# **TEXT S4**

# **Theory linking PCA, STRUCTURE and fineSTRUCTURE**

## Introduction

In this section, we give (in the form of propositions, later backed up empirically) a detailed technical description of results described in the main paper regarding the links between different approaches to infer and analyse population structure. All results in this section refer to *unlinked* markers and we assume *haploid* data, though all results also extend to the diploid case. In a number of situations that we will highlight, we have used these results to naturally extend to the linked case.

Since the subsequent material is somewhat detailed and technical, we begin with a summary of the results, which all relate to *unlinked* markers:

1. We can approximately regard our coancestry matrix (see main text) as a rescaled version of a matrix commonly used to perform principal components analysis (PCA) on genetic data (Price et al. 2006).
2. For all forms of the normal-approximation likelihood (covering a wide class of models) the PCA matrix contains all of the information available on population structure. The form of this likelihood demonstrates a close link between model-based analyses such as that performed by STRUCTURE (Pritchard et al. 2000), and PCA analyses of the type referred to in Proposition 1. Specifically, we show that for large datasets, the PCA matrix of Proposition 1 forms a set of sufficient statistics for the “STRUCTURE likelihood” – i.e. the likelihood of the data used by STRUCTURE’s “no linkage” model, and by other similar software applications. Thus, in practice we expect that almost all the information accessible to even model-based approaches is contained within the PCA matrix, and hence also within our coancestry matrix. The permitted population structure is very general and therefore includes both discrete population models and continuous models capturing admixture.
3. We derive an asymptotic approximation of the likelihood from Proposition 2 that takes a particularly simple form under the assumption of a large number of individuals and small drift.
4. Provided population structure is not very strong, and again for large datasets, the multinomial likelihood form that we use in fineSTRUCTURE gives the same asymptotic likelihood as that used by STRUCTURE (Pritchard et al. 2000) found in Proposition 3. The proof also leads to an explicit rescaling of the multinomial likelihood (our “” factor), and implies an algorithm to infer in general, whose efficacy we test via simulation in a range of scenarios.

The results of Propositions 1 and 4, in particular, naturally motivate our approaches for analysing linked data, which we describe within the relevant sections.

For clarity, this section has slightly modified notation. We begin by defining quantities used throughout this section.

### Notes on notation

We will make heavy use of vector notation. Matrices and vectors will be in bold, and scalar quantities will be in italics. For a matrix of dimension , selects an element, selects the th row and selects the th column. We define as the (appropriate length) vector containing all ones, hence is the vector of row sums, is the vector of column sums, and is the sum over all elements of . Where ambiguity is possible we denote the length of one-vectors by subscripts e.g. . Note that unlike in the main text we do not use lower case to denote elements.

### Definitions and Initial Assumptions

As in the main text, suppose we have data for individuals, drawn from populations, and let be the population of individual . Let be the total number of segregating sites (after appropriate filtering). Let be the total number of individuals from each population .

We define to be the *raw data matrix*. Specifically, elements are binary, taking the value 1 if haplotype carries that form at SNP and 0 otherwise. We assume no missing data for the purposes of our derivations. We also assume all sites are biallelic, with sites with one or fewer occurrences of 0 or 1 removed and regarded as uninformative.

We define to be the observed *coancestry matrix* for these data (see main text, where this is called ), where is the total *expected* number of chunks – each consisting of exactly one SNP in the unlinked case – that individual copies from individual . We view this expectation as calculated using the Li and Stephens (2003) model for genetic data applied with an infinite recombination rate, and arbitrarily small mutation rate in this setting, and thus this matrix is determined by the data . By the definition of the Li and Stephens algorithm, we have:

(Note that if a non-zero mutation rate is instead used, the matrix obtained is simply a rescaled version of this, with a constant term added, except at SNPs with seen only once in the sample, which we are assuming have been removed from the data.)

We define the *parameter matrix* to be the underlying donor matrix, where gives the expected values for each entry of and so gives the fraction of chunks that individual is expected to copy from individual . In particular, note we alter notation slightly from the text and view as an matrix (which will have a block-like structure if individuals are assigned to discrete populations).

We define the *Eigenstrat PCA matrix,* to be the matrix used for principal components analysis in the Eigenstrat approach (Price et al. 2006). To construct this matrix, viewing mutations as corresponding to columns in the data matrix, columns are first scaled to have zero mean and unit variance, and subsequently the rescaled data matrix is multiplied by its transpose, to give the following:

Here, we define to be the observed sample frequency of the “1” allele at SNP :

We will make use of several assumptions, which are summarized here and more fully defined when they are introduced.

1. **No linkage disequilibrium** between loci within populations**.** Loci are independent conditional on underlying population assignments such that within populations for .
2. **Large sample size.** We only require the leading order contribution in terms of where is the number of individuals sampled*.*
3. **Normally distributed drift.** Individuals are sampled in such a way that the drift can be approximated by a normal distribution. In population models, this means that the sample contains many individuals per population, and that rare SNPs are excluded.
4. **Large number of loci.** We only require the leading order contribution in terms of where is the number of loci.
5. **Weak drift.** We only require the leading order contribution of the drift, defined in Section S4.2.
6. (Technical assumption). The distribution of the (weighted) average frequency across all sampled individuals does not contain information on population structure.
7. **More loci than individuals.** The number of individuals is small in relation to the number of loci .
8. (Technical assumption). We require to integrate out the ancestral SNP frequencies, for which we assume that any prior on these is weak compared to the likelihood.

## S4.1: The coancestry matrix and the PCA matrix

The main result of this section is the following proposition.

Proposition 1: the PCA and coancestry matrices are related

For haploid variation data where sites are treated as unlinked (A1), the off-diagonal elements of the Eigentrat PCA matrix and the observed coancestry matrix are approximately related by the following equation:

#### Proof:

Let be the empirical frequency of all SNPs excluding individual . From the definition:

The proposition follows on rearrangement for .

Note that we can write .

Using properties of the PCA matrix, we can then show that this means the Eigenstrat PCA matrix and the observed coancestry matrix  have similar eigenvectors.

Corollary 1: the PCA and coancestry matrices have the same eigenvectors

For haploid variation data where sites are treated as unlinked, assuming large population size (A2) the PCA matrix and the observed coancestry matrix have approximately identical eigenvectors.

#### Proof:

First note that

since , the ancestral frequency of the SNP. Additionally, assuming that drift is small, we can replace the denominator by and the result follows. (Note that in many sensible models as becomes large, so the result may hold even when drift is moderate).

Now note that the row and column sums of are identically zero, so trivially the -vector is an eigenvector with eigenvalue 0. Similarly the same eigenvector has eigenvalue 1 for the co-ancestry matrix . Let **v** be any *other* eigenvector of with eigenvalue . We must then have and so

Therefore **v** is also an eigenvector of with transformed eigenvalue (when *N* is large, hence the approximation) as required.

#### Discussion of Proposition 1:

The above proposition suggests that a principal components approach based on the coancestry matrix should yield results comparable to those from using standard approaches. We confirmed this fact in the main text. Note that although the proposition excludes the diagonal of the PCA matrix, in practice since each row or column sum of this matrix is exactly zero, we do not expect there to be (much) information from these diagonals. In our coancestry matrix, each row and column automatically sums to 1, with diagonal entries set to zero, and so the information “lost” using this approach, relative to the standard PCA approach, is the value of the row and column sums in the PCA matrix. However, under weak assumptions we have shown that asymptotically the expected value of each element along the diagonal of the PCA matrix is , so the off-diagonal sums in the PCA matrix have expected value , and add little or no information about population structure.

In implementing these ideas in practice, we modified our matrix slightly to ensure that eigenvalues were ranked equivalently between our approach and the PCA matrix itself (with the most historically relevant eigenvalues taking the largest values). By corollary 1, the eigenvalues of are shifted and scaled relative to those from the PCA matrix, with one large eigenvalue equal to 1. To fix this, we first removed the shift by setting diagonal elements equal to the column sums. In the unlinked case, this is identically equal to for each column, and adding these diagonal values trivially leaves eigenvectors unchanged and increases eigenvalues by this constant, removing the shift. To rescale the large eigenvalue to the value zero, we next subtract column means from each entry of the matrix. Again, in the unlinked case this simply removes from each entry of the matrix, so has no effect on either the eigenvalues or eigenvectors, apart from the large eigenvalue corresponding to the *N*-vector . Finally, the resultant matrix will have eigenvalues that are simply rescaled by a factor relative to the PCA matrix.

Our new PCA approach obviously extends trivially to using the equivalent coancestry matrix in the *linked* case. For this case, we made one small final modification, to account for the fact that in the linked case our coancestry matrix need not be exactly symmetric, with the resultant drawback that left and right eigenvectors will differ slightly. To fix this, we note that if is a general symmetric matrix, then the eigenvectors of are identical. Motivated by this fact, and following other PCA approaches, we performed PCA in the general case on the matrix . In the linked case, this symmetrisation appeared to improve results slightly – other approaches we tried did not yield obvious improvements.

## S4.2: The coancestry matrix and model-based approaches to inferring structure

### Modelling population structure using a normally distributed drift matrix

In general, population structure between separated groups is often modelled using the concept of genetic drift between populations (for details, see e.g. Pritchard et al. (2000), Nicholson et al. (2002), Patterson et al. (2006)). We begin by defining to be the frequency of the mutation at the th locus in population . Several models assume a joint prior distribution on with some shared mean and variance matrix . Conceptually, can be thought of as representing the frequency of the mutation in an ancestral population before population-specific genetic drift results in new frequencies in each separate population group. The matrix defines the covariance structure of this drift between populations. This allows, in general, for correlated drift among populations. Let us go further and assume as in the original Nicholson et al. (2002) formulation (which approximates genetic diffusion models) that

This “Normal drift” approximation is expected to be quite accurate, provided drift is relatively small and (as below) provided it is applied to SNPs that are not at frequency very close to 0 or 1, a condition reasonably appropriate for many datasets involving ascertained SNPs, and which can be easily imposed more generally.

In all that follows, we use only the above assumption regarding drift, and thus our results apply across a fairly general range of settings. In particular, allowing for correlated drift means that if populations successively split, in a tree-like structure, the results still follow. Further, the no-linkage admixture model used by, for example, the program STRUCTURE (Pritchard et al. 2000) can be thought of as simply using an appropriate choice of (with each individual representing a population).

Now at an individual level, define to be the frequency for SNP in the population to which individual *i* belongs: . From the above, we have immediately that

where is the individual-level matrix giving drift (from ) between pairs of individuals. (This matrix is singular in general, but that will not affect our analysis.) We will formulate our likelihoods in terms of . (A notation reminder: is the vector of population SNP frequencies, and is the ancestral frequency).

### Constructing an approximate likelihood

For the remainder of this section, we study inference under normally distributed drift, using an approximation to the likelihood of the data. For this approximation to be valid, we assume (A3) that the number of individuals sampled from each population , and corresponding population-specific allele frequencies are such that we may use a Normal approximation to the binomial sampling likelihood of the observed data. In practice, this means should be large, at least ≥20, and should not be very close to 0 or 1.

Later, we will need an additional technical assumption (A6) that mean allele frequencies averaged over all populations are uninformative for population structure. Provided that drift is weak we expect that typically these can contain at most weak information about the underlying population assignments, justifying this assumption.

We spend the remainder of this section deriving a form for the approximate likelihood, under these two (main) assumptions.

Preliminary to Proposition 2: Single locus likelihood

Under assumptions A1-3, and assuming also that a weak prior is placed on the ancestral SNP frequency (A8), the likelihood for a single locus can be expressed in the following form:

Where is the (non-singular) covariance matrix.

#### Proof and derivation:

Conditional on the underlying allele frequencies, the observed allele counts in a (haploid) individual *i* from population are independent among individuals, and simply Bernoulli with mean and variance ). By the assumptions, we may approximate the likelihood by behaving as if data are taken from a Normal distribution. If drift is weak, the overall sample mean frequency can be used to approximate in the variance term and we obtain a joint distribution for the data vector for all individuals, at SNP :

Making the similar approximation

integrating out the population frequencies we have by properties of Normal distributions:

Using the notation , the likelihood is then:

Define

Let Clearly is normally distributed:

Thus, has mean and variance which decreases as we increase sample size.

We can rewrite the exponent term in our likelihood:

and substituting back in we find:

where only the final term in the likelihood depends on

Intuitively, we view as a weighted mean of the individual allele counts, specifically as an estimator of the ancestral allele frequency . We have shown that is a sufficient statistic for estimating Further we can then write

The ancestral allele frequency at an individual SNP is often not of direct interest in inferring structure, and is typically integrated out of the likelihood as a nuisance parameter. Therefore, suppose that we have placed some prior distribution on . (For example, the original implementation of STRUCTURE uses a prior for .) We will also suppose that in most practical settings, this prior is relatively diffuse while the data are more informative, so that over the support of we may view as unvarying: . We now may approximately integrate out the ancestral allele frequency to give unconditionally:

In our setting, we are interested in inferring structure. Thus, terms that concern us in the likelihood are only those that depend on this structure, which is wholly characterised by the covariance matrix and so up to a constant of proportionality:

In the large-sample size setting, this likelihood approximates the truth and contains almost all available information about population structure (ignoring information present in the overall average allele frequency at a given SNP).

We must take only one more simplifying step, and that is to rewrite the above in terms of rather than . Specifically, note that:

where in the last line, is simply the overall (unweighted) sample mean frequency. Finally, we have an approximate likelihood contribution for SNP :

Given a set of drift parameters, and observed data for SNP the likelihood can readily be calculated for inference purposes. However, the more useful feature of this approximation is that it makes it straightforward to combine information across loci, which we show leads to a strong link between certain forms of the principal components matrix, and model-based approaches using likelihoods of the general form discussed above.

### Linking the PCA/coancestry matrix and the model-based approach

Under our simplifying assumptions, we can now obtain a likelihood for the entire dataset by multiplying likelihoods across (unlinked) loci. The result is given by the following proposition, which easily extends to the genotype case.

Proposition 2

Suppose we have haploid variation data where sites are treated as unlinked (A1), that we sample  individuals from each of underlying populations, where each is large, and that we exclude very rare SNPs (A2-3). Consider a general model of normally distributed allele frequency drift at each locus, with an individual level drift covariance matrix . If the prior placed on the ancestral allele frequency is weak (A8), and if we assume there is no information about drift from the overall allele frequency of each mutation in the sample (A6), then the likelihood of the data can be approximated solely in terms of the “Eigenstrat” PCA matrix from Proposition 1:

where

Hence in this model, *all* available information regarding the underlying population structure is contained in the PCA matrix.

#### Proof:

The stated assumptions mean that the result derived in Preliminary Proposition 2 holds. Hence, multiplying the likelihood across independent sites we obtain:

(noting that the last term in the exponent is a constant, so does not affect proportionality: this is included to avoid a large term in the exponent for ), from which the required result is immediate by properties of determinants.

This result is general for the study of datasets containing mainly common mutations. It applies in both models incorporating discrete population structure, and also in no-linkage admixture models, where sample size is in a sense “large”, in that the Normal approximation to the likelihood still applies. An implication is that apart from the weak information contained in the average allele frequencies of mutations across groups, and neglecting information from occasional variants that reach loss or fixation in some groups (expected to occur for stronger drift), the Eigenstrat style PCA matrix contains all the available information about population structure. Thus in these models, we expect model-based approaches to generally only succeed if there is a signal of structure from the PCA approach, as has been observed previously in practice (S5).

## S4.3: Asymptotic behaviour of models of population structure, with weak drift

If population structure is strong, we anticipate that all reasonable approaches are likely to identify it. Of more interest is then the setting where population structure is weak (in a sense we will specify precisely below). In this section, we show that in the setting of weak structure the likelihood of a particular underlying drift matrix has a particularly simple approximate form. This likelihood is a function of our coancestry matrix. Further, we show that the likelihood form is approximately equivalent to that obtained by assuming that an appropriate rescaling of the coancestry matrix yields a multinomial distribution. Thus, in the weak drift case and for unlinked sites, the likelihood form we use in the main text (with an appropriate choice of scaling parameter) approximates the full likelihood of the data. (In the next section we will show how we generalise these ideas to the linked case, and describe our approach for estimating .)

Proposition 3: Asymptotic behavior for subtle population structure

Assume that the conditions of Proposition 2 hold (A1-3,A6,A8), and also assumptions A4 and A5, where (A5) is precisely that the drift is small in the sense that order and higher terms are negligible. The likelihood given by Proposition 2 can be simplified to

Where is the row and mean zero’d drift matrix with , and is the per-locus PCA matrix.

#### Proof:

We will approximate terms in the likelihood shown in Proposition 2. Recalling , under assumptionA5we then have correct to second order:

We start with the likelihood from Proposition 2:

Again correct to second order and after a little simplification we find:

Note that viewing the PCA matrix as estimating the drift, and evaluating the expected asymptotic values for this matrix without drift as becomes large, we have in large datasets, unless . For the case write . Again, for large datasets, is small. Then, correct to second order in we have:

Substituting into the likelihood:

and after simplification, we obtain

We now substitute ,which clearly has zero row and column means, so is the relative drift among individuals. Further, by expansion of the desired result it is clear that all other terms from cancel, and is a constant independent of the parameters so can be included or excluded in the proportionality. The result immediately follows.

#### Discussion of Proposition 3:

This is the key result; to second order the likelihood only depends on the data through the usual principal components matrix which is transformed to give , and the transformed drift matrix . Hence only the relative drift is identifiable from data, and the absolute overall value of drift cannot be inferred. Further, the likelihood behaves asymptotically as if the transformed entries in the principal components matrix are independent and normally distributed, with mean and variance .

We have used a series of approximations in deriving this result, which suppose essentially that the sample size is large, while drift is small. Examining the latter assumption in more detail, it can be seen that our approximations require the setting , so the average drift is small compared to . By Price et al. (2009), overall structure is strong if

In this setting, our assumptions may not hold but we expect (and see in practice) all competitive approaches to perform well and identify structure. In the case where structure is much weaker, then for some non-negative we can consider:

which will be considerably less than 1 provided , as is usually the case in current genetic studies where there are many more markers than individuals. Nevertheless, it is clearly important to evaluate the performance of both our approximation to the likelihood, and the resulting inference framework, via simulation, which we do extensively in Section S6. The results verify excellent agreement between the theory and observed results.

## S4.4: fineSTRUCTURE model

The likelihoods as written above produce a dimension reduction by avoiding the need to consider SNPs individually. However, this likelihood is not particularly convenient to work with directly – neither is it straightforward to extend to incorporate LD information. It is more natural to attempt to perform inference based on our coancestry matrices, which give expected counts, and *do* extend immediately to the LD case, while still giving a dimension reduction. We showed above that these matrices also relate closely to the PCA matrix, which our approximate likelihoods are defined in terms of.

For count data (and by extension our *expected* count data matrix), a natural model is the multinomial distribution, which is a member of the exponential family of distributions, enabling the use of computationally convenient conjugate prior choices. Although we have expected counts, we may nevertheless consider a model of the same form, where individuals are ordered. In general, we allow ourselves to multiply the count matrix by a constant before applying the likelihood. We can view this as calculating an *effective* number of loci across the genome.

In this section, we first attempt to identify a relationship between parameters in this multinomial likelihood, and the STRUCTURE model. The parameter in a multinomial model, for group , is the probability an observation falls in this bin. Suppose we consider a general multinomial distribution for our setting, where we have a mean probability that individual copies a SNP (or a chunk) from individual . It is immediate from Proposition 1 that if is large, , the expected count proportion for individual from individual is expressible in terms of the expectation of the PCA matrix and to order :

so we can relate this parameter to the individual by individual relative drift matrix . Note we could conceptually also include the case , even though we have disallowed self-copying in practice.

Proposition 4: Asymptotic behaviour of the multinomial likelihood

Assuming A1-6 and additionally (A7) that the whilst both the population size and number of loci are large, additionally

has the same asymptotic form found in Proposition 3, with (for haploids) the choice .

#### Proof and derivation:

The expected total number of chunks copied by individual from each other individual is large provided drift is small and the number of loci is large compared to the number of individuals.

 We again employ the central limit theorem, this time relying on the fact that the number of *loci* is large. We approximate the joint distribution of the counts using a multivariate normal distribution of dimension (because this entry is therefore removed from the likelihood). is the count vector, is the expected number of counts, and is the model covariance matrix for row with the other rows . Since , we can avoid having to use the generalised inverse directly, and have the standard likelihood form:

where is the empirical frequency of copying from individual from individual . Then since ,

Using Proposition 1, and noting that :

Substituting for , and discarding lower order terms in , we have:

where is defined as above, and where in the final line we include an additional term in the exponent corresponding to , assuming that the relative contribution of this single within individual term to the likelihood is small if is large.

The first term in this likelihood can be ignored for two reasons. Firstly, we can incorporate this term into the prior distribution on since it does not depend on the data or the number of loci or individuals except for a constant of proportionality, for example by fitting the prior variance in copying fractions based on the data. Secondly, the contribution of this term is small. Substituting we have:

(Here we have discarded higher order terms in *N* and , and noted that rows of sum to 0). The constant term can be incorporated into the proportionality, and the term inside the exponent is small compared to the remaining term in the likelihood. Choosing , we see that for small drift () we can discard the above term, leading to:

On simplification and discarding higher than second order terms in the drift we have

which precisely matches the form we derived in Proposition 3.

Thus by choosing , our multinomial likelihood can be viewed as approximating the likelihood used by STRUCTURE run on the same data, at least in the case of a large number of unlinked loci, many individuals, and weak drift. Extension to the diploid case is trivial by substituting into , giving in this case.

#### Discussion of Proposition 4:

The above results motivate the extension that modelling the coancestry matrix as multinomial may be appropriate even in the case where linkage disequilibrium is present, particularly when the number of loci is large, provided an appropriate value is chosen for . In the unlinked case, the value (for haploids) can be viewed as an adjustment for the fact that the variance of the entries in the count matrix is overestimated by the multinomial likelihood. Indeed, if structure is weak, the true counts variance is approximately a factor lower than that given in the multinomial model. Further, the symmetry of the matrix gives an additional factor 2 for the off-diagonal terms yielding . Noting that the multinomial distribution is generally approximated by a multivariate normal, we suggest that in general the multinomial model may still give a reasonable approximation, provided we use a value of equal to twice the ratio of the correct underlying variance of the coancestry matrix to the “multinomial” estimated variance (at least if structure is “weak” – in the strong structure case results are insensitive to the value of ).

In the general case it is not possible to analytically identify , so we instead estimate the true underlying variance of the number of chunks copied using a bootstrapping approach, calculate the required ratio using the genome-wide copying fractions, and substitute in this value of into the likelihood. The above theory demonstrates that this approach will work in the simple unlinked loci case at least. In Section S6 we demonstrate that the bootstrap approach leads to the same value of as the theory in the no-linkage case, and additionally works well in the linkage case using our copying model.