Genetics and population analysis

Probabilistic models of genetic variation in structured populations applied to global human studies

Wei Hao1,†, Minsun Song1,†,‡ and John D. Storey1,2,*

1Lewis-Sigler Institute for Integrative Genomics and 2Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08544, USA

*To whom correspondence should be addressed.
†The authors wish it be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
‡Present address: Department of Mathematics and Statistics, University of Nevada Reno, Reno, NV 89557, USA.

Received on January 5, 2015; revised on October 19, 2015; accepted on October 27, 2015

Abstract

Motivation: Modern population genetics studies typically involve genome-wide genotyping of individuals from a diverse network of ancestries. An important problem is how to formulate and estimate probabilistic models of observed genotypes that account for complex population structure. The most prominent work on this problem has focused on estimating a model of admixture proportions of ancestral populations for each individual. Here, we instead focus on modeling variation of the genotypes without requiring a higher-level admixture interpretation.

Results: We formulate two general probabilistic models, and we propose computationally efficient algorithms to estimate them. First, we show how principal component analysis can be utilized to estimate a general model that includes the well-known Pritchard–Stephens–Donnelly admixture model as a special case. Noting some drawbacks of this approach, we introduce a new ‘logistic factor analysis’ framework that seeks to directly model the logit transformation of probabilities underlying observed genotypes in terms of latent variables that capture population structure. We demonstrate these advances on data from the Human Genome Diversity Panel and 1000 Genomes Project, where we are able to identify SNPs that are highly differentiated with respect to structure while making minimal modeling assumptions.


Contact: jstorey@princeton.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

One of the most important goals of modern human genetics is to accurately model genome-wide genetic variation among individuals, as it plays a fundamental role in disease gene mapping and characterizing the evolutionary history of human populations. In this article, we develop latent variable probabilistic models and estimation methods of genetic variation that provide allele frequency estimates of each individual/SNP combination in the presence of arbitrarily complex population structure. Accurate estimates of allele frequencies in this setting allow for improved tests of genetic associations with complex traits and other population genetic analyses which do not rely on overly restricted models of population structure. For example, the models and methods developed here provide the key estimation step in the implementation of a new framework for association testing in the presence of arbitrarily complex structure (Song et al., 2015). Other applications we explore here are to
identify loci differentiated with respect to structure, test for random mating in the presence of structure, generalize the estimation of $F_{ST}$, and characterize the global distribution of allele frequencies of disease SNPs—all making minimal assumptions about the complexity and form of structure.

A longstanding problem has been to provide well-estimated probabilistic models of observed genotypes in the presence of complex population structure (see Raj et al., 2014 and references therein). A series of influential publications have proposed methods to estimate a model of admixture, where the primary focus is on the admixture proportions themselves (Alexander et al., 2009; Pritchard et al., 2000; Tang et al., 2005), which in turn may produce estimates of the allele frequencies of every genetic marker for each individual. Here, we significantly relax the assumptions made about the manifestation of structure to yield more general latent variable models of structure. Rather than targeting admixture proportions, we instead focus on the estimation of the individual-specific allele frequencies, and we show that we make significant gains over existing methods in the accuracy and computational efficiency in estimating these quantities. The individual-specific allele frequencies, rather than admixture proportions, are ultimately the key quantities that need to be estimated in the applications we discuss as well as in the association testing method of Song et al. (2015).

We propose two flexible genome-wide models of individual-specific allele frequencies as well as methods to estimate them. First, we develop a model that includes as special cases the aforementioned models; specifically, the Balding–Nichols (BN) model (Balding and Nichols, 1995) and its extension to the Pritchard–Stephens–Donnelly (PSD) model (Pritchard et al., 2000). However, we identify some limitations of our method to estimate this model. We therefore propose a second model based on the log-likelihood of the data that allows for rapid estimation of allele frequencies while maintaining a valid probabilistic model of genotypes.

The estimate of the first model is based on principal component analysis (PCA), which is a tool often applied to genome-wide data of genetic variation in order to uncover structure. One of the earliest applications of PCA to population genetic data was carried out by Menozzi et al. (1978). Exploratory analysis of complex population structure with PCA has been thoroughly studied (Manni, 2010; Menozzi et al., 1978; Novembre and Stephens, 2008; Rendine et al., 1999; Sokal et al., 1999). We show that a particular application of PCA can also be used to estimate allele frequencies in highly structured populations, although we have to deal with the fact that PCA is a real-valued operation and is not guaranteed to produce allele frequency estimates that lie in the unit interval $[0,1]$.

The estimate of the second model is based on generalized factor analysis approaches that directly model latent structure in observed data, including categorical data (Bartholomew et al., 2011) in which genotypes are included. We utilize a factor model of population structure (Engelhardt and Stephens, 2010) in terms of non-parametric latent variables, and we propose a method called ‘logistic factor analysis’ (LFA) that extends the PCA perspective toward likelihood-based probabilistic models and statistical inference (Collins et al., 2002). LFA is shown to provide accurate and interpretable estimates of individual-specific allele frequencies for a wide range of population structures. At the same time, this proposed approach provides visualizations and numerical summaries of structure similar to that of PCA, building a convenient bridge from exploratory data analysis to probabilistic modeling. LFA plays a key role in the aforementioned new test of genome-wide association of Song et al. (2015), called the genotype-conditional association test.

We compare our proposed methods with existing algorithms, ADMIXTURE (Alexander et al., 2009) and fastSTRUCTURE (Raj et al., 2014), and show that when the goal is to estimate all individual-specific allele frequencies, our proposed approaches are conclusively superior in both accuracy and computational speed. We apply the proposed methods to the Human Genome Diversity Project (HGDP) (Cann et al., 2002; Rosenberg et al., 2002, 2005) and 1000 Genomes Project (TGP) (1000 Genomes Project Consortium, 2010) datasets, which allows us to estimate allele frequencies of every SNP in an individual-specific manner. Using LFA, we are also able to rank SNPs for differentiation according to population structure based on the likelihoods of the fitted models. In both datasets, the most differentiated SNP is proximal to SLC24A5, and the second most differentiated SNP is proximal to EDAR. Variation in both of these genes has been hypothesized to be under positive selection in humans. In the TGP dataset, the second most different SNP is rs3827760, which confers a missense mutation in EDAR and has been recently experimentally validated as having a functional role in determining a phenotype (Kamburov et al., 2013). We also identify several SNPs that are highly differentiated in these global human studies that have recently been associated with diseases such as cancer, obesity and asthma.

## 2 Methods

### 2.1 Models of Allele Frequencies

It is often the case that human and other outbred populations are ‘structured’ in the sense that the genotype frequencies at a particular locus are not homogeneous throughout the population (Asle and Balding, 2009). Geographic characterizations of ancestry often explain differing genotype frequencies among subpopulations. For example, an individual of European ancestry may receive a particular genotype according to a probability different than an individual of Asian ancestry. This phenomenon has been observed not only across continents, but on very fine scales of geographic characterizations of ancestry. Recent studies have shown that population structure in human populations is quite complex, occurring more on a continuous rather than a discrete basis (Rosenberg et al., 2002). We can illustrate the spectrum of structural complexity with Figure 1, which shows dendrograms of hierarchically clustered individuals from the HapMap (phase II), HGDP and TGP datasets. The HapMap samples strongly indicate explicit membership of each individual to one of three discrete subpopulations (due to the intended sampling scheme). On the other hand, the clusterings of the HGDP and TGP individuals show a very complex configuration, more representative of random sampling of global human populations.

Let us introduce $Z$ as an unobserved variable capturing an individual’s structure, which we will estimate with dimension $d$. Let $x_{ij}$ be the observed genotype for SNP $i$ and individual $j$ ($i = 1, \ldots, m$, $j = 1, \ldots, n$), and assume that $x_{ij}$ is coded to take the values 0, 1, 2. We call the observed $m \times n$ genotype matrix $X$. For SNP $i$, the allele frequency can be viewed as a function of $Z$, i.e. $\pi_i(Z)$. For a sampled individual $j$ from an overall population, we have ‘individual-specific allele frequencies’ (Thornton et al., 2012) defined as $\pi_{ij} \equiv \pi_i(z_j)$ at SNP $i$. Each value of $\pi_{ij}$ informs us as to the expectation of that particular SNP/individual pair under the scenario we observed a new individual at that locus with the same structure, specifically as $E[x_{ij}]=2 \times \pi_{ij}$. If an observed SNP genotype $x_{ij}$ is treated as a random variable, then we assume that $\pi_{ij}$ serves to model $x_{ij}$ as a Binomial parameter: $x_{ij}(Z = z_j) \overset{\text{Binomial}}{\sim} (2, \pi_i(z_j))$. (We will drop the conditioning on $Z$ in the subsequent text for
The focus of this article is on the simultaneous estimation of the genetic population structure for these individuals since the allele frequencies are a linear combination of the rows of the matrix \( S \). The PSD model aims instead to estimate all individual-specific allele frequencies \( p_{ik} \) when \( d = 1 \), which have standalone interpretations. We define the distance matrix \( D \) as a linear basis, we have:

\[
D = S \Gamma S' \tag{1}
\]

where \( \Gamma \) is an \( m \times d \) matrix and \( S \) is an \( d \times n \) matrix with \( d \leq n \), and the entries of both matrices are unrestricted real numbers. The \( d \times n \) matrix \( S \) encapsulates the genetic population structure for these individuals since \( S \) is not SNP-specific. The \( m \times d \) matrix \( \Gamma \) maps how the structure \( S \) is manifested in the allele frequencies. Operationally, each SNP's allele frequencies are a linear combination of the rows of \( S \), where the linear weights for SNP \( i \) are contained in row \( i \) of \( \Gamma \). We define the dimension \( d \) so that \( d = 1 \) corresponds to the case of no structure: when \( d = 1 \), \( S = (1, 1, \ldots, 1) \) and \( \Gamma \) is the column vector of marginal allele frequencies.

This model is not necessarily the most effective way to estimate \( p_{ik} \) when working in the context of a probabilistic model or with the likelihood function given the data. Model 1 resembles linear regression, where the allele frequencies are treated as a real-valued response variable that is linearly dependent on the structure. A version of regression for the case of categorical response variables (e.g., genotypes) with underlying probability parameters is logistic regression. We developed an approach called logistic factor analysis (LFA),
which is essentially an extension of non-parametric factor analysis to \([0, 1, 2]-\)valued genotype data. The justification for LFA derives from that of generalized linear models (McCullagh and Nelder, 1989), where in our case observed covariates are instead replaced with unobserved latent variables that must also be estimated.

The log-likelihood is the preferred mathematical framework for representing the information the data contain about unknown parameters (Lehmann and Casella, 1998). Suppose that the model assumption holds such that \(x_{ij} \sim \text{Binomial}(2, \pi_i)\). We can write the log-likelihood of the data for SNP \(i\) and individual \(j\) as:

\[
\ell(\pi_i | x_{ij}) = \log \left( \Pr(x_{ij} | \pi_i) \right) = \log \left( \binom{2}{x_{ij}} (1 - \pi_i)^{2-x_{ij}} \right) = x_{ij} \log \left( \frac{\pi_i}{1 - \pi_i} \right) + 2 \log(1 - \pi_i).
\]

The log-likelihood of SNP \(i\) for all unrelated individuals is the sum:

\[
\sum_{j=1}^{n} \ell(\pi_i | x_{ij})
\]

where \(\pi_i = \logit(p_i)\), is the logit function and is written as \(\logit(p_i)\). \(\logit(p_i)\) is called the ‘natural parameter’ or ‘canonical parameter’ of the Binomial distribution and is the key component of logistic regression (McCullagh and Nelder, 1989). An immediate benefit of working with \(\logit(\pi_i)\) is that it is real valued, which allows us to directly model \(\logit(p_i)\) with a linear basis.

Let \(L\) be the \(m \times n\) matrix with \((i,j)\) entry equal to \(\logit(\pi_i)\). We form the following parameterization of \(L\):

\[
L = AH.
\]

where \(A\) is \(m \times d\) and \(H\) is \(d \times n\) with \(d \leq n\). In this case we can write

\[
\logit(p_i) = \sum_{k=1}^{d} a_{ik} b_k,
\]

where all parameters are free to span the real numbers. We choose the value of \(d\) by identifying the one that provides the best goodness-of-fit (Supplementary materials, Section S2).

We call the rows of \(H\) ‘logistic latent factors’ or just ‘logistic factors’ as they represent unobserved variables that explain the inter-individual differences in allele frequencies. In other words, the logit of the vector of individual-specific allele frequencies for SNP \(i\) can be written as a linear combination of the rows of \(H\):

\[
[\logit(\pi_1), \ldots, \logit(\pi_m)] = \logit(\pi_i) = \sum_{k=1}^{d} a_{ik} b_k,
\]

where \(b_k\) is the \(k\)th row of \(H\). Similarly, we can write:

\[
(\pi_1, \ldots, \pi_m) = \pi_i = \frac{\exp \left( \sum_{k=1}^{d} a_{ik} b_k \right)}{1 + \exp \left( \sum_{k=1}^{d} a_{ik} b_k \right)}.
\]

The relationship between our proposed LFA approach and existing approaches of estimating latent variables in categorical data is detailed in Supplementary materials, Section S6. Specifically, it should be noted that even though we propose calling the approach LFA, we do not make any assumptions about the distribution of the factors (which are often assumed to be normal). A technically more detailed name of the method is a ‘logistic nonparametric linear latent variable model for Binomial data.’

### 2.2 Estimation algorithms

The two models presented earlier make minimal assumptions as to the nature of the structure. For example, in Model 1, both \(F\) and \(S\) are permitted to be real valued. This allows us to apply a PCA-based algorithm directly to the genotype matrix \(X\), obtaining estimates of \(F\), \(F\), and \(S\). In essence, \(F\) is estimated by forming the projection of \(X/2\) onto the top \(d\) principal components of \(X\) with an explicit intercept for the \(d = 1\) case. One drawback of this approach is that because PCA is designed for continuous data, we have to take additional steps to constrain \(F\) to be in the range \([0, 1]\). However, we show in Results that \(F\) is still an extremely accurate estimate of the allele frequencies \(F\) for all formulations of \(F\) considered here, including the PSD model.

**Algorithm 1: Estimating \(F\) from PCA:**

1. Let \(\hat{\mu}_i\) be the sample mean of row \(i\) of \(X\). Set \(\pi^*_i = x_{ij} - \hat{\mu}_i\) and let \(X^*\) be the \(m \times n\) matrix with \((i,j)\) entry \(\pi^*_i\).
2. Perform singular value decomposition (SVD) on \(X^*\) which decomposes \(X^* = U \Lambda V^T\). Note that the rows of \(\Lambda V^T\) are the \(n\) row-wise principal components of \(X^*\) and \(U\) are the principal component loadings.
3. Let \(X^*_{d-1}\) be the projection of \(X^*\) on the top \(d - 1\) eigen-vectors of this SVD, \(X^*_{d-1} = U_{1 \ldots (d-1)} \Lambda_{1 \ldots (d-1)} V_{1 \ldots (d-1)}^T\).
4. Construct \(F\) by adding \(\hat{\mu}_i\) to row \(i\) of \(X^*_{d-1}\) (for \(i = 1, \ldots, n\)) and multiplying the resulting matrix by \(1/2\). In mathematical terms, \(F\) is defined to be the \((d-1)\) eigen-vectors of this SVD, \(F = \frac{1}{2} V_{1 \ldots (d-1)}^T\)

\[
F = \frac{1}{2} V_{1 \ldots (d-1)} \Lambda_{1 \ldots (d-1)} V_{1 \ldots (d-1)}^T
\]

where \(\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{d-1})\) and \(\lambda_i\) is the \(i\)th diagonal entry of \(U^T\).

5. Since it may be the case that some \(\theta^*_i\) are such that \(\theta^*_i < 0\) or \(\theta^*_i > 1\), we truncate these. The final PCA based estimate of \(F\) is formed as \(\hat{F}\) where the \((i,j)\) entry \(\hat{\theta}_{ij}\) is defined to be

\[
\hat{\theta}_{ij} = \begin{cases} 
C & \text{if } \theta^*_i < C \\
\theta^*_i & \text{if } C < \theta^*_i < 1 - C \\
1 - C & \text{if } \theta^*_i \geq 1 - C 
\end{cases}
\]

for some \(C \approx 0\). An estimate of \(L\) can be formed as \(\hat{L} = \logit(\hat{F})\), where we used \(C = \frac{1}{20}\), which is the minimum resolution of the data given \(2n\) alleles are observed. In summary, \(F\) is a projection of \(X\) into its top principal components, scaled by \(1/2\), and truncated so that all values lie in the interval \((0, 1)\).

For Model 2, we propose a method for estimating the latent variable \(H\). Starting from the output of Algorithm 1, we apply the logit transformation to the subset of rows that had no truncation, i.e. no
values where \( \tilde{\pi}_i \leq C \) or \( \tilde{\pi}_j \geq 1 - C \). We then extract the right singular vectors of this transformed subset. As long as the subset is large enough to span the same space as the row space of \( L \), this approach accurately estimates the basis of \( H \). Next, we calculate the maximum likelihood estimation of \( A \) parameterized by \( H \) to yield \( \tilde{A} \), and then set \( L = \tilde{A}H \). This involves performing a logistic regression of each SNP’s data on \( H \). In order to estimate the individual-specific allele frequency matrix \( F \), we calculate \( F = \text{logit}^{-1}(L) \). An important property to note is that all \( \tilde{\pi}_i \in [0, 1] \) due to the fact that we are modeling the natural parameter.

**Algorithm 2:** Estimating Logistic Factors

1. Apply steps 1–4 of Algorithm 1 to obtain the estimate \( \tilde{F} \) from Step 4.
2. Recalling that \( \tilde{\pi}_i \) is the \((i, j)\) entry of \( \tilde{F} \), we choose some \( C \geq 0 \) and form
\[
S = \{i : C < \tilde{\pi}_i < 1 - C, \forall j = 1, \ldots, n\}.
\]

\( S \) identifies the rows of \( \tilde{F} \) where the logit function can be applied stably. Here we use \( C = \frac{C}{2} \).
3. Define \( F_S \) to be the corresponding subset of rows of \( \tilde{F} \), and calculate \( L_S = \text{logit}(F_S) \). Let \( L_{S'} \) be the row-wise mean centered and standard deviation scaled matrix \( L_S \).
4. Perform SVD on \( L_{S'} \) resulting in \( L_{S'} = TAW^T \). Set \( H \) to be the \( d \times n \) matrix composed of the top \( d - 1 \) right singular vectors of the SVD of \( L_{S'} \) stacked on the row \( n \)-vector (1, 1, \ldots, 1):
\[
H = \begin{pmatrix}
1 & 1 & \ldots & 1 & 1 \\
\begin{pmatrix}
W_1^T(d-1) \\
1 & 1 & \ldots & 1 & 1
\end{pmatrix}
\end{pmatrix}
\]
\[
= \begin{pmatrix}
1 & 1 & \ldots & 1 \\
1 & 1 & \ldots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \ldots & 1
\end{pmatrix}
\]

**Algorithm 3:** Estimating \( F \) and \( L \) from LFA

1. Apply Algorithm 2 to \( X \) to obtain \( H \).
2. For each SNP \( i \), perform a logistic regression of the SNP genotypes \( x_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \) on the rows of \( H \), specifically by maximizing the log-likelihood
\[
\ell(\pi_i|X, H) = \sum_{j=1}^{n} x_{ij} \log \left( \frac{\pi_i}{1 - \pi_i} \right) + 2 \log(1 - \pi_i)
\]
under the constraint that \( \text{logit}(\pi_i) = \sum_{k=1}^{d} \alpha_k \hat{b}_{ki} \). It should be noted that an intercept is included because \( \hat{b}_{ki} = 1 \) \( \forall \) \( j \) by construction.
3. Set \( \hat{\alpha}_k \) \( (j = 1, \ldots, n) \) to be equal to the maximum likelihood estimates from the above model fit, for each of \( i = 1, \ldots, m \). Let \( L = AH \), \( F = \text{logit}^{-1}(L) \), and \( \tilde{\pi}_j \) be the \((i, j)\) entry of \( F \):
\[
\tilde{\pi}_j = \exp\left\{ -\frac{1}{1 + \exp\left\{ \sum_{k=1}^{d} \alpha_k \hat{b}_{ki} \right\}} \right\}
\]

PCA-based estimation of Model 1 requires one application of SVD and LFA requires two applications of SVD. We leverage the fact that \( n \gg d \) to utilize Lanczos bidiagonalization which is an iterative method for computing the SVD of a matrix (Baglama and Reichel, 2006). Lanczos bidiagonalization excels at computing a few of the largest singular values and corresponding singular vectors of a sparse matrix. While the sparsity of genotype matrices is fairly low, we find that in practice using this method to perform the above estimation algorithms is more effective than using methods that require the calculation of all the singular values and vectors. This results in a substantial reduction of the computational time needed for the implementation of our methods.

**3 Results**

We applied our methods to a comprehensive set of simulation studies and to the HGDP and TGP datasets.

**3.1 Simulation studies**

To directly evaluate the performance of the estimation methods (see Section 2.2), we devised a simulation study where we generated synthetic genotype data with varying levels of complexity in population structure. Genotypes were simulated based on allele frequencies subject to structure from the BN model, the PSD model, spatial structure populations and real datasets. For the first three types of simulations, the allele frequencies were parameterized by Model 1, while for the real-data simulations, the allele frequencies were taken from model fits on the data themselves.

A key property to assess is how well the estimation methods capture the overall structure. One way to evaluate this is to determine how well \( S \) from the PCA-based method (Algorithm 1) estimates the true underlying \( S \), and similarly how well \( H \) from LFA estimates the true \( H \). Note that even though the genotype data were generated from the F of Model 1, we can evaluate \( H \) by converting with \( L = \text{logit}(F) \). To evaluate PCA, we regressed each row of \( F \) on \( S \) and calculated the average \( R^2 \); similarly, for LFA we regressed each row of \( L \) on \( H \) and calculated the average \( R^2 \) value. The results are presented in Table 1.

**Table 1. Accuracy in estimating linear bases for \( S \)**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Mean ( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGP fit by PCA</td>
<td>0.9998</td>
</tr>
<tr>
<td>TGP fit by LFA*</td>
<td>0.9912</td>
</tr>
<tr>
<td>HGDP fit by PCA</td>
<td>0.9996</td>
</tr>
<tr>
<td>HGDP fit by LFA*</td>
<td>0.9835</td>
</tr>
<tr>
<td>BN</td>
<td>0.9999</td>
</tr>
<tr>
<td>PSD ( z = 0.01 )</td>
<td>0.9998</td>
</tr>
<tr>
<td>PSD ( z = 0.1 )</td>
<td>0.9998</td>
</tr>
<tr>
<td>PSD ( z = 0.5 )</td>
<td>0.9996</td>
</tr>
<tr>
<td>PSD ( z = 1 )</td>
<td>0.9993</td>
</tr>
<tr>
<td>Spatial ( \alpha = 0.1 )</td>
<td>0.9999</td>
</tr>
<tr>
<td>Spatial ( \alpha = 0.25 )</td>
<td>0.9999</td>
</tr>
<tr>
<td>Spatial ( \alpha = 0.5 )</td>
<td>0.9999</td>
</tr>
<tr>
<td>Spatial ( \alpha = 1 )</td>
<td>0.9998</td>
</tr>
</tbody>
</table>

Column 1 shows the scenario from which the data were simulated. Columns 2 and 3 display the estimation accuracy of the PCA-based method (Column 2) and LFA (Column 3). Column 2 shows the mean \( R^2 \) value when regressing the true \( \pi_1, \pi_2, \ldots, \pi_n \) on \( S \) from PCA, averaging across all SNPs. Column 3 shows the mean \( R^2 \) value when regressing the true \( \text{logit}(\pi_1), \text{logit}(\pi_2), \ldots, \text{logit}(\pi_n) \) on \( H \) from LFA, averaging across all SNPs. All estimated standard errors fell between \( 10^{-6} \) and \( 10^{-5} \) so these are not shown. Note for each scenario, \( R^2 \) values are higher for the method from which the true \( F \) matrix was generated. All but the two scenarios marked with an asterisk (*) are from Model 1, while the two marked scenarios are from Model 2, where we took \( F = \text{logit}^{-1}(L) \).
We specifically note that when the PSD model was utilized to simulate structure, we were able to recover the structure very well (Supplementary Fig. S2) without needing to employ the computationally intensive and assumption-heavy Bayesian model fitting techniques from Pritchard et al. (2000). In addition, it seems that the PSD largely captures the geometry of S where it may be the case that S can be recovered with a high degree of accuracy by transforming S back into the simplex. By comparing the results on the real data (Fig. 2) with the simulated data (Supplementary Fig. S2), one is able to visually assess how closely the assumptions of the PSD model resemble real datasets. When structure was simulated that differed substantially from the assumptions of the PSD model, our estimation methods were able to capture that structure just as well (Supplementary Fig. S3). This demonstrates the flexibility of the proposed approaches.

We also compared PCA and LFA to two methods of fitting the PSD model, ADMIXTURE (Alexander et al., 2009) and fastSTRUCTURE (Raj et al., 2014), by determining how well the methods estimated the individual-specific allele frequencies $p_{ij}$. A subset of results is shown in Table 2, and the full set of results is shown in Supplementary Table S1. For the real data scenarios, we simulated genotypes based on estimates of F from the four different methods, thus giving each method an opportunity to fit its own simulation. The methods were compared by computing three different error metrics with respect to the oracle F: Kullback-Leibler divergence, absolute error and root mean squared error (Supplementary materials, Section S4). PCA and LFA significantly outperformed ADMIXTURE and fastSTRUCTURE, which confirms the intuitive understanding of the differences between the models: the goal of Models 1 and 2 is to estimate the allele frequencies $p_{ij}$, while the PSD model provides a probabilistic interpretation of the structure by modeling them as admixture proportions.

The computational time required to perform the proposed methods was also significantly better than ADMIXTURE and fastSTRUCTURE. Both proposed methods completed calculations on average over 10 times faster than ADMIXTURE and fastSTRUCTURE, with some scenarios as high as 150 times faster. This is notable in that both ADMIXTURE and fastSTRUCTURE are described as computationally efficient implementations of methods to estimate the PSD model (Alexander et al., 2009; Raj et al., 2014).

### Table 2. A comparison of accuracy in estimating $p_{ij}$ parameters where data were simulated from the PSD model for varying $\alpha$

<table>
<thead>
<tr>
<th>Method</th>
<th>PCA</th>
<th>LFA</th>
<th>ADX</th>
<th>FS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 0.01$</td>
<td>$7.2 \times 10^{-3}$</td>
<td>$7.6 \times 10^{-3}$</td>
<td>$1.7 \times 10^{-1}$</td>
<td>$1.7 \times 10^{-1}$</td>
</tr>
<tr>
<td>$\alpha = 0.1$</td>
<td>$7.2 \times 10^{-3}$</td>
<td>$9.3 \times 10^{-3}$</td>
<td>$2.4 \times 10^{-1}$</td>
<td>$2.4 \times 10^{-1}$</td>
</tr>
<tr>
<td>$\alpha = 0.5$</td>
<td>$7.3 \times 10^{-3}$</td>
<td>$9.0 \times 10^{-3}$</td>
<td>$1.8 \times 10^{-1}$</td>
<td>$1.8 \times 10^{-1}$</td>
</tr>
<tr>
<td>$\alpha = 1.0$</td>
<td>$7.4 \times 10^{-3}$</td>
<td>$8.4 \times 10^{-3}$</td>
<td>$2.2 \times 10^{-1}$</td>
<td>$2.2 \times 10^{-1}$</td>
</tr>
</tbody>
</table>

Methods used are the proposed PCA-based method (Algorithm 1) and LFA method (Algorithms 2 and 3), and two competing methods, ADMIXTURE (ADX) and fastSTRUCTURE (FS), that directly fit the PSD model. The values reported are root mean squared error in the $p_{ij}$ parameter. See Supplementary Table S1 for more extensive comparisons.

3.2 Analysis of the HGDP and TGP data

We analyzed the HGDP and TGP data using the proposed methods. The HGDP data consisted of $n = 940$ individuals and $m = 431,345$ SNPs, and the TGP data consisted of $n = 1500$ and $m = 339,100$ (see Supplementary materials, Section S1 for details). We first applied PCA and LFA to these datasets and made bi-plots of the top three PCs and top three LFs (Fig. 2). It can be seen that PCA and LFA provide similar visualizations of the structure present in these data. In addition, the structures estimated by these methods are related, but not identical, to the population labels provided in the original studies. We next chose a dimension $d$ for the LFA model (Model 2) for each dataset. This was done by identifying the value of $d$ that provides the best overall goodness of fit (Supplementary materials, Section S2). We identified $d = 15$ for HGDP and $d = 7$ for TGP based on this criterion.

One drawback of utilizing a PCA-based approach (Algorithm 1) for estimating the individual-specific allele frequencies F is that we are not guaranteed that all values of the estimates lie in $[0, 1]$, so some form of truncation is necessary. We found that 65.4% of the SNPs in the HGDP dataset and 26.5% in the TGP dataset resulted in at least one estimated individual-specific allele frequency <0 or >1 before the truncation was applied. Therefore, the truncation in forming the estimate F is necessary when employing Algorithm 1 to estimate F from Model 1. On the other hand, due to the formulation of Model 2, all estimated allele frequencies fall in the valid range when applying LFA (Algorithms 2 and 3).

The LFA framework provides a natural computational method for ranking SNPs according to how differentiated they are with...
respect to structure. Accurately ranking SNPs according to this differen-
tiation is a technique often used to identify genetic polymor-
phisms that are strong candidates for instances of positive selection
(Coop et al., 2009). Note that existing methods typically require one
to first assign each individual to one of K discrete subpopulations
(as done in Coop et al., 2009) which may make unnecessary assump-
tions on modern datasets such as HGDP and TGP. In order to rank
SNPs for differentiation, we calculate the deviance statistic when
performing a logistic regression of the SNPs genotypes on the logis-
tic factors. Specifically, we calculated the deviance by comparing the
models \( \logit(p_i) = a_i b_j \) versus \( \logit(p_i) = \sum_{k=1}^{d} a_i b_k \), where the
former model is intercept only (i.e. \( d = 1 \), no structure).

Our application of LFA to identify SNPs with allele frequencies
differentiated according to structure can be developed further. First,
the recently proposed ‘jackstraw’ approach (Chung and Storey,
2015) provides a manner in which statistical significance can be as-
signed to these SNPs. Assigning statistical significance to the popula-
tion differentiation of SNPs has traditionally been a difficult
problem (Akey et al., 2002). Second, we found the deviance measure
tends to have more extreme values for SNPs with larger minor allele
frequencies (MAFs). Therefore, the ranking of SNPs may be made
more informative if MAF is taken into account. Third, although this
ranking is identifying differentiation and not specifically selection, it
may provide a useful starting point in understanding methods that
attempt to detect selection.

The most differentiated SNPs (Supplementary Tables S2 and S3)
reveal some noteworthy results, especially considering the flexible
approach to forming the ranking. SNPs located within or very close
to \( SLC24A5 \) were the top ranked in both HGDP and TGP. This
gene is well known to be involved in determining skin pigmentation
in humans (Lamason et al., 2005) and is hypothesized to have been
subject to positive selection (Sabeti et al., 2007). The next most
highly ranked SNPs in both studies are located in \( EDAR \), which
plays a major role in distinguishing phenotypes (e.g. hair follicles)
among Asians. SNP rs3827760 is the second most differentiated
SNP in the TGP data, which has also been hypothesized to be under
positive selection in humans and whose causal role in the hair fol-
licle phenotype has been verified in a mouse model (Kamberov et al.,
2013). SNPs corresponding to these two genes for both studies
are plotted in increasing order of \( \hat{p}_ij \) values, revealing subtle vari-
ation within each major ancestral group in addition to coarser dif-
ferences in allele frequency (Fig. 3). Other noteworthy genes with
highly differentiated proximal SNPs include:

- \( FOXP1 \), which is a candidate gene for involvement in tumor pro-
gression and plays an important regulatory role with \( FOXP2 \)
  (Banham et al., 2001; Shigekawa et al., 2011);
- \( TBC1D1 \) in which genetic variation has been shown to confer
  risk for severe obesity in females (Stone et al., 2006);
- \( KIF3C \), a novel kinesin-like protein, which has been hypothe-
sized to be involved in microtubule-based transport in neuronal
  cells (Sardella et al., 1998);
- \( KCNMA1 \), a recently identified susceptibility locus for obesity
  (Jiao et al., 2011);
- \( CTNNA3 \) in which genetic variation has been shown to be asso-
ciated with disocyanate-induced occupational asthma (Bernstein
et al., 2013);
- \( PTK6 \), breast tumor kinase (Brk), which is known to function in
  cell-type and context-dependent processes governing normal dif-
  ferentiation (Ostrander et al., 2010).

We have provided information on the 5000 most differentiated
SNPs for both TGP and HGDP as Supplementary material files.
4 Discussion

We have investigated two latent variable models of population structure to simultaneously estimate all individual-specific allele frequencies from genome-wide genotyping data. Model 1, a direct model of allele frequencies, can be estimated by using a modified PCA and Model 2, a model of the logit transformation of allele frequencies, is estimated through a new approach we called LFA. For both models, the latent variables are estimated in a non-parametric fashion, meaning we do not make any assumptions about the underlying structure captured by the latent variables. These models are general in that they allow for each individual’s genotype to be generated from an allele frequency specific to that individual, which includes discretely structured populations, admixed populations and spatially structured populations. In LFA, we construct a model of the logit of these allele frequencies in terms of underlying factors that capture the population structure. We have proposed a computationally efficient method to estimate this model that requires only two applications of SVD. This approach builds on the success of PCA and Model 2, a model of the logit transformation of allele frequencies lying between 0 and 1.

Although PCA has become very popular for genome-wide genotype data, it should be stressed that PCA is fundamentally a method for characterizing variance and special care should be taken when applying it to estimate latent variables. The authoritative treatment of PCA (Jolliffe, 2010) eloquently makes this point throughout the text and considers cases where factor analysis is more appropriate than PCA through examples reminiscent of the population structure problem. Here, we have shown that modeling and estimating population structure can be understood from the factor analysis perspective, leading to estimates of individual-specific allele frequencies through their natural parameter on the logit scale. At the same time, we have avoided some of the difficulties of traditional parametric factor analysis by maintaining the relevant non-parametric properties of PCA, specifically in making no assumptions about the underlying probability distributions of the logistic factors that capture population structure.

Funding

This research was supported in part by NIH grant R01 HG006448.

Conflict of Interest: none declared.

References


