]^T$ of dimensions $M \times K$.

add description of influencing parameters

The mixing weights of each individual are subsequently applied over the mutation probabilites of all populations at each locus, which equals to the multiplication of both established matrices $\mathbf{P} = \mathbf{QF}$. The resulting matrix $\mathbf{P}$ of dimension $N \times L$ holds the mutation probabilities of each individual for each locus. By using each entry of $\mathbf{P}$ to sample a value from a bernoulli **why bernoulli???** distribution, the simulated SNP genotype data for each individual is obtained.

calculate centroids

Individuals can lie outside of the population simplex, however only due to the natural variance of the bernoulli sampling. The allele probabilites assigned to individuals from which their genetic information is sampled from, all lie within the simplex.

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 values of the genetic information of an individual lies on a lattice of values zeros and ones. Or more concrete, let N be the number of loci used as genetic markers, then $\{0,1\}^N \subseteq \mathbb{R}^N$ is the set containing all possible values for the genetic information of an individual. The measure of genetic distance is

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

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. 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:

$$\begin{aligned}\mathbb{E}[D] &= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[|l_n^i - l_n^j|] \\ &= \frac{1}{N}\sum_{n=1}^{N} p^i(l)(1-p^j(l)) + p^j(l)(1-p^i(l))\end{aligned}$$

The result is intuitive. It is the the empirical mean of sampling diffirent allele values.

$$\frac{2}{N}\sum_{n=1}^{N}(p(l) - p(l)^2)$$

This is an intuitive result as it
add modell relevance to real life ???
The presented modell is strongly related to a modell more commonly known under the name latent dirichlet allocation (LDA) **blei2003latent**. 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.

3 Theory

3.1 ABC

To infer the number of populations K expressed in a given dataset X the conditional probability $P(K|X)$ with respect to K is maximized. Since the large dimensionalities of the used datasets pose substantial computational difficulties, the datasets are summarized 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

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

The calculation of the likelihood $P(sum(X)|K)$ however is intractable because **???** To circumvent this problem the likelihood is implicitly calculated by employing a supervised learning method to estimate the posterior, such as a neural network or boosting decision trees. These methods are trained by trying to link summary statistics of datasets to the corresponding values for the number of populations. The prior $P(K)$ can be implicitly adjusted by changing the proportions of K in the training data. Gradient boosting with decision trees is chosen in this case, for it has demonstrated good results for various classification problems citation needed. **??? Furthermore, some inuitive reasoning exists, as explained later on, for the use of decision trees in this particular case.**

3.2 Choosing the Summary Statistics

The choice of an adequate summary statistics is essentiell to obtain significant results. On the one hand a summary of the data is necessary to handle its large dimensionality, on the other hand the manner of summarisation has to be chosen carefully to sustain the vital information necessary for the inference, because each summarisation consequently forfeits some of the principal information.

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

The only non-constant factor influencing the entropy of a multivariate gaussian is the determinant of the respective correlation 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 correlation matrix are responsible for the magnitude of the entropy. Furthermore, the multivariate Gaussian for a given mean and variance is the maximum entropy distribution, for a proof the reader is referred to **cover2012elements**.

Assume Data is linearly correlated??? connection of normal distribution to PCA

3.2.1 PCA - Introduction

introduction Principle component analysis is a statistical method that performs a basis transformation on a given data set, such that no linear correlations are anymore expressed by the data. Since the direction of a linear correlation corresponds to the direction of the highest variance in a concerning subspace, a new axis musst be aligned according that particular direction. **citation needed ???** This construction of the axes is done by requiering 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 already established axes.

The differences between data points makes them destingiushable, but also determines the magnitude of the observed variance. Decreasing the variance in a data set by projecting it into a subspace, thus endagers destinguishability (in certain features or even in total), which is 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 **singular values ???** of the empirical covariance matrix, as is subsequently shown.

Let $S = \frac{1}{N}XX^T - \overline{X}\overline{X}^T$ denote the emperical covariance matrix of the data matrix X. Then the expression u^TSu is the emprical variance of u^TX, which is the data X projected on to the vector u.

$$\begin{aligned} u^TSu &= \frac{1}{N}u^TXX^Tu - u^T\overline{X}\overline{X}^Tu \\ &= \frac{1}{N}u^TX(u^TX)^T - \overline{u^TX}(\overline{u^TX})^T \\ &= \frac{1}{N}(u^TX)^2 - (\overline{u^TX})^2 \end{aligned}$$

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

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

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.

3.2.2 In context to clustering

Cluster analysis intends to group similar data points together. **what are the clustering assumptions??? distribution? group criteria? bounderies?** Many connections between PCA and cluster analysis have been established **cite and give examples**.

Data that exhibits reasonable clustering possess a considerably unique structure. This structure also reveals itself to some degree in the orientation of the eigenvectors and the magnitude of the corresponding eigenvalues, such that they can be utilised to infer certain properties of the data, like the number of clusters as is the current intention.
Intuitively, for inferring the number of clusters, it is assumable **cluster assumptions** in a reasonable setting, including for example that each cluster has a similar amount of members, that the in-between variance between two distinct clusters is significantly greater than the variance whith-in a particular cluster. The in-between variance constitutes itself through the variance of the with-in variance of both concerning clusters and the distance between the clusters (under the assumption that outliers are possible, so cluster membership is not compulsive). While the with-in variance of a cluster is solely confined to the space assigned to that cluster, which concludes to a significantly smaller variance, espacially considering that the distance from the mean of each data point has a squared impact on the variance (definition of variance). **Moreover, the with-in variance of a cluster is always the same (for $n \to \infty$) regardless of the vector it is projected on (is this true??? proof necessary - RMT), so each cluster can be summarised to its centroid**. For these reasons a new principal component will orient itself in such a way, that it effectively captures the remaining in-between variance of the clusters.
For K many different population clusters $K-1$ significant PCs are obtained, thus allowing an inference of the number of populations. Considering only two clusters, a single significant PC would be observed that is oriented along a line connecting the two centroids of each cluster, as this would maximise the distance of the clusters after a projection on the PC and therefore maximise the variance after projection. In a general setting with K clusters, the PCs would arrange themselves as linear combinations of the in-between variances, as the overall variance is maximised so all in-betwenn variances are taken into account. Since every cluster participates in $K-1$ in-between variances and capturing these in-between variances corresponds to determining the **exact??? (the centroids???** relative positions of the other cluster, a linear combination of exactly $K-1$ vectors are needed to locate the relative positions of the other clusters **proof necessary??? less: would mean a cluster is admixed, admixture = linear combination of existing clusters → contradiction | more: some in-between variance was not captured - contradiction to maximising variance, sufficient???. The centroids lie in the span of the first $K-1$ PCs ???**
The past mentions of population clusters solely referred to clusters that are not admixed. The introduction of admixed population clusters, however does not alter the previously established theory under certain assumptions. Admixed clusters are sampled from allele probability values that are subject to a weighting of the allele probability values of the non-admixed populations according to their involvement in the admixture **refer back to modell**. This is simply a linear combination, restricted to the coefficients being proportions, of the centroids of the other populations, hence it is situated in between non-admixed populations, meaning the centroid of an admixed population lies also in the span of the PCs spanning the non-admixed populations. **In general, the centroids of the non-admixed populations constitute the corners of a simplex, that determines if a population cluster is admixed - proof ???**. If a cluster lies within the simplex it is admixed (No cluster can lie outside the simplex). Since admixed clusters lie in between the non-admixed clusters, their influence concerning the maximisation of the variance is neglegible **further explanation needed???**.

3.3 Difficulties

3.4 RMT

Let $\mathbf{A} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ be the empirical covariance matrix of $\mathbf{X}$ with $\mathbf{X}$ being an $m \times n$ matrix. $\lambda_1, \lambda_2, \ldots, \lambda_m$ are the corresponding eigenvalues of $\mathbf{A}$. The empirical spectral distribution (**ESD**) for $\mathbf{M}$ is then given by:

$$F^M(x) = \frac{1}{m} \mid \{\lambda_i \leq x \mid i \leq m\} \mid$$

Whereby $| \cdot |$ denotes the size of a set.

By assuming a theoretical setting in which $m, n \to \infty$ while $y = \frac{m}{n} \to (0, \infty)$ the Marchenko-Pastur Law extends the ESD to the continuous case.
Under the assumption that the entries of $\mathbf{X}$ are random variables iid distributed with mean 0, it states that the probability density of the eigenvalues is given by:

$$p^M(x) = \frac{1}{2\pi x y \sigma^2}\sqrt{(\rho_+ - x)(x - \rho_-)}$$

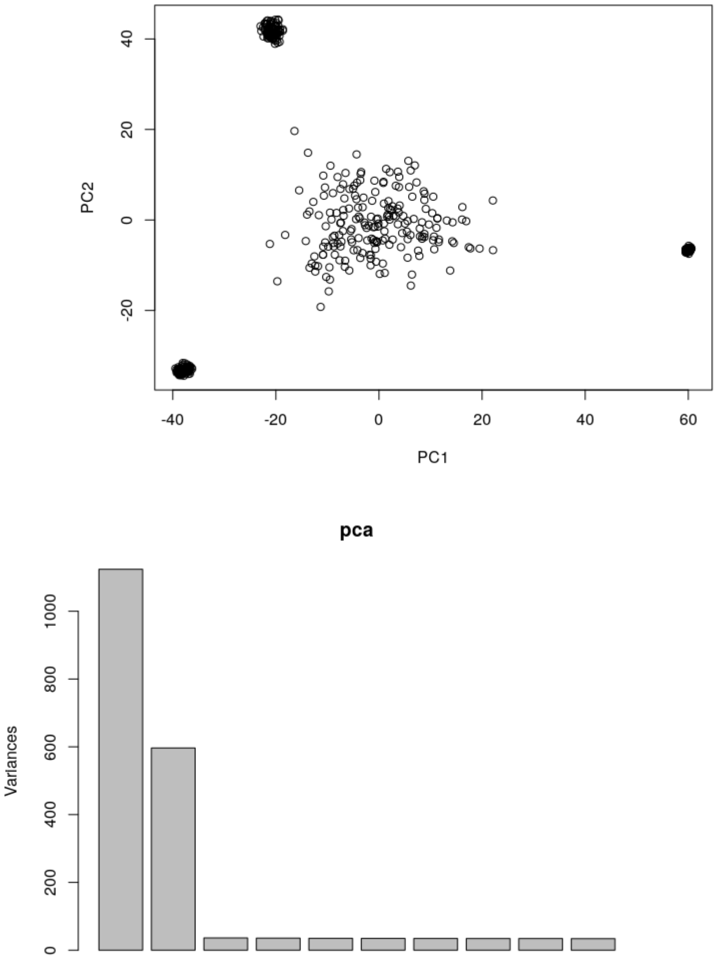
where $\rho_\pm = \sigma^2(1 \pm \sqrt{y})^2$ and σ^2 is the variance of the random variables.
Insert Plot of distribution

4 Gradient Boosting

5 Results

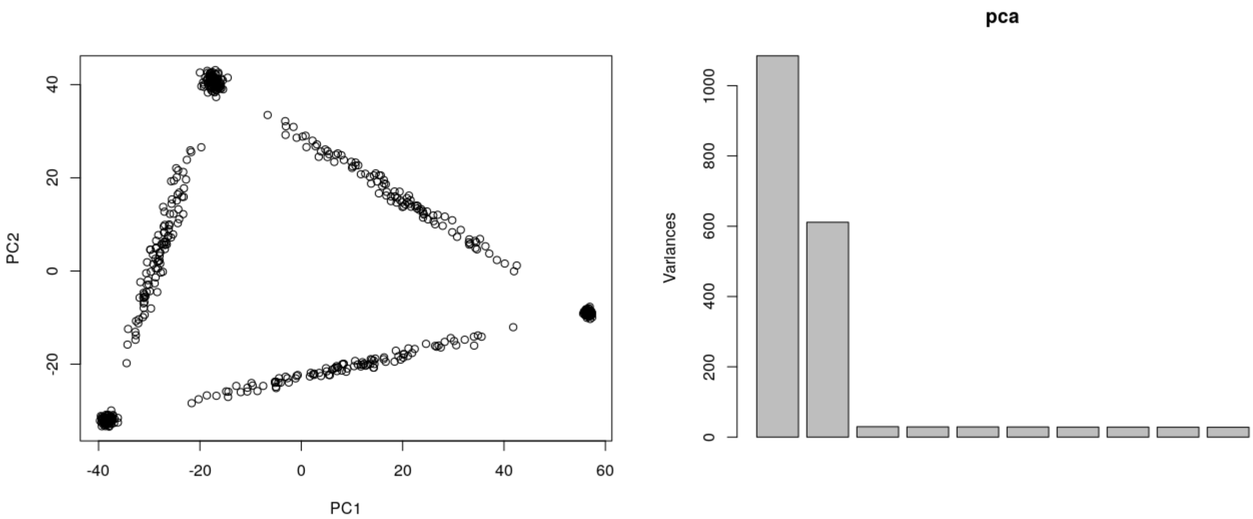
6 Discussion

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



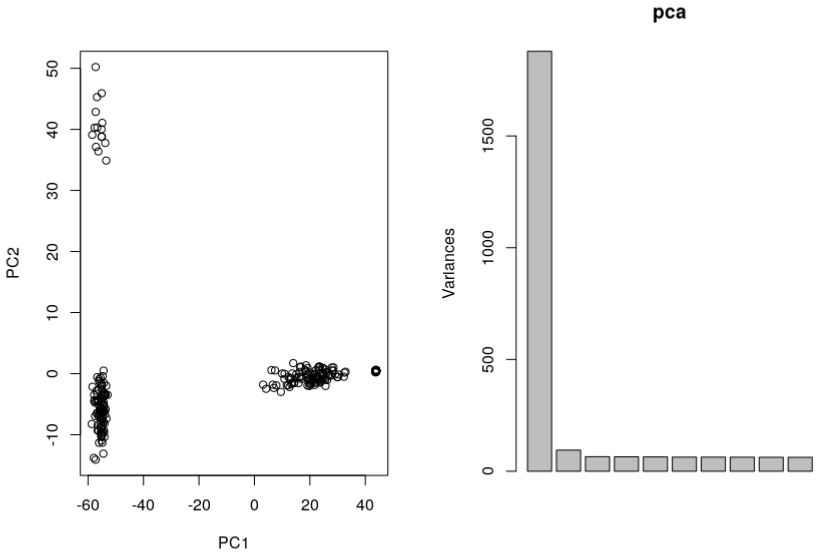
Three populations with F-values of $0.1, 0.5, 0.9$ and 100 individuals each were sampled. The admixed population derived the proportional weightings for its allele probabilites by sampling for each allele from a $Dir(8, 8, 8)$ distribution. 200 individuals were sampled for the admixed. For this simulation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

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 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.

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.