Statistical Methods for Inferring Population Structure with Human Genome Sequence Data

Jennifer L Kirk

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Timothy Thornton, Chair

Ali Shojaie

Michael Wu

Ann Vander Stoep, GSR

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Statistical Methods for Inferring Population Structure with Human Genome Sequence Data

Jennifer L Kirk

Chair of the Supervisory Committee:
Chair Timothy Thornton
Biostatitics

Population structure is systematic variation in the human genome due to non-random mating because of physical or cultural barriers. Population structure is of interest in several fields of medicine, including population genetics, medical genetics, and personalized genomics. Advances in sequencing technology have lead to a precipitous drop in the cost to sequence the human genome, which has lead to a plethora of sequencing studies in recent years. This increase in the availability of genotype data has led to a commensurate increase in the number of statistical methods for analyzing sequence data. To date, the majority of these new methods have focused on association testing, with relatively little work on inferring population structure, despite the importance of population structure inference. There are several challenges to inferring population structure with sequencing data, including: an abundance of rare variants (loci where there is little variation across human populations) and the large number of loci. Existing methods are not directly applicable to rare variants and few computationally feasible methods exist. This dissertation considers the problem of inferring population structure with human genome sequence data. We present new statistical methods, with theoretical justification, extensive simulation studies, and applications to the 1000 Genomes Project data. We also develop extensions of the methods that are computationally feasible for large sequencing data sets and that allow for the use of reference population samples to better elucidate population structure from sequence data.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

There are so many people who without their help, I would not have made it through UW. First, and foremost, I would like to thank my advisor, Tim, for his steadfast enthusiasm and encouragement. When I started graduate school, I dreaded the thought of doing research, but with Tim's guidance and encouragement, I have enjoyed my dissertation work immensely.

I would like to thank my committee members, Ali, Mike, Ann and Joe, for taking time out of their busy schedules to support me. Their critical feedback greatly improved the quality of my dissertation. I would also like to thank the Biostatistics department, especially Gitana, for helping me complete my degree.

Along the way, I have had many mentors. The kind biostatisticians who worked with me at NIAID, especially Mike Fay, Mike Proschan, and Pam Shaw, took great care to make sure I was prepared for graduate school. I learned so much in my two years at NIAID, which gave me a great foundation for graduate school. Debra Kaysen and Isaac Rhew provided invaluable mentorship and support, allowing me to take on more responsibility as my knowledge and confidence grew.

I have been fortunate to have many, many friends in both the Biostatistics and Statistics programs, plus many other departments at UW. Thank you all for the homework help, happy hours, holiday parties and hugs!

Finally, without my family I would not have made it past the first day of classes. Dad, thank you for encouraging me when I felt out of place and out of my depth. Mom, thank you for patiently listening to an endless litany of complaints. Scott, my husband, thank you for everything. Your love and kindness have made me the person I am.

DEDICATION

To Mom, who sacrificed her dreams so that mine might live.

ACRONYMS

ACB African Caribbeans from Barbados

ASW Americans of African Ancestry in the Southwest United States

BEB Bengali from Bangladesh

CDX Chinese Dai in Xishuangbanna, China

CEU Utah Residents with Northern and Western European Ancestry

CHB Han Chinese in Beijing, China

CHS Southern Han Chinese

CLM Colombians from Medellin, Colombia

ESN Esan in Nigeria

FIN Finnish in Finland

GBR British in England and Scotland

GIH Gujarati Indian from Houston, Texas

GRM genetic relatedness matrix

GWD Gambians in Western Divisions in the Gambia

HMM hidden Markov model

HWE Hardy-Weinburg equilibrium

IBS Iberian Population in Spain

ITU Indian Telugu from the United Kingdom

JPT Japanese in Tokyo, Japan

KHV Kihn in Ho Chi Minh City, Vietnam

LD linkage disequilibrium

LWK Luhya in Webuye, Kenya

MAF minor allele frequency

MDS Multi-dimensional Scaling

MSL Mende in Sierra Leone

MXL Mexican Ancestry from Los Angeles, California

PCA principal component analysis

PEL Peruvians from Lima, Peru

PJL Punjabi from Lahore, Pakistan

PUR Puerto Rican

SDR spectral dimension reduction

SNP single nucleotide polymorphism

SNV single nucleotide variant

SSVD stochastic singular value decomposition

STU Sri Lankan Tamil from the United Kingdom

SVD singular value decomposition

TSI Toscani in Italy

YRI Yoruba in Ibadan, Nigeria

Chapter 1

INTRODUCTION

Since the first human genome was sequenced in 2003, there has been a precipitous decrease in the time and cost to sequence the full human genome (see Figure 1.1, based on data from [1]). Furthermore, the cost to sequence the whole human exome decreased below $1000 in 2015, and the cost to sequence the whole human genome decreased by 63% over the second half of 2015 to $1500 [1]. As the cost of sequencing has decreased, there has been a commensurate increase in sequencing studies. In particular, since sequence data has an abundance of rare variants (loci with little to no variation within human populations) and association studies with common variants have not fully accounted for the observed heritability of many diseases, scientists have been interested in identifying associations between disease susceptibility and rare variants, because it is believed that rare variants may explain the some of the missing heritability. While there has been a plethora of methods published for association testing with sequencing data, particularly rare variant association testing methods, relatively few methods exist for inferring population structure with sequence data. In addition, several papers have demonstrated inflated type I error rates when using rare variant association tests adjusted for population structure of common variants inferred using existing methods, as rare variants can have different structure than common variants.

Sequencing data presents several challenges to inferring population structure. Existing methods for inferring population structure are not suitable for use with rare variants. Current model-based methods assume subjects have ancestry from a fixed number of discrete populations. It is unclear whether this assumption is accurate for rare variants, which appear to delineate fine-scale structure. EIGENSTRAT, the leading machine-learning based method, is numerically unstable when applied to rare variants. Furthermore, these methods are not

generally computationally feasible for extremely large genomic data sets. Hence, there is a need for computationally feasible methods to infer population structure from sequence data, including rare variants.

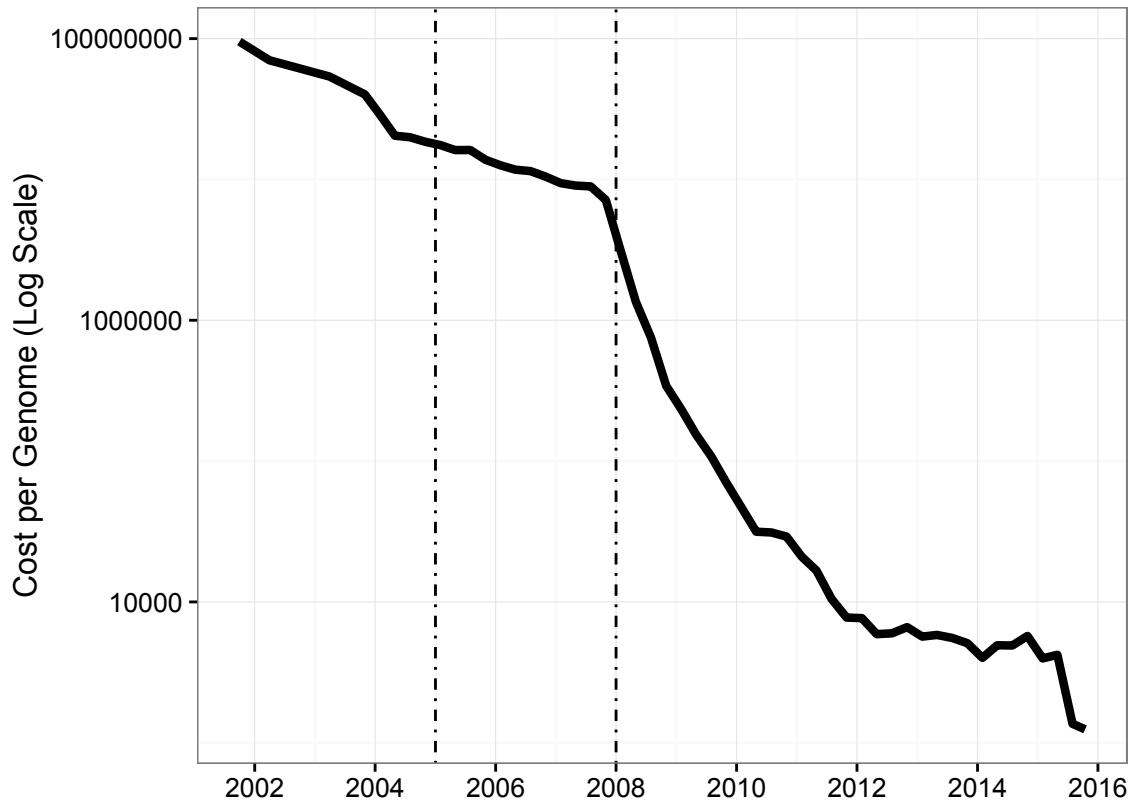


Figure 1.1: The cost to sequence a full human genome in US dollars, based on data from the National Human Genome Research Institute. The first dashed line at 2005 indicates the release of the first next-generation sequencing platform, 454 pyrosequencing, and the second dashed line indicates start of the rapid decrease in costs as the popularity of sequencing increased.

1.1 Aims

Given the increased interest in analyzing full sequence data and the evidence to suggest that there are inflated type 1 error rates when rare variant association tests are only adjusted for common variant population structure, the goal of this dissertation is to develop computationally feasible methods for inferring population structure from sequence data.

The rest of this chapter presents general background information on population genetics and sequence data, along with a general discussion of EIGENSTRAT, a population inference method that is referred to extensively throughout this dissertation. Chapter 2 presents a new method, PCA-seq, for inferring population structure from sequence data that can be applied to rare variants. Chapter 3 introduces a modification to PCA-seq that significantly improves the speed of computation for large data sets. Chapter 4 extends PCA-seq by proposing a method for deriving weights that are informative for population structure. Chapter 5 concludes with a summary and a discussion of limitations and future work.

1.2 Background

While each chapter has a separate background section, some concepts and methods are central to this dissertation and are referenced in multiple chapters. This section provides background on these topics, along with general background on the 1000 Genomes Project data set, which is used in the applied examples.

1.2.1 Population Structure

Population structure is systematic variation in the allele frequencies of the human genome due to non-random mating. There are many different causes of non-random mating among humans, including physical separation, migration, and cultural norms. The patterns of systematic variation vary widely, depending on the populations and the forces that created these differences. Population structure is typically referred to as continental, referring to ancestry differences comparing across continents, and fine-scale, referring to ancestry differences

within a single continent. Continental structure was originally driven by humans migrating out of Africa and moving to new continents, leading to discrete populations on each continent. People with ancestry from multiple, discrete continental populations are referred to as admixed. Fine-scale structure is more often conceived of as continuous variation in allele frequencies across a continent, such as a cline, although fine-scale population structure can also be caused by physical barriers.

1.2.2 EIGENSTRAT

EIGENSTRAT is one of the most commonly used methods for inferring population structure from common variants. First published in 2006 by Price et al. [2], EIGENSTRAT summarizes the genotype data in a genetic relatedness matrix (GRM) then uses principal component analysis (PCA), which we discuss in the following section, to uncover structure in the GRM.

Suppose we have a sample of N unrelated subjects, indexed by $n \in 1 \dots N$, for which we have genotype data from M loci. Let $m \in 1 \dots M$ indicate the mth locus, let g_{im} be the number of copies subject i has of the minor allele at locus m, and let $\hat{p}_m = \frac{1}{2N} \sum_{n=1}^{N} g_{nm}$ be the observed (sample) frequency of the minor allele at locus m. With EIGENSTRAT, the entry for the ith and jth subjects in the GRM is

$$\tilde{\psi}_{ij} = \frac{1}{M} \sum_{m=1}^{M} \frac{(g_{im} - 2\hat{p}_m)(g_{jm} - 2\hat{p}_m)}{2\hat{p}_m(1 - \hat{p}_m)}. \tag{1.1}$$

The terms in this sum are the empirical pairwise correlations, under the assumption that subjects are sampled from a single homogenous population. If a population is composed of several subpopulations, with potentially admixed subjects, then this quantity will not be the true empirical pairwise correlation. We will refer to this quantity as the *empirical homogenous correlation*, to indicate that it is only truly a correlation under the assumption of homogeneity. When $\hat{p}_m$ is near zero or one, the denominator is approximately zero, and in turn, $\tilde{\psi}_{ij}$ is very large. Subjects with more copies of very rare alleles have relatively large GRM entries, and due to the well-known fact that PCA is sensitive to outliers, the principal components detect these subjects instead of population structure.

Furthermore, this problem cannot be detected by the proportion of variance explained by each principal component. When applied to rare variants, EIGENSTRAT can have a first principal component with a relatively high proportion of variance explained, yet the principal component can be completed uncorrelated with the true admixture (see Figure 2.4 and §2.3). To avoid the problems caused by rare variants, it is common to use EIGENSTRAT on a subset of loci which have a minor allele frequency (MAF) greater than a certain threshold, typically between 0.005 and 0.01.

1.2.3 PCA

Principal component analysis (PCA) is an exploratory data analysis technique that is used to visualize underlying structure in data sets with large numbers of variables. PCA takes a set of potentially correlated variables and transforms them into a new set of uncorrelated variables. Each set of observations in the original data set has a corresponding set of observations in the new variables. The new, uncorrelated variables are constructed so that they are linear combinations of the original variables such that the variance of the new variables is maximized. The new variables are referred to as the principal component scores, and the coefficients that relate the new variables to the old variables are referred to as the loadings. Principal component analysis (PCA) is typically applied to data that has been centered to have mean zero and scaled if the original variables are measured on different scales. Because PCA maximizes the variance of the uncorrelated variables, it is sensitive to outliers in the original data and to how the original variables are scaled.

The principal components of a data set can be found in several different ways. Geometrically, PCA is equivalent to projecting the original data onto a new set of axes such that the variance of the projected data is maximized, while the residuals are minimized. Therefore, the principal components can be found via solving the least-squares regression equation for each principal component subject to the maximized variance constraint. However, it is easier to find the principal components via the eigendecomposition of the empirical covariance matrix or the singular value decomposition (SVD) of the original matrix.

1.2.4 1000 Genomes Project

The 1000 Genomes Project was a world-wide effort to create a basic reference data set for the human genome [3]. This project collected genotype data from 2,504 people from 26 populations around the world. The geographic locations from which subjects were recruited were chosen to capture the breadth of variation between human populations. Subjects were recruited from the locations shown in the map below (Figure 1.2), and were grouped into 5 super-populations. These populations and their samples sizes are given in Table 1.1. The phase III data is whole genome sequence data for the complete sample.

Figure 1.2: Map of the 1000 Genome Project Populations [3]. Dots indicate from where the populations were recruited: African (yellow), American (red), East Asian (green), European (blue) and South Asian (purple).

Table 1.1: 1000 Genomes Project Population Samples Sizes

Super-population	Population	Sample Size
African (AFR)	ACB	123
	ASW	112
	ESN	173
	GWD	180
	LWK	116
	MSL	128
	YRI	186
American (AMR)	CLM	148
	MXL	107
	PEL	130
	PUR	150
East Asian (EAS)	CDX	109
	CHB	108
	CHS	171
	JPT	105
	KHV	124
Europeans (EUR)	CEU	183
	GBR	107
	FIN	125
	IBS	162
	TSI	112
South Asians (SAS)	BEB	144
	GIH	113
	ITU	118

Continued on next page

Table 1.1 – continued from previous page

Super-Population	Population	Sample Size
	PJL	158
	STU	128

Chapter 2

INFERRING POPULATION STRUCTURE WITH HUMAN GENOME SEQUENCE DATA

Inferring population structure is important in several areas of genetics, including population genetics, medical genetics, and personalized genomics. Advances in high-throughput sequencing technology have made large sequencing studies affordable and feasible. Sequencing data sets include large numbers of rare variants, which comprise 50-70% of the human genome [4]. This increased availability combined with the failure to fully explain heritability through common variants has led to increased interest in rare variant associating testing [5]. While there has been considerable work on methods for inferring population structure and association testing with common variants, most work on rare variants has focused on association testing.

A series of recent papers suggest that rare variant population structure confounds rare variant association testing, even after adjusting for common variant population structure. For the Genetic Analysis Workshop 17, several working groups analyzed a subset of the 1000 Genomes Project exon data and found inflated type 1 error, even after adjusting rare variant association tests for common variant population structure [6]. Mathieson and McVean [7] showed through simulations that fine scale spatial population structure can lead to inflated type 1 error rates for rare variants, and that EIGENSTRAT fails to adjust for this confounding. Similarly, O'Connor et al. [8] compared nine different rare variant association tests under fine-scale population structure and found spurious association rates as high as 40%, even after adjusting for common variant population structure. Finally, Liu et al. [9] show varying levels of inflated type 1 error in the C-alpha and burden tests due to population structure.

Furthermore, several recent papers demonstrate that rare variants may have different

population structure than common variants. Rare variants capture fine-scale, recent population structure that is not as easily captured by common variants [10]. Leslie et al. [11] have demonstrated that rare variants decompose fine-scale structure of the British Isles and continental European populations. Galinsky et al. [12] found that rare variants were more informative for population structure than common variants alone.

In this chapter, we develop a method for inferring population structure using sequence data that generalizes the EIGENSTRAT estimator. We review existing methods for common and rare variant population structure inference in the context of adjusting association tests. We then present our new method, PCA-seq, and demonstrate its benefits theoretically and through simulation studies. Finally, we demonstrate that PCA-seq is useful when applied to human genetics data by analyzing the phase III 1000 Genomes Project data.

2.1 Existing Methods for Inferring Population Structure

2.1.1 Common Variant Structure

Confounding in association studies due to population structure has been a well known problem since at least the mid-1990s, and methods for inferring population structure have been published since at least that time [13–15]. Since then, a variety of methods have been proposed for inferring common variant population structure in the context of adjusting association tests. These methods can be broadly grouped into two classes: model-based methods and machine-learning methods. Broadly, these classes represent a trade-off between models, which make more assumptions but yield scientifically interpretable information, and machine learning algorithms, which make fewer assumptions but yield less interpretable results.

Model-based Methods

Model-based methods include both frequentist and Bayesian methods, such as STRUCTURE, FRAPPE and ADMIXTURE. The common feature among these methods is their reliance on a likelihood for the observed genotype data which allows estimation of population-specific

allele frequencies and subject-specific admixture proportions.

STRUCTURE is the oldest of these methods, first proposed in 2000 [16]. Pritchard et al. developed STRUCTURE to cluster subjects into homogenous groups to address cryptic relatedness and confounding by population structure in common variant association testing. The grouping is done via a Bayesian clustering method, which assumes there are a fixed number of potentially unknown populations and that each population is characterized by a unique set of allele frequencies. Assuming that within each population the loci are in linkage equilibrium and Hardy-Weinburg equilibrium (HWE), STRUCTURE places a posterior distribution on the number of populations, the admixture proportions, and the allele frequencies within each population. Bayesian methods are used to estimate these parameters, which can then be used to cluster subjects and adjust for population structure.

FRAPPE and ADMIXTURE are related methods; both assume the same likelihood as STRUCTURE, but use frequentist methods to maximize the likelihood and estimate the population structure parameters [17]. Therefore, like STRUCTURE, FRAPPE and ADMIXTURE rely on the assumption that subjects have ancestry from a fixed number of discrete populations. This assumption does not necessarily hold with common variants, and with rare variants, it is unclear whether this assumption ever holds.

Machine-Learning Based Methods

There are two common machine-learning based methods: Multi-dimensional Scaling (MDS) and EIGENSTRAT. MDS maps the genotype data into a lower dimensional space while preserving the distances between subjects [18, 19]. While MDS is not inherently inappropriate for rare variants or sequence data, it is equivalent to EIGENSTRAT, the other machine-learning based method, when used with an Euclidean distance metric. In population structure inference, MDS is most often used with the Euclidean distance metric, as this has the same interpretation as EIGENSTRAT.

The most popular method for inferring population structure to adjust for association testing is EIGENSTRAT, which estimates the genome-wide average pairwise correlation be-

tween each subject's genome and then uses PCA to discover axes of variation [2]. These axes are often correlated with population structure when applied to common variants, although they can uncover other structure in the genotype data, such as batch effects. Unfortunately, EIGENSTRAT can perform poorly when trying to estimate population structure with rare variants. Therefore, it is common to remove rare variants before inferring population structure with EIGENSTRAT. We discuss the specifics of why this occurs in § 2.2.

2.1.2 Rare Variant Population Structure

While there is evidence that common variant methods do not accurately capture the population structure in rare variants [6–9], only two methods to infer population structure using rare variants have been proposed to date. Both of these methods, like EIGENSTRAT, apply PCA to a matrix that describes the pairwise average genetic similarity of subjects.

The first method, fineSTRUCTURE [20], uses a hidden Markov model (HMM) to estimate a coancestry matrix using haplotype data. For a given subject, the genome is split into small segments, and for each segment, an HMM model is used to identify another subject in the sample who is most closely related to the given subject at that segment. The corresponding row of the coancestry matrix for the given subject is the number of recombination events separating the given subject from each of the other subjects. This coancestry matrix is then standardized and PCA is applied. When the coancestry is calculated per locus, fineSTRUCTURE is asymptotically equivalent to EIGENSTRAT. Since this method requires phased data unless the coancestry is calculated per locus, this method is functionally equivalent to EIGENSTRAT for sequence data and performs similarly in simulations.

The second method, spectral dimension reduction (SDR) [21], applies PCA to a matrix of genetic similarity, $\boldsymbol{\Delta}$, with the form:

$$\boldsymbol{\Delta} = \boldsymbol{I} - \boldsymbol{T}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{T}^{-\frac{1}{2}}$$

where $\boldsymbol{B}$ is a symmetric matrix with entries $b_{ij} = \sqrt{2M\tilde{\psi}_{ij}}\mathbb{I}(\tilde{\psi}_{ij} > 0)$, M is the number of loci and $\tilde{\psi}_{ij}$ is the EIGENSTRAT GRM entry for subjects i and j. $\boldsymbol{T}$ is a diagonal matrix with

entries $t_{ii} = \sum_{n=1}^{N} b_{in}$. A single entry in the SDR genetic similarity matrix has the form

$$\begin{aligned}
\delta_{ij} &= \mathbb{I}(i=j) - \frac{b_{ij}}{\sqrt{t_{ii}t_{jj}}} \\
&= \mathbb{I}(i=j) - \frac{\sqrt{2M\tilde{\psi}_{ij}}\mathbb{I}(\tilde{\psi}_{ij} > 0)}{\sqrt{t_{ii}t_{jj}}} \\
&= \left(\frac{\mathbb{I}(i=j)}{\tilde{\psi}_{ij}} - \frac{\sqrt{2M}\mathbb{I}(\tilde{\psi}_{ij} > 0)}{\sqrt{t_{ii}t_{jj}\tilde{\psi}_{ij}}} \right) \tilde{\psi}_{ij}
\end{aligned}$$

The first term in the product weights the entries of the EIGENSTRAT GRM using a function of the genotype data (for full derivation, see Appendix A.1). Given that the EIGENSTRAT GRM estimator tends towards infinity when applied to rare variants (see §2.2), this weighting function would need to counteract this tendency. In simulations from the SDR paper, the type I error was inflated after adjusting for population structure estimated with SDR using only rare variants, which suggests that this weight does not deflate large GRM entries enough. Furthermore, SDR captured continental structure with rare and common variants, but did not capture within-continent structure using only rare variants from the 1000 Genomes Project data.

2.2 PCA-seq

As an alternative to the EIGENSTRAT estimator, we propose a weighted estimator for constructing the GRM:

$$\hat{\psi}_{ij} = \frac{1}{M} \sum_{m=1}^{M} w_m (g_{im} - 2\hat{p}_m)(g_{jm} - 2\hat{p}_m), \tag{2.1}$$

where w_m is a locus-specific weight. The benefit of this estimator is that for appropriately chosen weights, we need not exclude rare variants.

There are many possible choices for the weights. For the current discussion, we focus on

the flexible family of weights defined by the Beta distribution:

$$\begin{aligned}\sqrt{w_m} &= Beta(\hat{p}_m; \alpha, \beta) \\ &= \frac{1}{B(\alpha, \beta)} \hat{p}_m^{\alpha-1} (1 - \hat{p}_m)^{\beta-1},\end{aligned}$$

where the square root is taken for computational convenience. Several special cases exist that are worth noting. If $\alpha = 0.5$ and $\beta = 0.5$, then the weights are proportional to the Madsen and Browning weights [22] and the PCA-seq estimator is equivalent to the EIGENSTRAT estimator. If $\alpha = 1$ and $\beta = 1$, then the weights are uniform and all loci are weighted equally, making the GRM the empirical covariance matrix. The PCA-seq estimator with uniform weights is biased for the coancestry, or correlation between subjects due to shared population structure, but in the limit, has lower variance than the EIGENSTRAT estimator (see §2.2.1). Alternatively, we can consider all of the loci at once but give varying weight based on MAF, such as the Wu weights [23] which give greater weight to rare variants than common variants. Such a weighting scheme would not require us to subset the loci by MAF, which could be beneficial if we expect population structure to vary smoothly by MAF.

Figure 2.1 shows the weights by MAF for the EIGENSTRAT, uniform and Wu weights. At higher MAF, all three weighting schemes are similar and put approximately the same weight on loci. However, EIGENSTRAT places greater weight on the variants with the lowest MAFs, while the Wu weights place more weight on variants with MAFs between 0 and 0.1.

We can motivate this generalized estimator by considering the values each estimator takes under different genotypes and allele frequencies. Specifically, we compare the EIGENSTRAT estimator ($\alpha = 0.5$, $\beta = 0.5$) to the uniform estimator ($\alpha = 1$, $\beta = 1$). The uniform estimator is an extreme case as it treats all loci as equally informative for the population structure of interest. We expect any weights that focus on loci associated with the structure of interest to perform better than the uniform estimator.

Figure 2.2 shows the value of each estimator at a single locus across a range of MAFs (0–0.5) under certain genotypes for two individuals. We take the minor allele to be the one we are counting, so that a genotype of 2 indicates homozygosity for the minor allele. When

both alleles are relatively common (Figures 4.1a and 4.1b), the two weights are similar, as are their contributions to the GRM. If both subjects share one copy of the minor allele, then under both weights, a locus with a smaller MAF contributes more to the GRM than a locus with a larger MAF. Similarly, if only one subject is heterozygous for the minor allele, under both weights, a locus with a smaller MAF contributes evidence of dissimilar ancestry (negative GRM contribution). However, if the shared allele is rare (Figures 2.2c and 2.2d), then a locus with a very small MAF receives more weight than a locus with a larger MAF under EIGENSTRAT. When both subjects are heterozygous for a very rare minor allele, this effect is very pronounced, as the EIGENSTRAT estimator tends to infinity (as noted previously). This excessive weighting of loci with very low MAFs under the EIGENSTRAT estimator causes subjects with very rare alleles to be outliers relative to the rest of the sample, which in turn causes PCA to fail to detect population structure.

2.2.1 Properties of PCA-*seq*

In this section, we demonstrate that the PCA-seq estimator is a first order Taylor approximation of the genome-wide average empirical homogenous correlation. Under the assumption of discrete populations and independent loci, we demonstrate that the EIGENSTRAT estimator is unbiased for the coancestry and that the PCA-seq estimator is biased. However, we demonstrate that that the PCA-seq estimator, for sufficiently large sample sizes, has smaller variance than the EIGENSTRAT estimator.

Theorem 2.2.1. *The* PCA-*seq estimator is a first-order Taylor approximation of the genome-wide empirical homogenous correlation, and this approximation is exact when the covariance between the* PCA-*seq estimator and the genome-wide average variance assuming* HWE *is zero.*

Proof. We rewrite the PCA-seq estimator with uniform weights as

$$\hat{\psi}_{ij} = \frac{\frac{1}{M}\sum_{m=1}^{M}(g_{im} - 2\hat{p}_m)(g_{jm} - 2\hat{p}_m)}{\frac{1}{M}\sum_{m=1}^{M}\hat{p}_m(1-\hat{p}_m)},$$

which is equivalent to the form given previously, up to the multiplicative constant in the denominator $\frac{1}{M}\sum_{m=1}^{M}\hat{p}_m(1-\hat{p}_m)$. Note that PCA is invariant to a constant scale factor applied to all loci.

Define T_M to be the numerator of $\hat{\psi}_{ij}$ and B_M to be the denominator. Then $\hat{\psi}_{ij}$ is equivalent to $h(T_M, B_M) = T_M/B_M$. Furthermore, we note that B_M cannot be zero unless $\hat{p}_m = 0$ for all loci, so the ratio is well-defined. We now approximate h around the point $(E[T_M], E[B_M])$ using a first order Taylor series:

$$\begin{aligned} h(t,b) = h(E[T_M], E[B_M]) + h'_t(\theta)(t - E[T_M]) + h'_b(\theta)(b - E(B_M)) \\ + O(|t - E(T_M)|^2) + O(|b - E(B_M)|^2). \end{aligned}$$

If we take the expectation of both sides of the equation above with respect to the genotype, the last two terms are $O(\mathrm{Var}[T_M])$ and $O(\mathrm{Var}[B_M])$. Since both T_M and B_M are averages, their variances are $\sigma^2_{T_M}/M$ and $\sigma^2_{B_M}/M$, both of which converge to zero as $M \to \infty$. And since the expectation of $t - E(T_M)$ and $b - E(B_M)$ are both zero,

$$\mathrm{E}[h(t,b)] \approx h(E[T_M], E[B_M]).$$

Noting that $\mathrm{E}[h(t,b)]$ is the genome-wide empirical homogeneous correlation, we have

$$\mathrm{E}[\hat{\psi}_{ij}] \approx \frac{\mathrm{E}[T_m]}{\mathrm{E}[B_M]}.$$

Therefore, PCA-seq is a first order Taylor approximation to the genome-wide empirical homogeneous correlation.

This approximation will be exact when $\mathrm{Cov}(T_M/B_M, B_M) = 0$ as,

$$\begin{aligned} \mathrm{Cov}\left(\frac{\bar{T}_M}{\bar{B}_M}, \bar{B}_M\right) &= \mathrm{E}\left(\frac{\bar{T}_M}{\bar{B}_M}\bar{B}_M\right) - \mathrm{E}\left(\frac{\bar{T}_M}{\bar{B}_M}\right)\mathrm{E}(\bar{B}_M) \\ &= \mathrm{E}(\bar{T}_M) - \mathrm{E}\left(\frac{\bar{T}_M}{\bar{B}_M}\right)E(\bar{B}_M) \end{aligned}$$

which implies if the covariance is zero,

$$\mathrm{E}\left(\frac{\bar{T}_M}{\bar{B}_M}\right) = \frac{E(\bar{T}_M)}{E(\bar{B}_M)}.$$

□

Theorem 2.2.1 illustrates that the PCA-seq and EIGENSTRAT estimators are both estimating the same population parameter, but are derived in different ways. In the next two theorems, we show why that the PCA-seq estimator is the more desirable estimator.

Theorem 2.2.2. *Assuming that the loci are independent, the true allele frequencies are known, and the subjects are unrelated, the* EIGENSTRAT *estimator is unbiased for the coancestry or correlation due to population structure between the genotypes, θ_{ij}.*

$$\tilde{\psi}_{ij} \to \theta_{ij}$$

The PCA*-seq estimator is biased for the coancestry θ_{ij}, but we can bound the bias:*

1. *for uniform weights ($w_m = 1$)*

$$\hat{\psi}_{ij} \to \tilde{\theta}_{ij} \leq \theta_{ij}$$

2. *for non-negative, finite weights ($w_m \in [0, w^{(M)}]$):*

$$\hat{\psi}_{ij} \to \tilde{\theta}_{ij} \leq w^{(M)}\theta_{ij}$$

Proof. Consider a set of N subjects, indexed by $n \in \{1 \dots N\}$, with ancestry from S populations. Let a_n denote the ancestry of the nth subject, where $a_n = [a_1 \dots a_S]$ is a vector of ancestry proportions that sum to one (i.e., $\sum_{s=1}^{S} a_s = 1$). At each of M loci, indexed by $m \in \{1 \dots M\}$, let $\boldsymbol{p}_m = [p_1 \dots p_S]$ be the vector of population-specific allele frequencies. We treat the ancestry proportions as fixed, but allow the allele frequencies to be random, with

$$\begin{aligned} \mathrm{E}(\boldsymbol{p}_m) &= p_m \mathbf{1} \\ \mathrm{Cov}(\boldsymbol{p}_m) &= p_m(1 - p_m)\boldsymbol{\Theta}_S \end{aligned}$$

where this holds for all m. Under this assumption, each subject has a subject-specific allele frequency, defined as the expected allele frequency for subject n at locus m: $\mu_{nm} = \boldsymbol{a}_n \boldsymbol{p}_m$.

This subject-specific allele frequency has expectation $\mathrm{E}(\mu_{nm}) = p_s$ and

$$\begin{aligned}
\mathrm{E}(\mu_{im}\mu_{jm}) &= \sum_{s=1}^{S}\sum_{s'=1}^{S} \boldsymbol{a}_i^s \boldsymbol{a}_j^{s'} \mathrm{E}(p_m^s p_m^{s'}) \\
&= \sum_{s=1}^{S}\sum_{s'=1}^{S} \boldsymbol{a}_i^s \boldsymbol{a}_j^{s'} \left[(p_m)^2 + \mathrm{E}(p_m^s p_m^{s'}) - (p_m)^2\right] \\
&= \sum_{s=1}^{S}\sum_{s'=1}^{S} \boldsymbol{a}_i^s \boldsymbol{a}_j^{s'} (p_m)^2 + \sum_{s=1}^{S}\sum_{s'=1}^{S} \boldsymbol{a}_i^s \boldsymbol{a}_j^{s'} \mathrm{E}(p_m^s p_m^{s'} - (p_m)^2) \\
&= (p_m)^2 \sum_{s=1}^{S} \boldsymbol{a}_i^s \sum_{s'=1}^{S} \boldsymbol{a}_j^{s'} + \sum_{s=1}^{S}\sum_{s'=1}^{S} \boldsymbol{a}_i^s \boldsymbol{a}_j^{s'} \mathrm{E}(p_m^s p_m^{s'} - (p_m)^2) \\
&= (p_m)^2 + \sum_{s=1}^{S}\sum_{s'=1}^{S} \boldsymbol{a}_i^s \boldsymbol{a}_j^{s'} \mathrm{Cov}(p_m^s, p_m^{s'}) \\
&= (p_m)^2 + p_m(1-p_m)\theta_{ij}
\end{aligned}$$

where $\theta_{ij} = \boldsymbol{a}_i^{\mathsf{T}} \boldsymbol{\Theta}_S \boldsymbol{a}_j$ is the coancestry coefficient due to population structure for the pair of subjects i and j. The coancestry is the correlation between a random pair of alleles from a specific subpopulation, relative to the total population.

Let Y_{ij} be the set of most recent common ancestors for the pair of subjects i and j. This set includes more than one person if i and j have multiple ancestors with the same familial relationship (i.e. two parents if i and j are siblings). Lest ℓ_{ij} be the length of the pedigree path from subject i to subject j via their most recent common ancestor. Define $\ell_{ii} = 1$. Then the kinship coefficient can be written in terms on the path lengths:

$$\begin{aligned}
\phi_{ij} &= \sum_{y \in Y} \left(\frac{1}{2}\right)^{\ell_{iy}+\ell_{jy}-1} (1+f_y) \\
&= \sum_{y \in Y} \phi_{ij|y}
\end{aligned}$$

Let x_{im_r} be the indicator that subject i has the reference allele at their allele copy $r \in$

$\{1, 2\}$ at locus m. Then their genotype at this locus, $g_{im} = x_{im_1} + x_{im_2}$ and

$$\begin{aligned} \mathrm{E}(g_{im} g_{jm} | \boldsymbol{p}_s) &= 4\mathrm{E}(x_{im_r} x_{jm_{r'}} | \boldsymbol{p}_s) \\ &= 4 \sum_{y \in Y_{ij}} \left[(\phi_{ij|y}) \mu_{ym} (1 - \mu_{ym}) \right] + 4 \mu_{im} \mu_{jm}. \end{aligned}$$

If we take the expectation of this quantity with respect to $\boldsymbol{p}_m$, then we have

$$\mathrm{E}(g_{im} g_{jm}) = 4(p_m)^2 + 4p_m(1 - p_m) \left[\phi_{ij} + \theta_{ij} - \sum_{y \in Y_{ij}} \phi_{ij|y} \theta_{yy} \right]$$

where $\theta_{yy} = \boldsymbol{a}_y^{\mathsf{T}} \boldsymbol{\Theta}_S \boldsymbol{a}_y$.

We can now derive the limiting expectation of the EIGENSTRAT and PCA-seq estimators. We make several assumptions:

1. The true ancestral population allele frequencies and subject-specific allele frequencies are known.

2. The ancestral population allele frequencies are independent and identically distributed, although we need not know the distribution.

3. Genotypes across loci are independent.

Under these assumptions, the EIGENSTRAT estimator will converge to its expectation.

Therefore,

$$\begin{aligned}
\mathrm{E}(\tilde{\psi}_{ij}) &= \mathrm{E}\left[\frac{1}{M}\sum_{m=1}^{M}\frac{(g_{im}-2p_m)(g_{jm}-2p_m)}{2p_m(1-p_m)}\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\mathrm{E}\left[\frac{(g_{im}-2p_m)(g_{jm}-2p_m)}{2p_m(1-p_m)}\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\frac{1}{2p_m(1-p_m)}\mathrm{E}\left[(g_{im}-2p_m)(g_{jm}-2p_m)\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\frac{1}{2p_m(1-p_m)}\mathrm{E}\left[g_{im}g_{jm}-2p_mg_{im}-2p_mg_{jm}+4(p_m)^2\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\frac{1}{2p_m(1-p_m)}\left[\mathrm{E}(g_{im}g_{jm})-\mathrm{E}(2p_mg_{im})-\mathrm{E}(2p_mg_{jm})+\mathrm{E}(4(p_m)^2)\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\frac{1}{2p_m(1-p_m)}\left[\mathrm{E}(g_{im}g_{jm})-2p_m\mathrm{E}(g_{im})-2p_m\mathrm{E}(g_{jm})+4(p_m)^2\right]
\end{aligned}$$

Using the expectations derived previously, we have

$$\begin{aligned}
\mathrm{E}(\tilde{\psi}_{ij}) &= \frac{1}{M}\sum_{m=1}^{M}\frac{1}{2p_m(1-p_m)}4(p_m)^2+4p_m(1-p_m)[\phi_{ij}+\theta_{ij} \\
&\quad -\sum_{y\in Y_{ij}}\phi_{ij|y}\theta_{yy}]-4(p_m)^2-4(p_m)^2+4(p_m)^2 \\
&= \frac{1}{M}\sum_{m=1}^{M}\frac{1}{2p_m(1-p_m)}4p_m(1-p_m)\left[\phi_{ij}+\theta_{ij}-\sum_{y\in Y_{ij}}\phi_{ij|y}\theta_{yy}\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\phi_{ij}+\theta_{ij}-\sum_{y\in Y_{ij}}\phi_{ij|y}\theta_{yy} \\
&= \phi_{ij}+\theta_{ij}-\sum_{y\in Y_{ij}}\phi_{ij|y}\theta_{yy}
\end{aligned}$$

where we have removed the summation in the last line, as each of the quantities is constant with respect to locus. Since we are interested in estimating the coancestry between subjects i and j, the other two terms are bias. If i and j are unrelated, then $\tilde{\psi}_{ij} \to \theta_{ij}$, unless there is no population structure (i.e. i and j are from a single homogeneous population).

We preform a similar derivation for the PCA-seq estimator. Under these assumptions, the PCA-seq estimator will converge to its expectation. Therefore, treating the weights as known functions of the allele frequencies:

$$\begin{aligned}
\mathrm{E}(\hat{\psi}_{ij}) &= \mathrm{E}\left[\frac{1}{M}\sum_{m=1}^{M} w_m(g_{im}-2p_m)(g_{jm}-2p_m)\right] \\
&= \frac{1}{M}\sum_{m=1}^{M} \mathrm{E}\left[w_m(g_{im}-2p_m)(g_{jm}-2p_m)\right] \\
&= \frac{1}{M}\sum_{m=1}^{M} w_m\mathrm{E}\left[(g_{im}-2p_m)(g_{jm}-2p_m)\right] \\
&= \frac{1}{M}\sum_{m=1}^{M} w_m\mathrm{E}\left[g_{im}g_{jm}-2p_mg_{im}-2p_mg_{jm}+4(p_m)^2\right] \\
&= \frac{1}{M}\sum_{m=1}^{M} w_m\left[\mathrm{E}(g_{im}g_{jm})-\mathrm{E}(2p_mg_{im})-\mathrm{E}(2p_mg_{jm})+\mathrm{E}(4(p_m)^2)\right] \\
&= \frac{1}{M}\sum_{m=1}^{M} w_m\left[\mathrm{E}(g_{im}g_{jm})-2p_m\mathrm{E}(g_{im})-2p_m\mathrm{E}(g_{jm})+4(p_m)^2\right]
\end{aligned}$$

Using the expectations derived previously, we have

$$\begin{aligned}
\mathrm{E}(\hat{\psi}_{ij}) &= \frac{1}{M}\sum_{m=1}^{M} w_m 4(p_m)^2 + 4p_m(1-p_m)[\phi_{ij}+\theta_{ij} \\
&\quad -\sum_{y\in Y_{ij}}\phi_{ij|y}\theta_{yy}] - 4(p_m)^2 - 4(p_m)^2 + 4(p_m)^2 \\
&= \frac{1}{M}\sum_{m=1}^{M} w_m 4p_m(1-p_m)\left[\phi_{ij}+\theta_{ij}-\sum_{y\in Y_{ij}}\phi_{ij|y}\theta_{yy}\right] \\
&= \frac{1}{M}\sum_{m=1}^{M} w_m 4p_m(1-p_m)\left[\phi_{ij}+\theta_{ij}-\sum_{y\in Y_{ij}}\phi_{ij|y}\theta_{yy}\right] \\
&= \frac{1}{M}\left[\phi_{ij}+\theta_{ij}-\sum_{y\in Y_{ij}}\phi_{ij|y}\theta_{yy}\right]\sum_{m=1}^{M} 4p_m(1-p_m)w_m
\end{aligned}$$

We can consider the effect of various weights on this estimator. First, we consider the effect

of uniform weights. If $w_m = 1$, then

$$\begin{aligned}
\mathrm{E}(\hat{\psi}_{ij}) &= \frac{1}{M}\left[\phi_{ij} + \theta_{ij} - \sum_{y \in Y_{ij}} \phi_{ij|y}\theta_{yy}\right] \sum_{m=1}^{M} 4p_m(1-p_m) \\
&\leq \frac{1}{M}\left[\phi_{ij} + \theta_{ij} - \sum_{y \in Y_{ij}} \phi_{ij|y}\theta_{yy}\right] \sum_{m=1}^{M} 1 \\
&= \phi_{ij} + \theta_{ij} - \sum_{y \in Y_{ij}} \phi_{ij|y}\theta_{yy}
\end{aligned}$$

which suggests that the PCA-seq estimator with uniform weights has an expectation in the limit that is equal to or less than that of EIGENSTRAT. These two estimators will be equal if every locus has $p_m = 0.5$ (as in this case the two estimators are equivalent). In general, the PCA-seq estimator is smaller than the EIGENSTRAT estimator. When the two subjects are unrelated, the coancestry estimates will be smaller from PCA-seq than from EIGENSTRAT.

Now we consider the more general case where $w_m \in [0, w^{(M)}]$, where $w^{(M)}$ is the maximum weight. In this case, we have

$$\begin{aligned}
\mathrm{E}(\hat{\psi}_{ij}) &= \frac{1}{M}\left[\phi_{ij} + \theta_{ij} - \sum_{y \in Y_{ij}} \phi_{ij|y}\theta_{yy}\right] \sum_{m=1}^{M} 4p_m(1-p_m)w_m \\
&\leq \frac{1}{M}\left[\phi_{ij} + \theta_{ij} - \sum_{y \in Y_{ij}} \phi_{ij|y}\theta_{yy}\right] \sum_{m=1}^{M} w^{(M)} \\
&= w^{(M)}\left[\phi_{ij} + \theta_{ij} - \sum_{y \in Y_{ij}} \phi_{ij|y}\theta_{yy}\right]
\end{aligned}$$

This estimate is biased, and may be larger than the estimate from EIGENSTRAT, if $w^{(M)}$ is greater than 1. In the case where $w^{(M)}$ is 1, we have the same results as the uniform weights.

□

Theorem 2.2.2 elucidates several key properties of the PCA-seq estimator. Relative to EIGENSTRAT, the PCA-seq estimator is biased. As the bound on the bias indicates, the weights chosen have a strong effect on the bias. When the maximum weight is extremely

large, the PCA-seq estimator may have large bias. Under uniform weights, the PCA-seq estimator is always biased downwards, relative to the EIGENSTRAT estimator.

We demonstrate that the PCA-seq estimator with uniform weights has a variance that is smaller than the variance of the EIGENSTRAT estimator for a sufficiently large sample size.

Theorem 2.2.3. *The variance of the* PCA*-seq estimator with uniform weights is less than that of the* EIGENSTRAT *estimator with a sufficiently large sample size.*

Proof. As before, we assume that the true ancestral population allele frequencies are known and that they are independent and identically distributed. We also assume that the genotypes are independent across loci. We begin by deriving the variance of the EIGENSTRAT estimator.

$$\begin{aligned}
\mathrm{Var}(\tilde{\psi}_{ij}) &= \mathrm{Var}\left[\frac{1}{M}\sum_{m=1}^{M}\frac{(g_{im}-2p_m)(g_{jm}-2p_m)}{2p_m(1-p_m)}\right] \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\mathrm{Var}\left[\frac{(g_{im}-2h_m)(g_{jm}-2p_m)}{2p_m(1-p_m)}\right] \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\frac{1}{4p_m^2(1-p_m)^2}\mathrm{Var}\left[(g_{im}-2p_m)(g_{jm}-2p_m)\right]
\end{aligned}$$

Since p_m is a probability, $p_m^2(1-p_m)^2$ achieves its maximum of $\frac{1}{16}$ at $p_m = 0.5$. This implies that $\frac{1}{p_m^2(1-p_m)^2}$ is greater than or equal to 16. Therefore, the above is bounded from below by

$$\begin{aligned}
\mathrm{Var}(\tilde{\psi}_{ij}) &= \frac{1}{M^2}\sum_{m=1}^{M}\frac{1}{4p_m^2(1-p_m)^2}\mathrm{Var}\left[(g_{im}-2p_m)(g_{jm}-2p_m)\right] \\
&> \frac{4}{M^2}\sum_{m=1}^{M}\mathrm{Var}\left[(g_{im}-2p_m)(g_{jm}-2p_m)\right]
\end{aligned}$$

Now we consider the PCA-seq estimator with uniform weights.

$$\begin{aligned}
\mathrm{Var}(\hat{\psi}_{ij}) &= \mathrm{Var}\left[\frac{1}{M}\sum_{m=1}^{M}(g_{im}-2p_m)(g_{jm}-2p_m)\right] \\
&= \frac{1}{M^2}\sum_{m=1}^{M}\mathrm{Var}\left[(g_{im}-2p_m)(g_{jm}-2p_m)\right]
\end{aligned}$$

Hence, we have the inequality we desire.

$$\begin{aligned}
\mathrm{Var}(\tilde{\psi}_{ij}) &= \frac{1}{M^2}\sum_{m=1}^{M}\mathrm{Var}\left[(g_{im}-2p_m)(g_{jm}-2p_m)\right] \\
&< \frac{4}{M^2}\sum_{m=1}^{M}\mathrm{Var}\left[(g_{im}-2p_m)(g_{jm}-2p_m)\right] \\
&< \mathrm{Var}(\hat{\psi}_{ij})
\end{aligned}$$

□

Theorem 2.2.3 demonstrates that the PCA-seq estimator with uniform weights will have lower variance than the EIGENSTRAT estimator. As M increase, the two variances will become more and more similar, as the constant multiplicative factor is dominated by the summation and the M^2 term; however, PCA-seq will always have the smaller variance. Together Theorems 2.2.2 and 2.2.3 suggest that the PCA-seq estimator with uniform weights will have better mean-squared error than the EIGENSTRAT estimator when the number of loci is sufficiently large. Given that sequence data includes millions of loci, we expect the variance to dominate the mean-squared error, indicating that PCA-seq is the better estimator for sequence data.

2.3 Simulations

We performed several simulations to demonstrate the performance of PCA-seq when applied to rare and common variants under a range of population structures (F_{st}), samples sizes, and rare variant definitions (MAF). In all the simulations, we fixed the proportion of rare variants as a percent of the total sample size. We simulated genotype data from individuals that were admixed between two ancestral populations with allele frequencies drawn from Balding-Nichols model [24] under F_{st} of 0.01 to 0.15. In these simulations, the population structure was the same for both the rare and common variants. Simulation results are presented as the average and empirical 95% confidence interval of the correlation (r^2) between

the true admixture proportion and the first principle component from either EIGENSTRAT or PCA-seq.

To demonstrate that PCA-seq is equivalent to EIGENSTRAT for common variants (MAF $>$ 0.025) and better than EIGENSTRAT for inferring population structure with rare (MAF $\leq$ 0.025) variants, we simulated 200 unrelated individuals who were admixed between two populations with frequencies simulated under F_{st} ranging from 0.01 to 0.15. We simulated 10,000 loci, of which 60% were rare variants. Figure 2.3 shows the average and empirical 95% confidence interval for the correlation between the true admixture proportion and the first principal component from EIGENSTRAT and PCA-seq with uniform weights. Using only the common variants to infer population structure, PCA-seq performs comparably to EIGENSTRAT, particularly for large F_{st} values. Both methods have high correlation between the true population structure and the first principal component. However, with rare variants, PCA-seq consistently has a higher average correlation. Furthermore, the variation in the correlation is considerably smaller for PCA-seq. In some simulations, there is almost no correlation between the first principal component from EIGENSTRAT and the true admixture.

We also looked at the correlation between the r^2 values from these simulations and proportion of variance explained by the first principal component. Figure 2.4 shows these results. The correlation displayed on the plot is the overall correlation of the points in the plot. Both EIGENSTRAT and PCA-seq when applied to common variants have relatively high correlation between the r^2 and the percent of variation explained for the first principal component ($\rho \approx 0.8$). However, when EIGENSTRAT is applied to rare variants, there are several data sets in which the proportion of variance explained by the first principal component is relative high, even though the r^2 value is small. This indicates that the proportion of variance explained is not a good indicator of whether the principal components from EIGENSTRAT reflect outliers in the data or true structure. However, the proportion of variance explained by the first principal component from PCA-seq applied to rare variants is highly correlated with the r^2 value, although not quite as highly correlated as when PCA-seq is applied to common variants.

In our second set of simulations, we again simulated unrelated subjects from two populations, but we only simulated rare variants with MAFs ≤ 0.01. We fixed the F_{st} at 0.01, and varied the number of loci from 6,000 to 60,000 for sample sizes of 200, 500, and 1,000. In these simulations, we aimed to understand the effect of increasing the number of samples and loci on each method's ability to infer population structure. This is important, as sequencing the human genome gives us an abundance of loci, while the cost of sequencing studies limits the number of samples available.

Figure 2.5 shows the results of these simulations. The average correlation with the true admixture proportion increases with increasing numbers of rare variants for all 3 sample sizes. With 200 subjects, PCA-seq with uniform weights consistently has higher average r^2 and significantly less variation in the correlation between the first principal component and the true admixture, even with 100,000 loci. However, even with less than 15,000 loci, PCA-seq does have very high average correlation between the first principal component and the true admixture. With 500 unrelated subjects, both methods have higher average correlation, although PCA-seq still performs better on average and has less variable r^2. For the largest sample size, the two methods are very similar, with at least 15,000 loci. Although with less than 15,00 loci, even with 1000 subjects, EIGENSTRAT still has considerably more variability than PCA-seq with uniform weights. These results suggest that even with relatively small numbers of subjects, with enough loci, PCA-seq can be used to estimate population structure well. However, adequate numbers of subjects are needed to estimate structure when very little genetic data is available.

In the results from the first set of simulations, we found that in some cases, the first principal component from EIGENSTRAT had little to no correlation with the true admixture proportions, while PCA-seq had relatively high correlation with the true admixture proportions. We selected a few of the simulation data sets where this occurred, and found that in each data set, there was one subject with a unique haplotype. That is, one subject had a unique set of single nucleotide variants (SNVs), which were highly associated with the first principal component from EIGENSTRAT. In our final simulation, we attempted to replicate

this scenario, as this mimics the effects of genotyping errors or subjects with unique ancestry, relative to the rest of the sample.

We simulated 1,000 data sets from 200 unrelated individuals who were admixed between 2 populations with frequencies simulated under F_{st} values of 0.01 to 0.15. We simulated 10,000 loci in each data, where 75% of the loci were rare variants (MAF ≤ 0.025). We then created a subject with 20 SNVs that were unrelated to ancestry by randomly selecting a subject and 20 monomorphic loci. For this subject, we changed their genotype data to indicate a single copy of the previously missing allele at those twenty loci. The results of these simulations are shown in Figure 2.6. Compared to the results in Figure 2.3, both methods have lower average correlation, although PCA-seq has a smaller decrease in average correlation and a smaller increase in variability. EIGENSTRAT has a significant decrease in average correlation, as in some cases, the first principal component from EIGENSTRAT is uncorrelated with the true admixture. This demonstrates that PCA-seq is robust to genotyping errors and outliers in rare variants, while EIGENSTRAT is not.

2.4 Application to 1,000 Genomes Project Data

To demonstrate the utility of PCA-seq, we analyzed the Phase III data from the 1,000 Genomes Project [3]. The data was subset into the 5 super-populations: Africans, Americans, East Asians, Europeans, and South Asians. For each super-population, we subset the data into rare variants, those with MAF ≤ 0.05 in the global sample, and common variants, those with MAF > 0.5. Before running EIGENSTRAT or PCA-seq, we removed monomorphic loci. We did not filter by linkage disequilibrium. Both EIGENSTRAT and PCA-seq with uniform weights were applied to the rare and common variant datasets from each super-population. The results of this analysis are shown as plots of the first two principal components from each method, and as parallel coordinate plots for the first ten principal components from each method. We present the results for the African super-population; results for the other super-populations can be found in Appendix A.2.

Figure 2.7 shows the first two principal components from analyzing the rare (MAF ≤ 0.05)

and common (MAF > 0.05) variants from the African super-population separately using EIGENSTRAT and PCA-seq with uniform weights. When EIGENSTRAT and PCA-seq are applied to common variants, the results are very similar. The first principal component separates the African populations along an East-West cline, while the second principal component describes the non-African ancestry of the African Caribbeans and African Americans. Both EIGENSTRAT and PCA-seq capture continental level population structure when applied to common variants.

When EIGENSTRAT is applied to rare variants, we see a similar pattern as with the common variants, although the second principal component primarily separates one African American subject from the rest of the African American subjects. When PCA-seq with uniform weights is applied to the rare variants, the first two principal components decompose the African populations into three distinct groups which reflect the geographic clustering of these populations: (1) the Kenyans (LWK), (2) the Nigerians (ESN, YRI), and (3) the Sierra Leonians and Gambians (MSL, GWD). Furthermore, the African Caribbeans and African Americans show admixture between these three groups. PCA-seq with uniform weights has captured fine-scale, within-continent structure in rare variants that is not captured by EIGENSTRAT when applied to rare or common variants.

Figure 2.8 shows parallel coordinate plots for the first 10 principal components from EIGENSTRAT and PCA-seq applied to rare and common variants. The parallel coordinate plots show many of the same features we saw in Figure 2.7. EIGENSTRAT and PCA-seq are very similar when applied to common variants. For both methods, principal components beyond the fourth or fifth do not delineate population structure, and instead higher principal components separate subjects that are outliers from the rest of the sample. The higher principal components do not describe fine-scale population structure.

If we repeat the above analysis, but exclude the two admixed populations (African Caribbeans from Barbados (ACB), Americans of African Ancestry in the Southwest United States (ASW)), we see that EIGENSTRAT and PCA-seq uncover the fine-scale population structure when applied to common variants. EIGENSTRAT applied to rare variants still detects

subjects who are outliers, although the effect is much less pronounced. This suggests that in the presence of admixed populations, PCA-seq may be the better estimator for uncovering within-continent population structure.

2.5 Conclusion

In this chapter, we have shown that PCA-seq, which generalizes EIGENSTRAT by allowing for arbitrary weighting of loci outperforms EIGENSTRAT when applied to both rare and common variants. Furthermore, we have shown that under appropriate weights, PCA-seq is robust to outliers, whether due to genotyping errors or due to subjects with substantially different ancestry compared to the rest of the sample. Finally, we have shown that the usual metric for assessing the quality of principal components, percent of variation explained, fails to indicate that EIGENSTRAT has not captured population structure with rare variants.

PCA-seq is a broadly applicable method which can be used anywhere EIGENSTRAT is currently used. Just as rare variant association testing has been extended to incorporate common variant population structure and relatedness, we could extend rare variant association testing to incorporate rare variant population structure, either through a random effect with a covariance matrix equal to the GRM or using the principal components as fixed effects.

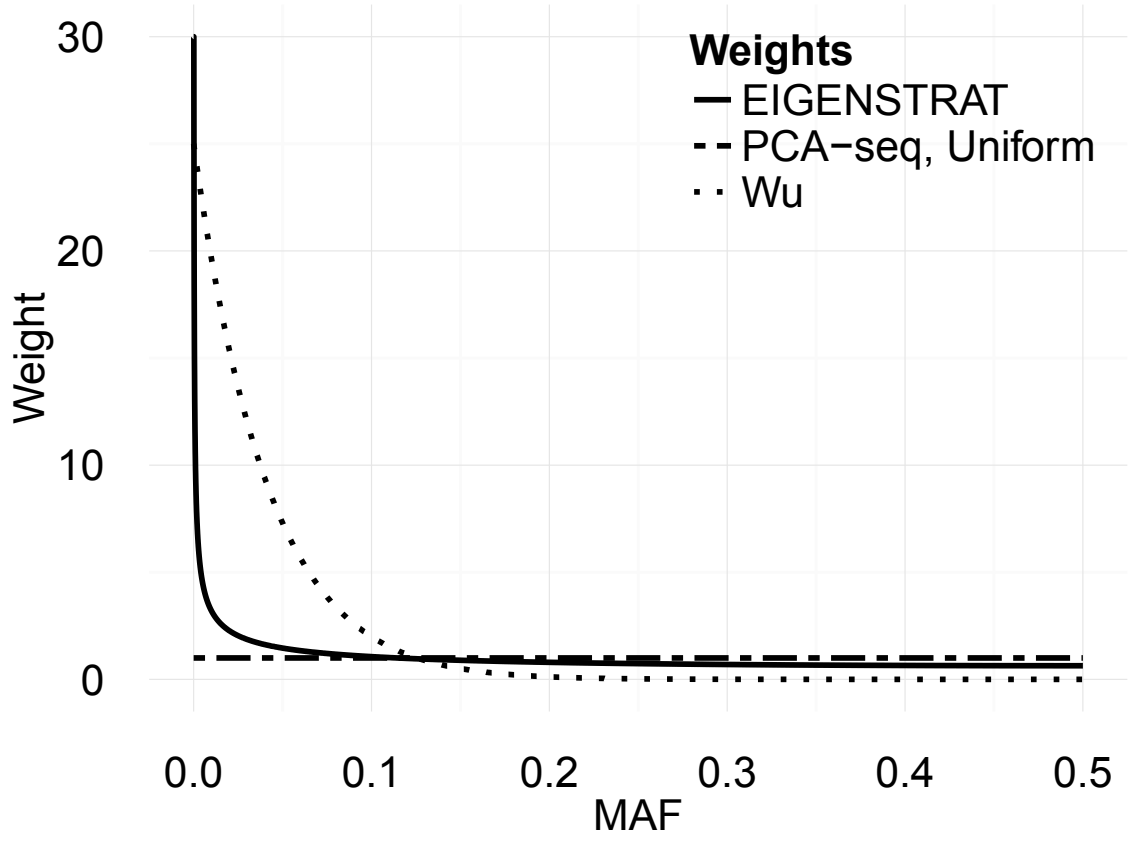
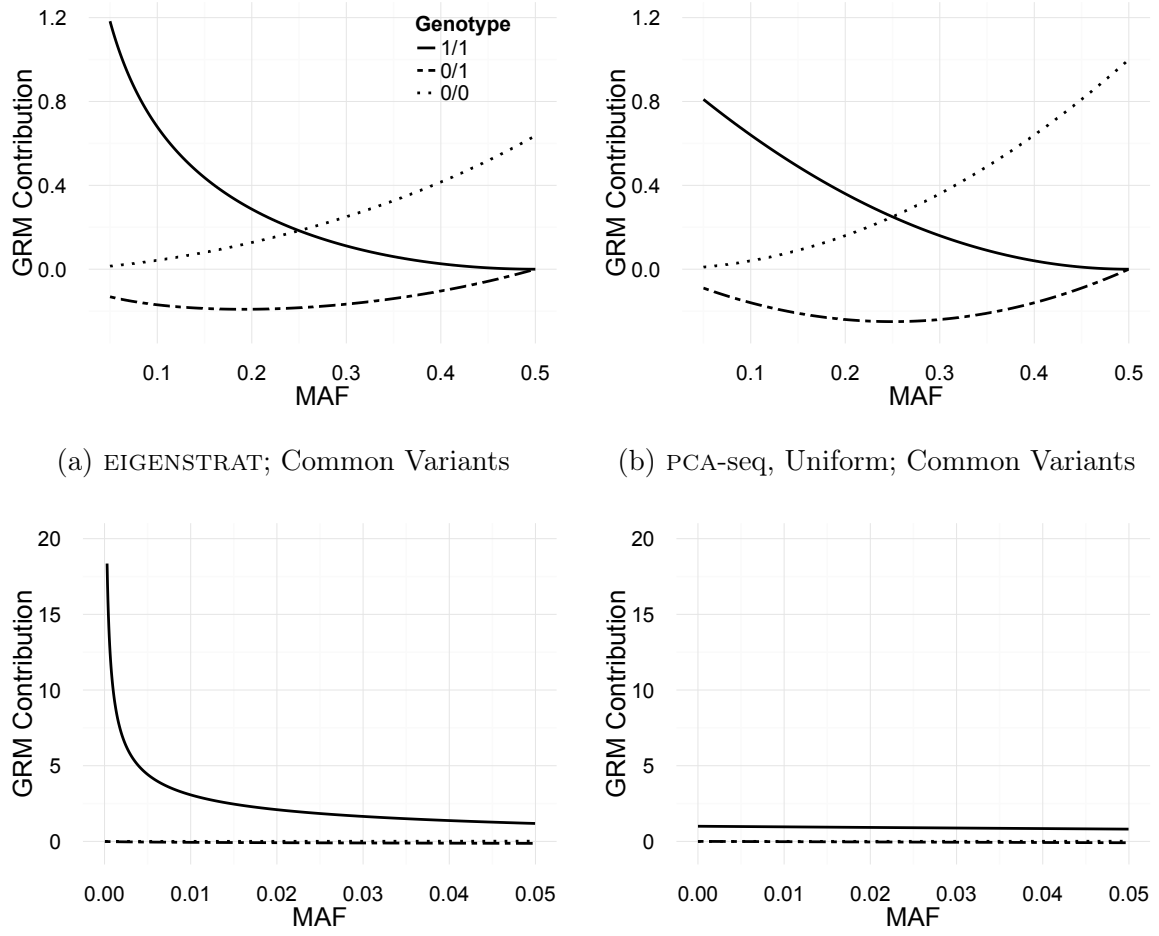


Figure 2.1: Three possible weighting schemes using the beta distribution by minor allele frequency (MAF): $\alpha = 0.5, \beta = 0.5$ (EIGENSTRAT); $\alpha = 1, \beta = 1$ (Uniform), $\alpha = 1, \beta = 25$ (Wu). All 3 weighting schemes are similar for common variants, but for rare variants, the EIGENSTRAT weights give the most weight to variants with the lowest minor allele frequencies.



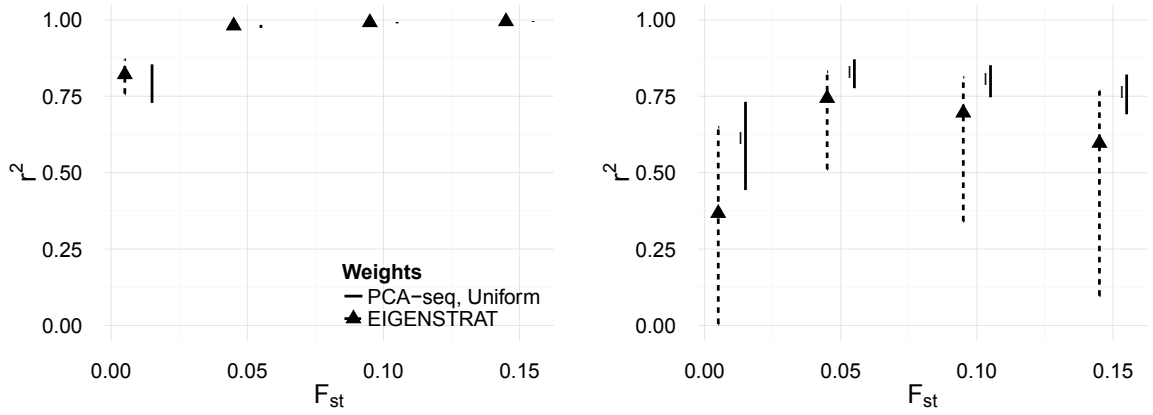
(a) EIGENSTRAT; Common Variants

(b) PCA-seq, Uniform; Common Variants

(c) EIGENSTRAT; Rare Variants

(d) PCA-seq, Uniform; Rare Variants

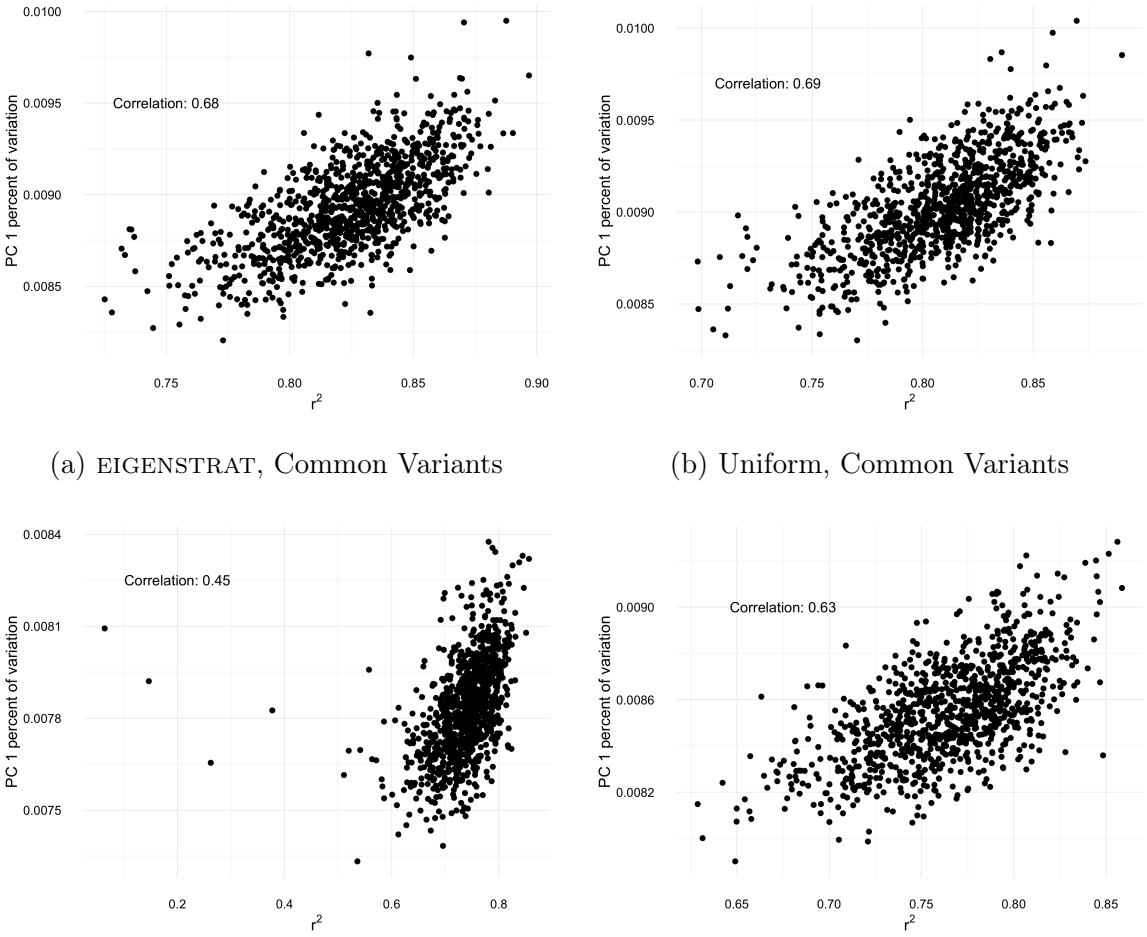
Figure 2.2: These figures show the contribution one locus makes to the entry in the GRM by minor allele frequency (MAF) for a pair of subjects under three different genotypes for the two subjects: both subjects are heterozygous for the minor allele (1/1), one subject is heterozygous and one subject is homozygous (0/1) for the minor allele, and neither subject is heterozygous for the minor allele (0/0). A common variant has a similar contribution under EIGENSTRAT (a) and PCA-seq with Uniform weights (b). However, a rare variant has very different contributions to the GRM under EIGENSTRAT (c) and PCA-seq with Uniform weights (d).



(a) Common Variants

(b) Rare Variants

Figure 2.3: The average and 95% confidence interval for the correlation between the first principal component and the true admixture from 1000 simulation replicates analyzed with EIGENSTRAT and PCA-seq with uniform weights. When applied to common variants (a), both methods perform similarly, although EIGENSTRAT has a slightly higher average correlation. When applied to rare variants (b), PCA-seq with uniform weights has significantly higher average correlation with the true admixture and less variability in the correlation, compared to EIGENSTRAT.



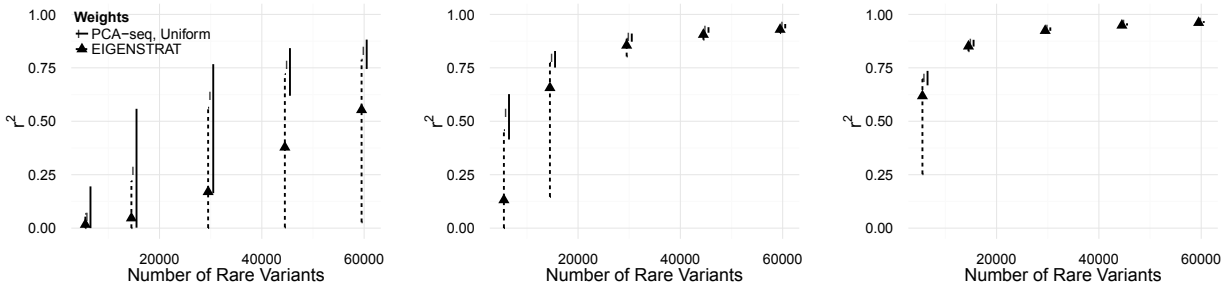
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

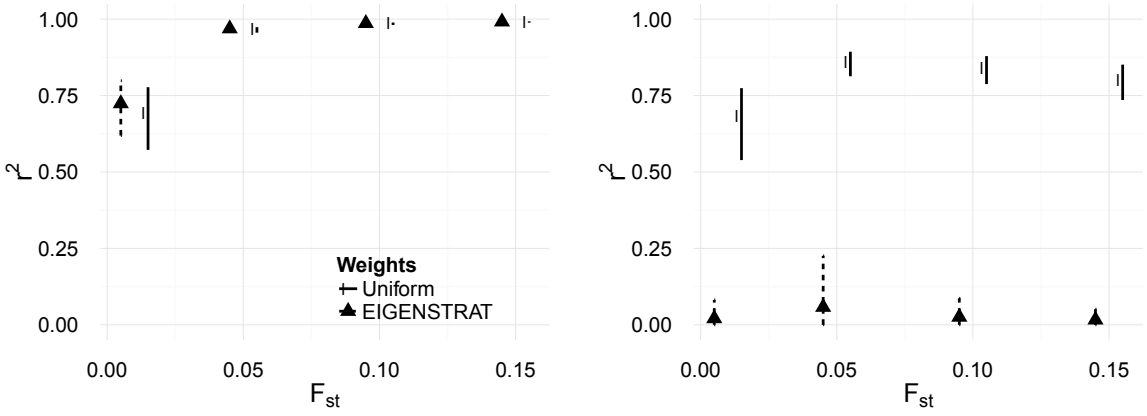
(d) Uniform, Rare Variants

Figure 2.4: These figures show the percent of variation explained by the first principal component from either EIGENSTRAT or PCA-seq with uniform weights by the correlation between the first principal component and the true admixture from 1000 simulation replicates.



(a) $N = 200$ (b) $N = 500$ (c) $N = 1000$

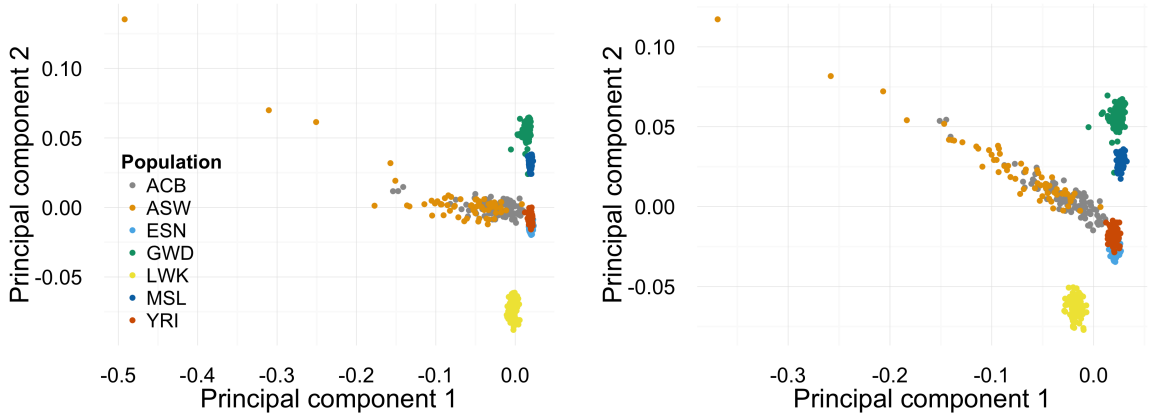
Figure 2.5: The average and 95% confidence interval for the correlation between the first principal component and the true admixture from 1000 simulation replicates of rare variants (MAF ≤ 0.025) analyzed with EIGENSTRAT and PCA-seq with uniform weights. With increasing numbers of rare variants, both methods have greater average correlation and less variability. Similarly, with increasing numbers of subjects, both methods have greater average correlation and less variability. However, PCA-seq with uniform weights consistently performs better than EIGENSTRAT with smaller sample sizes.



(a) Common Variants

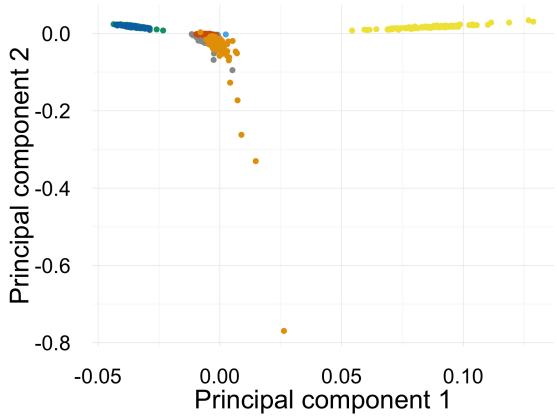
(b) Rare Variants

Figure 2.6: The average and 95% confidence interval for the correlation between the first principal component and the true admixture from 1000 simulation replicates in which one subject has a unique haplotype at 20 loci analyzed with EIGENSTRAT and PCA-seq with uniform weights. Rare variant genotyping errors have no effect on the correlation between the true admixture and the first principal component from either method when common variants are used infer population structure (a). When rare variants are used to infer population structure (b), there is almost no correlation between the first principal component from EIGENSTRAT and the true admixture. In the presence of genotyping errors, PCA-seq with uniform weights does not have a similar decrease in the correlation between the first principal component and the true admixture.

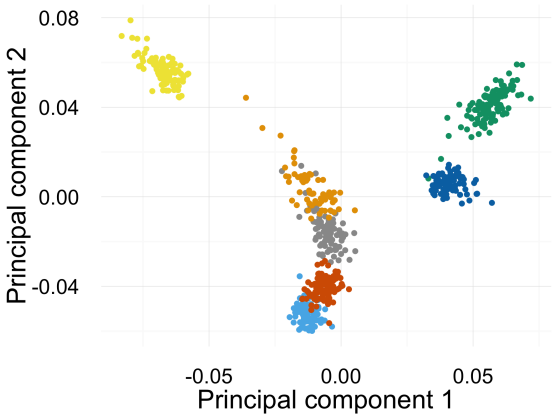


(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

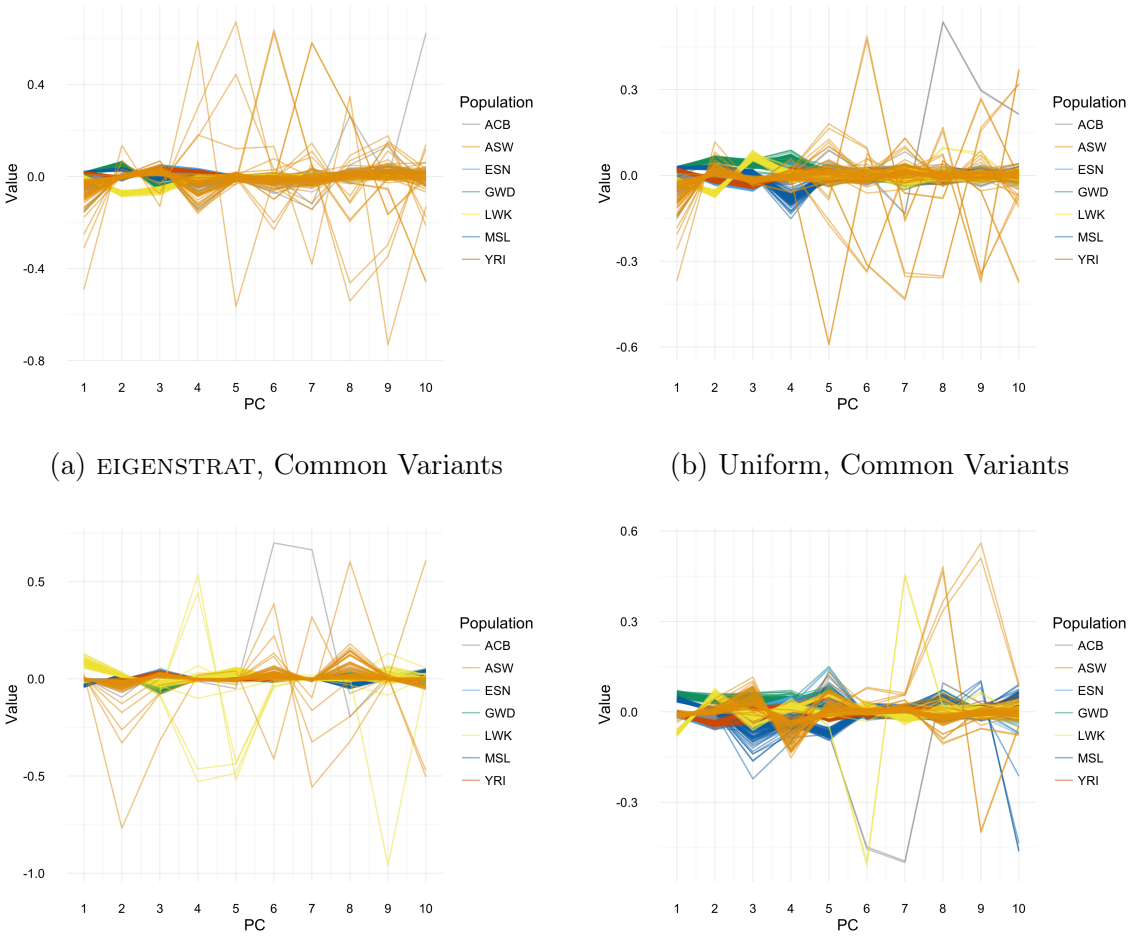


(c) EIGENSTRAT, Rare Variants



(d) Uniform, Rare Variants

Figure 2.7: The first and second principal components from the 1000 Genomes Phase 3 African super-population: African Caribbeans from Barbados (ACB), Americans of African Ancestry in the Southwest United States (ASW), Esan in Nigeria (ESN), Gambians in Western Divisions in the Gambia (GWD)), Luhya in Webuye, Kenya (LWK), Mende in Sierra Leone (MSL), and Yoruba in Ibadan, Nigeria (YRI)). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF ≤ 0.05, all other variants are common.



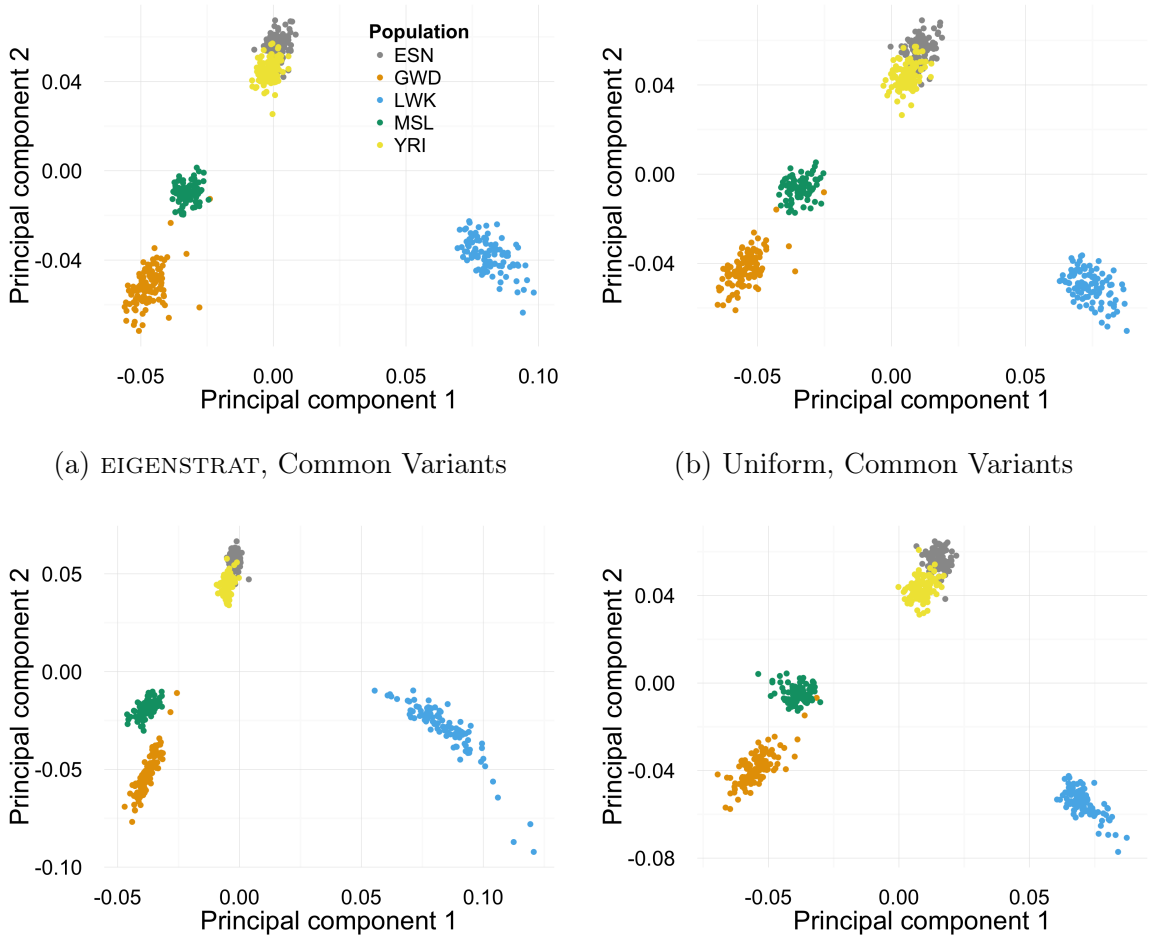
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure 2.8: Top 10 Principal Components from the 1000 Genomes Phase 3 African Ancestry Subpopulations: African Caribbeans from Barbados (ACB), Americans of African Ancestry in the Southwest United States (ASW), Esan in Nigeria (ESN), Gambians in Western Divisions in the Gambia (GWD), Luhya in Webuye, Kenya (LWK), Mende in Sierra Leone (MSL), and Yoruba in Ibadan, Nigeria (YRI). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure 2.9: First and Second Principal Components from the 1000 Genomes Phase 3 African Ancestry Subpopulations: Esan in Nigeria (ESN), Gambians in Western Divisions in the Gambia (GWD), Luhya in Webuye, Kenya (LWK), Mende in Sierra Leone (MSL) , and Yoruba in Ibadan, Nigeria (YRI). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.

Chapter 3

FAST APPROXIMATE INFERENCE OF POPULATION STRUCTURE

In Chapter 2, we presented a method for uncovering population structure using genotype data from sequence data. The phase III 1000 Genomes Project data set, which is one of the largest publicly available genomic datasets, includes 80 million SNPs from 2,504 people and is already challenging the limits of existing software. Given that the full human genome is 3 billion basepairs and that the cost of sequencing a full human genome has continued to decrease, the size of available datasets and the need for computationally efficient methods will only grow.

In this chapter, we address the specific challenge of implementing a computationally efficient version of PCA-seq. The main source of computational difficulty in PCA-seq is calculation of the GRM, which requires multiplying the extremely large standardized genotype data set by its transpose. This multiplication is extremely expensive in terms of time and memory. We can exploit a fundamental property of PCA to avoid this computation and directly decompose the standardized genotype matrix. While this removes the computationally difficult GRM calculation, it leads to an equally expensive matrix decomposition. To address this issue, we apply efficient algorithms for approximate or random matrix decompositions and examine the effects of approximation on PCA-seq.

We first present the necessary background on exact and random matrix decompositions in the context of PCA-seq, including a discussion of existing algorithms that improve upon EIGENSTRAT and their computational complexity. Next, we give a fast implementation of PCA-seq that makes use of the stochastic singular value decomposition (SSVD). We demonstrate the performance of this fast implementation and discuss the relevant issues that arise

when applying approximate matrix decompositions to rare variant data.

3.1 Asymptotic Time Complexity

The theoretical computation time of an algorithm can be derived as a function of the size of the inputs to the algorithm and is known as the time complexity of the algorithm. Time complexity is typically expressed as the asymptotic time complexity, where constants are ignored (indicated with big O notation). Time complexity is a useful method for comparing algorithms, but like all asymptotic theory, it does not necessarily reflect the actual performance of an algorithm on specific finite data sets. In particular, an algorithm may run faster than its worst case complexity on the specific data sets we encounter. In addition, since asymptotic time complexity ignores constant factors (for example due to implementation choices), one algorithm may run faster than another in practice. This means that an algorithm may perform well in theory but not in practice, or vice versa, depending on the implementation details. That said, asymptotic time complexity allows us to make fair comparisons between methods without favoring a method with a better implementation or data structure. Table 3.1 gives the asymptotic time complexity for the mathematical operations we will consider in our exact and approximate matrix decomposition algorithms. EIGENSTRAT and PCA-seq, which consist of a matrix multiplication to construct the GRM plus the eigendecomposition of the GRM, have asymptotic complexity $\mathrm{O}(N^2M + N^3)$ if $N \leq M$ (see Appendix A.4 for full details). Note that for all of the derivations, we have assumed $N < M$, as for sequence data this can reasonably be expected.

3.2 Existing Methods for Fast EIGENSTRAT

There are several existing methods for computing EIGENSTRAT in a computationally efficient manner. Beyond updates to the `smartpca` method in the `EIGENSOFT` package, which is the original implementation of EIGENSTRAT, there are `shellfish` [25], `flashpca` [26], and `FastPCA` [12]. `shellfish` [25], a parallel implementation of EIGENSTRAT, is considerably faster than `smartpca`, but only if the necessary computing hardware is available. `flashpca`

Table 3.1: Asymptotic Time Complexity of Mathematical Operations

Operation	Asymptotic Time Complexity
Matrix Multiplication: $[N \times M][M \times P]$	$\mathrm{O}(NMP)$
QR-decomposition: $[N \times M]$	$\mathrm{O}(NM \min\{N, M\})$
SVD: $[N \times M]$	$\mathrm{O}(NM \min\{N, M\})$
Eigendecomposition: $[N \times M]$	$\mathrm{O}(NM \min\{N, M\})$
Column-wise ℓ_2-norm standardization: $[N \times M]$	$\mathrm{O}(NM)$
Random multivariate normal matrix: $[N \times M]$	$\mathrm{O}(NM)$

and `FastPCA` both make use of random matrix decompositions, which provide improvements in speed whether they are implemented in serial or parallel. In general, random matrix decompositions approximate the top k principal components in two parts: first, a matrix is found that yields a low-rank approximation to the input matrix, then the appropriate decomposition of the low-rank approximation is taken, yielding the desired k principal components. This strategy replaces the single computationally expensive matrix decomposition with several approximation steps plus a faster matrix decomposition that together are much faster than the single decomposition.

`flashpca` is given in Algorithm 1. In the first part of `flashpca`, a low-rank approximation of the standardized genotype data, $\boldsymbol{B}$, is constructed by finding a matrix $\boldsymbol{Q}$ (steps 1-10). This low-rank approximation is used to find an approximation to the GRM, $\boldsymbol{S}$, and the eigendecomposition of $\boldsymbol{S}$ yields the desired principal components. Table 3.2 shows the asymptotic computational complexity of each step in the `flashpca` algorithm. The overall complexity of the algorithm is $\mathrm{O}(N^2M + NM)$, which is less than the complexity of EIGENSTRAT, if $N^2 > M$ (see Appendix A.4 for full details). For full sequence data, where M is on the order of 80 million, N would need to be greater than 8,900 for EIGENSTRAT to be computationally more efficient.

While `flashpca` is faster than `shellfish` and `smartpca` for computing EIGENSTRAT, it still has several computationally expensive steps which can be avoided. As we will show in §3.3, the explicit construction of the GRM can be avoided, as can the construction of $\boldsymbol{S}$. Furthermore, as the `flashpca` authors themselves note, the rescaling of the principal components in the last steps (13-15) is not strictly necessary.

Algorithm 1 flashpca Algorithm

Given a standardized $N \times M$ genotype matrix $\tilde{\boldsymbol{G}}$, and integers k, l and C, this algorithm computes the top k principal components of $\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}$.

1: $\boldsymbol{\Psi} \leftarrow \tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}$
2: Find a randomized $M \times (k+l)$ multivariate normal matrix: $\boldsymbol{\Omega}$
3: $\boldsymbol{A}^{(0)} \leftarrow \tilde{\boldsymbol{G}}\boldsymbol{\Omega}$
4: Normalize $\boldsymbol{A}^{(0)}$ column-wise using the ℓ_2-norm
5: **for** $c = 1$ **to** C **do**
6: $\quad \boldsymbol{A}^{(c)} \leftarrow \boldsymbol{\Psi}\boldsymbol{A}^{(c-1)}$
7: $\quad$ Normalize $\boldsymbol{A}^{(c)}$ column-wise using the ℓ_2-norm
8: **end for**
9: Take the QR-decomposition of $\boldsymbol{A}^{(C)}$: $\boldsymbol{QR}$
10: $\boldsymbol{B} \leftarrow \boldsymbol{Q}^\mathsf{T}\tilde{\boldsymbol{G}}$
11: $\boldsymbol{S} \leftarrow \boldsymbol{B}\boldsymbol{B}^\mathsf{T}$
12: Take the eigendecomposition of $\boldsymbol{S}$: $\boldsymbol{UDU}$
13: $\tilde{\boldsymbol{U}} \leftarrow \boldsymbol{QU}$
14: Square root the diagonal elements of $\boldsymbol{D}$ and divide them by $\sqrt{N-1}$: $\tilde{\boldsymbol{D}}$
15: $\boldsymbol{Z} \leftarrow \tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}$. The first k columns of $\boldsymbol{Z}$ are the top k principal components of $\boldsymbol{\Psi}$

`FastPCA` is another fast version of EIGENSTRAT, which provides several improvements over `flashpca`. The `FastPCA` algorithm is given in Algorithm 2. This algorithm eliminates several computationally intense steps found in the `flashpca` algorithm. The `FastPCA` algo-

Table 3.2: Asymptotic Complexity of the `flashpca` Algorithm

Step	Asymptotic Time Complexity
1	$\mathrm{O}(N^2M)$
2	$\mathrm{O}(M[k+l])$
3	$\mathrm{O}(NM[k+l])$
4	$\mathrm{O}(NM)$
5-8	$\mathrm{O}(C[N^2(k+l)]+CNM)$
9	$\mathrm{O}(N^2M)$
10	$\mathrm{O}(NM[k+l])$
11	$\mathrm{O}(M[k+l]^2)$
12	$\mathrm{O}([k+l]^3)$
13	$\mathrm{O}(N[k+l]^2)$
14	$\mathrm{O}(N)$
15	$\mathrm{O}(N[k+l]^2)$
Total	$\mathrm{O}(N^2M+NM)$

rithm does not explicitly compute the GRM, avoiding a computationally expensive matrix multiplication. Furthermore, assuming $N < M$, `FastPCA` uses a smaller initial random matrix, which makes constructing $\boldsymbol{\Omega}$ and multiplying it by the genotype data faster. `FastPCA` omits the construction of the approximate GRM, $\boldsymbol{S}$, and instead computes the unscaled principal components directly from the low-rank approximation of the genotype data using the singular value decomposition (SVD). Finally, `FastPCA` omits the unnecessary rescaling of the principal components.

`flashpca` and `FastPCA` take very different approaches to constructing the matrix that projects the genotype matrix into a low-rank space. `flashpca` iteratively projects the GRM into a low-rank space, then takes the QR-decomposition of the final projection and uses the $\boldsymbol{Q}$ matrix from the QR-decomposition to project the genotype data into a low-rank space. `FastPCA` iteratively projects the genotype data into a low-rank space along both dimensions, constructing a series of low-rank approximations. This series of low-rank approximations to the genotype data is combined into a single block matrix, and the $\boldsymbol{U}$ matrix from the SVD of this block matrix yields the projection matrix. By omitting unnecessary steps and using a different method of approximating the genotype data, `FastPCA` improves the asymptotic time complexity to $\mathrm{O}(NM + M)$. For $N < M$, `FastPCA` has significantly faster time complexity than EIGENSTRAT and `flashpca`. Therefore, as the `FastPCA` paper demonstrates, `FastPCA` performs much better in practice.

3.3 Exact Matrix Decompositions

As discussed in Chapter 1, PCA is a convenient method for summarizing the key features of a data set by decomposing the data into a set of uncorrelated axes of maximum variation. If we have a standardized genotype matrix $\tilde{\boldsymbol{G}}$, then the kth principal component for subject i is a normalized linear combination of the entries associated with subject i:

$$z_{ki} = \sum_{m=1}^{M} \eta_{nk} \tilde{g}_{in},$$

Algorithm 2 FastPCA Algorithm

Given a standardized $N \times M$ genotype matrix $\tilde{\boldsymbol{G}}$, and integers k, l and C, this algorithm computes the top k principal components of $\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}$.

1: Find a random $N \times (k+l)$ multivariate normal matrix $\boldsymbol{\Omega}$
2: $\boldsymbol{A}^{(0)} \leftarrow \tilde{\boldsymbol{G}}^T\boldsymbol{\Omega}$
3: **for** $c = 1$ **to** C **do**
4: $\quad \tilde{\boldsymbol{A}}^{(c)} \leftarrow \frac{1}{M}\tilde{\boldsymbol{G}}\boldsymbol{A}^{(c-1)}$
5: $\quad \boldsymbol{A}^{(c)} \leftarrow \tilde{\boldsymbol{G}}^T\tilde{\boldsymbol{A}}^{(c)}$
6: **end for**
7: $\boldsymbol{A} \leftarrow [\boldsymbol{A}^{(0)} \dots \boldsymbol{A}^{(C)}]$
8: Take the SVD of $\boldsymbol{A}$: $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$
9: $\boldsymbol{B} \leftarrow \boldsymbol{U}^\mathsf{T}\tilde{\boldsymbol{G}}$.
10: Take the SVD of $\boldsymbol{B}$: $\tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{V}}^\mathsf{T}$. The first k columns of $\tilde{\boldsymbol{V}}$ are the first k principal components of $\boldsymbol{\Psi}$.

Table 3.3: Asymptotic Complexity of the `FastPCA` Algorithm

Step	Asymptotic Time Complexity
1	$\mathrm{O}(N[k+l])$
2	$\mathrm{O}(NM[k+l])$
3–6	$\mathrm{O}(2CNM[k+l])$
7	—
8	$\mathrm{O}(M[C+1]^2[k+l]^2)$
9	$\mathrm{O}(NM[C+1][k+l])$
10	$\mathrm{O}(N[C+1][k+l]^2)$
Total	$\mathrm{O}(NM+M)$

where η_{nk} is the nth loading of the kth principal component and $\sum_{n=1}^{N} \eta_{nk}^2 = 1$. By solving for the η's and calculating the z_k's, we can uncover systematic structure in the data that is associated with high variation.

There are several methods for finding the principal components of a matrix $\tilde{\boldsymbol{G}}$. The principal components of a matrix can be found by taking the eigendecomposition of the empirical covariance matrix, $\tilde{\boldsymbol{\Phi}} = \tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}$ (see proof of Theorem A.3.1 in Appendix §A.3). This is the method used to compute the principal components in EIGENSTRAT and PCA-seq.

Since the eigendecomposition of a square matrix is equivalent to both the spectral decomposition and the singular value decomposition (SVD), the principal components of a matrix can also be found via the SVD. This is sometimes referred to as a generalization of PCA to non-square matrices. We prove the relationship between the SVD of $\tilde{\boldsymbol{\Phi}}$ and the SVD of $\tilde{\boldsymbol{G}}$, as we will make use of this relationship in our approximate matrix decomposition algorithms to avoid unnecessary matrix multiplication.

Theorem 3.3.1. *If $\tilde{\boldsymbol{G}}$ is a matrix with singular value decomposition $\tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{V}}$, then the matrix $\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}$ has SVD $\tilde{\boldsymbol{V}}^\mathsf{T}\boldsymbol{D}\tilde{\boldsymbol{V}}$, where $\boldsymbol{D}$ is a diagonal matrix whose non-zero entries are the square of the non-zero entries in $\tilde{\boldsymbol{D}}$.*

Proof. First, we note that if $\tilde{\boldsymbol{G}}$ has SVD $\tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{V}}$, then $\tilde{\boldsymbol{G}}^\mathsf{T}$ has SVD

$$\begin{aligned}\tilde{\boldsymbol{G}}^\mathsf{T} &= (\tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{V}})^\mathsf{T} \\ &= \tilde{\boldsymbol{V}}^\mathsf{T}\tilde{\boldsymbol{D}}^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}.\end{aligned}$$

Therefore,

$$\begin{aligned}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T} &= \tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{V}}^\mathsf{T}\tilde{\boldsymbol{D}}^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}) \\ &= \tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{V}}\tilde{\boldsymbol{V}}^\mathsf{T}\tilde{\boldsymbol{D}}^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T} \\ &= \tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{D}}^\mathsf{T}\tilde{\boldsymbol{U}}^\mathsf{T}\end{aligned}$$

as $\tilde{\boldsymbol{V}}$ is an orthogonal matrix, implying $\tilde{\boldsymbol{V}}^\mathsf{T}\tilde{\boldsymbol{V}} = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix. Since $\tilde{\boldsymbol{D}}$ and $\tilde{\boldsymbol{D}}^\mathsf{T}$ are both diagonal matrices with identical non-zero entries, $\tilde{\boldsymbol{D}}^\mathsf{T}\tilde{\boldsymbol{D}}$ is a diagonal

matrix with non-zero entries equal to the square of the non-zero entries in $\tilde{\boldsymbol{D}}$. Call this diagonal matrix $\boldsymbol{D}$. Then we have

$$\tilde{\boldsymbol{G}}^{\mathsf{T}}\tilde{\boldsymbol{G}} = \tilde{\boldsymbol{U}}^{\mathsf{T}}\boldsymbol{D}\tilde{\boldsymbol{U}}$$

Since $\tilde{\boldsymbol{U}}^{\mathsf{T}}$ is an orthogonal matrix, $\boldsymbol{D}$ is a diagonal matrix, and $\tilde{\boldsymbol{U}}$ is an orthogonal matrix (as each are matrices from an SVD), this is the SVD of $\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^{\mathsf{T}}$. □

Theorem 3.3.1 allows us to avoid explicitly finding the GRM in order to find the principal components of our genotype data. This reduces the computation time of our algorithm, as it eliminates the costly matrix transpose and multiplication needed to calculate the GRM. However, taking the SVD of $\tilde{\boldsymbol{G}}$ is still costly, implying we need a computationally feasible method of finding the SVD of a large matrix.

3.4 Approximate Matrix Decompositions

Approximate matrix decompositions provide computationally efficient algorithms for approximating the matrix decompositions of a large matrix. While there are many such algorithms, we focus on the stochastic singular value decomposition (SSVD), as we wish to combine approximate matrix decompositions with the efficiency of using the SVD to find the principal components of genotype data. For a thorough discussion of approximate matrix decompositions and a more general treatment of the algorithms presented here, see Halko et al. [27].

The general idea behind the SSVD is to approximate the genotype data with a low-rank matrix and then take the SVD of the low-rank matrix. Since the low-rank matrix will be much smaller than the full genotype data, finding the SVD will be much less computationally expensive. If the steps necessary to find the low-rank approximation are relatively fast, then we can construct a good approximation in less time than is necessary to take the full SVD. Finding the low-rank matrix and calculating the SVD can be optimized to improve the speed or memory use (or both) of the algorithm. There is a trade-off between computation time and the quality of approximation in both steps. Better approximations to the genotype data

require more computation time, reducing the speed of the algorithm. Full computation of the SVD increases the accuracy of the method, but can be slow for extremely large data sets. In this section, we primarily address improvements in speed related to approximating the genotype data. Approximating the SVD is primarily useful when the number of subjects is extremely large, resulting in a large matrix, even after projecting into a low-rank space.

The general stochastic SVD algorithm is given in Algorithm 3. Starting with a standardized $M \times N$ genotype matrix, $\tilde{\boldsymbol{G}}$, we multiply it by a random matrix $\boldsymbol{\Omega}$. Multiplying the genotype by a random matrix is a computationally efficient way of finding a low-rank approximation to $\tilde{\boldsymbol{G}}$, as a set of random vectors is likely to have full rank. To ensure that we have a set of vectors with the desired rank, we inflate the size of this random matrix beyond the number of principal components we wish to recover. We then find the $\boldsymbol{Q}$ matrix from the QR-decomposition of $\tilde{\boldsymbol{G}}\boldsymbol{\Omega}$ to obtain a a set of orthonormal vectors that approximate $\tilde{\boldsymbol{G}}$ in a lower dimensional space. A single iteration yields a reasonable approximation, and the error bound can be improved by iterating further. Once we have a good approximation to $\tilde{\boldsymbol{G}}$, we can project $\tilde{\boldsymbol{G}}$ into this space and take the SVD of this approximation. This algorithm, like `FastPCA`, is significantly faster than EIGENSTRAT, PCA-seq, and `flashpca`. The primary difference between `FastPCA` and this algorithm is the repeated normalization via QR-decompositions. These repeated QR-decompositions do not significantly increase the time complexity of the algorithm and improve the numerical stability of the algorithm.

Algorithm 3 can be further optimized by choosing an $\boldsymbol{\Omega}$ which can be multiplied by $\tilde{\boldsymbol{G}}$ quickly without sacrificing the accuracy of the approximation [27]. We use the subsampled random fast Hadamard transform (SRFHT), which is a structured random matrix [28]. This matrix is defined as

$$\boldsymbol{\Omega} = \sqrt{\frac{M}{k+l}}\boldsymbol{R}\boldsymbol{H}\boldsymbol{D}$$

where $\boldsymbol{R}$ is a matrix of vectors that randomly sample $l+k$ entries of length M vectors, $\boldsymbol{H}$ is the Fast Hadamard Transform matrix of the appropriate dimension, and $\boldsymbol{D}$ is a diagonal matrix of random signs (i.e. ± 1). Since the Fast Hadamard Transform is the discrete analog

to the Fast Fourier Transform, multiplying by $\boldsymbol{\Omega}$ essentially smoothes the genotype data out and then randomly samples this smoothed data to create a low-rank approximation. The SRFHT preserves the structure or distances in the matrix it projects and can be multiplied by another matrix in $O[2^{\lceil \log(M) \rceil} \log(2^{\lceil \log(M) \rceil})]$ time. By using the SRFHT, we can further reduce the computational burden of one of the multiplications in fast PCA-seq. Furthermore, when implemented, we do not need to explicitly construct $\boldsymbol{\Omega}$, eliminating the initial step to construct this matrix.

Algorithm 3 General SSVD Algorithm

Given a standardized $N \times M$ genotype matrix $\tilde{\boldsymbol{G}}$, and integers k, l and C, this algorithm computes the top k principal components of $\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^{\mathsf{T}}$.

1: Find a random $M \times (k+l)$ matrix: $\boldsymbol{\Omega}$.
2: $\boldsymbol{A}^{(0)} \leftarrow \tilde{\boldsymbol{G}}\boldsymbol{\Omega}$.
3: Take the QR-factorization of $\boldsymbol{A}^{(0)}$: $\boldsymbol{Q}^{(0)}\boldsymbol{R}^{(0)}$
4: **for** $c = 1$ **to** C **do**
5: Form $\boldsymbol{A}^{(c)} = \tilde{\boldsymbol{G}}\boldsymbol{Q}^{(i-1)}$
6: Take the QR-decompostion of $\boldsymbol{A}^{(c)} = \tilde{Q}^{(i)}\tilde{R}^{(i)}$
7: Form $\tilde{\boldsymbol{G}}\tilde{\boldsymbol{Q}}^{(i)}$.
8: Find the QR factorization, $Q^{(i)}R^{(i)}$.
9: **end for**
10: Form $\boldsymbol{B} = \boldsymbol{Q}^{(C)\mathsf{T}}\tilde{\boldsymbol{G}}$
11: Find the SVD of $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}}$

The error bound for this algorithm, which was derived by Halko et al. [27], is given in Theorem 3.4.1. The error bound on the algorithm is a function of the number of extra dimensions specified (l), the number of iterations (q), and the $k+1$ eigenvalue or singular value ($\lambda^{(k+1)}$). From this error bound, we can see that both the parameters q and l and the structure of the genotype data ($\tilde{\boldsymbol{G}}$) affect the accuracy of the algorithm. If the singular values decrease slowly, then the $k+1$ singular value will be large, as will the error bound. We have

Table 3.4: Asymptotic Complexity of the General SSVD Algorithm

Step	Asymptotic Time Complexity
1	$\mathrm{O}(M[k+l])$
2	$\mathrm{O}(NM[k+l])$
3	$\mathrm{O}(M[k+l]^2)$
4–9	$\mathrm{O}(2CNM[k+l]+C[N+M][k+l])$
10	$\mathrm{O}(NM[k+l])$
11	$\mathrm{O}(N[k+l]^2)$
Total	$\mathrm{O}(NM+M+N)$

observed that the singular values of the rare variants decay very slowly in our simulations in Chapter 2. If the singular values (eigenvalues) of the genotype decay quickly, we will need a better approximation of genotype data to get a small error bound. Increasing the number of iterations used to generate the projection matrix $\boldsymbol{Q}$ decreases the bound much more quickly than increasing the number of extra dimensions used to approximate $\tilde{\boldsymbol{G}}$. Unfortunately, increasing the number of iterations makes the algorithm much slower, relative to increasing the number of dimensions.

Theorem 3.4.1. *(From Halko et al. [27]) Suppose that $\tilde{\boldsymbol{G}}$ is a real $M \times N$ matrix. Select an exponent q and a number of singular vectors k, where $2 \leq k \leq 0.5\min(M,N)$. If Algorithm 3 is used to find the rank-$2k$ approximation $\boldsymbol{B}$ to $\tilde{\boldsymbol{G}}$, then the bound on the error is*

$$E||\tilde{\boldsymbol{G}}-\boldsymbol{B}|| \leq \lambda^{(k+1)} + \left[1+4\sqrt{\frac{2\min(M,N)}{k-1}}\right]^{\frac{1}{2q+1}} \lambda^{(k+1)}$$

where the expectation is taken with respect to the random matrix $\boldsymbol{\Omega}$ and $\lambda^{(k+1)}$ is the $k+1$ largest singular value of $\tilde{\boldsymbol{G}}$.

3.5 Simulations

To better understand the trade-offs in speed and accuracy between increasing the number of dimensions and increasing the number of iterations used to approximate the genotype data, we performed a series of simulations that are similar to those performed in Chapter 2. We considered the effect of varying population structure (F_{st}), the number of subjects and single nucleotide polymorphisms (SNPs). In all the simulations, we fixed the proportion of rare variants as a percent of the total number of loci. We simulated genotype data from individuals that were admixed between two ancestral populations with allele frequencies drawn from Balding-Nichols model [24] under F_{st} of 0.01 to 0.15. In these simulations, the population structure was the same for both the rare and common variants.

To compare the accuracy of the approximate matrix decompositions for both EIGENSTRAT and PCA-seq with the exact matrix decompositions, we simulated 200 unrelated individuals who were admixed between two populations with frequencies simulated under F_{st} ranging from 0.01 to 0.15. We simulated 10,000 loci, of which 60% were rare variants. For the approximate algorithm, we considered estimating 10 and 20 extra principal components (l) and 1 and 10 iterations (q). For each data set, we ran the approximate algorithm 10 times and averaged across the 10 replications.

Figure 3.1 shows the average and empirical 95% confidence interval for the correlation between the true admixture proportion and the first principal component from EIGENSTRAT and PCA-seq with uniform weights under both the exact and approximate algorithms for common variants. Figures (a) and (c) are replicates, as the exact method is the same regardless of q and l. Both approximate methods perform similarly for all but the lowest F_{st}, and in general the approximations are very similar to the exact results, except at the lowest F_{st}. Increasing l and q has no effect on the average correlation or the variance of the average correlation. This is probably due to the relatively simple population structure and the lack of noise in the data. We would expect to improvement in the approximation if we were using real genotype data.

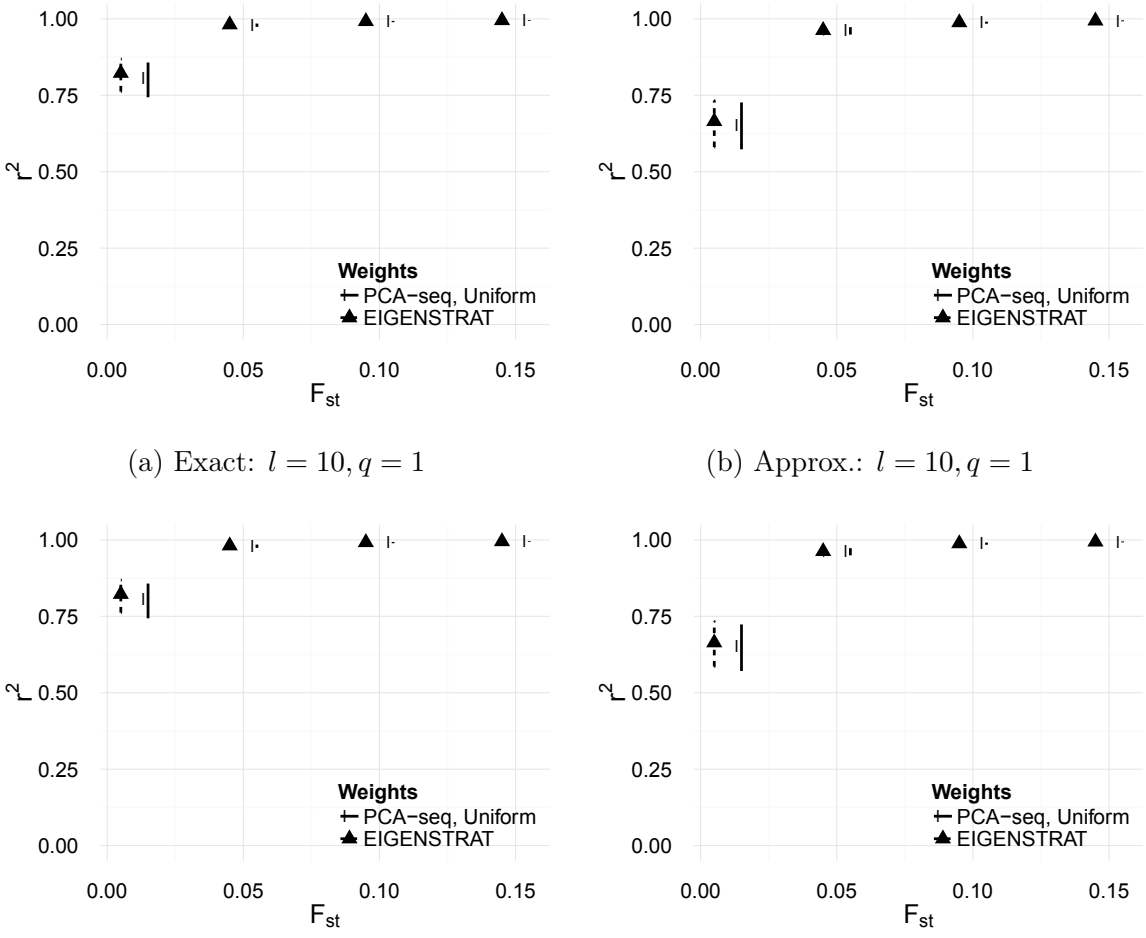
Figure 3.2 shows the results for the rare variants. There is a noticeable decrease in the average correlation under the approximate methods and an increase in the variability of the correlation with the true admixture. In particular the performance at the lowest F_{st} is significantly lower for the approximate methods. PCA-seq still outperforms EIGENSTRAT, even in when the approximate methods are used.

The sample sizes in the previous simulations were relatively small. Realistically, we would apply the approximate methods to data sets with many magnitudes more loci, and quite a few more subjects. Figure 3.3 shows the results for rare and common variants with increased sample sizes for $l = 10$ and $q = 1$. If we increase the number of loci to 100,000 and the number of subjects to 1500, we see that for both rare and common variants, the approximate methods perform nearly as well as the exact methods, even at the lowest F_{st}. This fits with our previous simulations, which suggested that increasing the sample size, in terms of both people and loci, greatly improves our ability to infer population structure.

Finally, we can look at the average and 95% confidence interval for the variance of correlation between the first principal component of each approximate decomposition and the true admixture proportion (Figure 3.5). For common variants at F_{st} values above 0.01, there is almost no variation between in replicates of the approximate methods. However, for rare variants, there is quite a bit of variation, although it decreases with increasing F_{st}. PCA-seq has lower variance on average, particularly for rare variants. Increasing l and q has little effect on the variance compared to the effect of increasing the sample size.

3.6 Conclusion

In this chapter we proposed a fast version of PCA-seq that approximates the principal components of interest. By approximating the principal components of interest, we can speed up the computation time considerably. While we have demonstrated the utility of this method using simulations, a highly scalable software package is future work.



(a) Exact: $l = 10, q = 1$

(b) Approx.: $l = 10, q = 1$

(c) Exact: $l = 10, q = 10$

(d) Approx: $l = 10, q = 10$

Figure 3.1: The average and 95% confidence interval for the correlation between the first principal component and the true admixture from 1000 simulation replicates analyzed with EIGENSTRAT and PCA-seq with uniform weights (Exact) and the average and 95% confidence interval for the average of 10 approximate matrix decomposition replicates. Each simulation replicate used simulated data from 4,000 loci with MAF greater than 0.05 and 200 subjects. l indicates the number of extra dimensions that were estimated and q is the number of iterations to estimate the $\boldsymbol{Q}$ matrix.

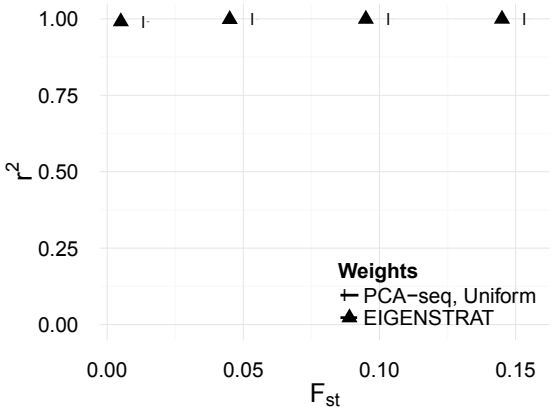
(a) Exact: $l = 10, q = 1$

(b) Approx.: $l = 10, q = 1$

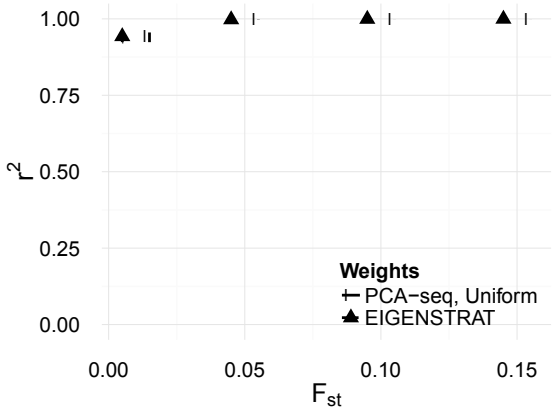
(c) Exact: $l = 10, q = 10$

(d) Approx: $l = 10, q = 10$

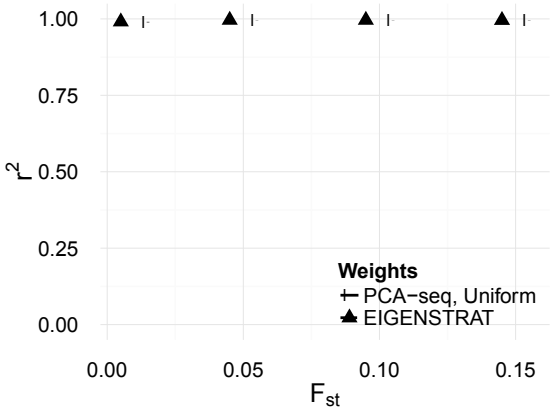
Figure 3.2: The average and 95% confidence interval for the correlation between the first principal component and the true admixture from 1000 simulation replicates analyzed with EIGENSTRAT and PCA-seq with uniform weights (Exact) and the average and 95% confidence interval for the average of 10 approximate matrix decomposition replicates. Each simulation replicate used simulated data from 6,000 loci with MAF less than or equal to 0.05 and 200 subjects. l indicates the number of extra dimensions above 10 that were estimated and q is the number of iterations to estimate the $\boldsymbol{Q}$ matrix.



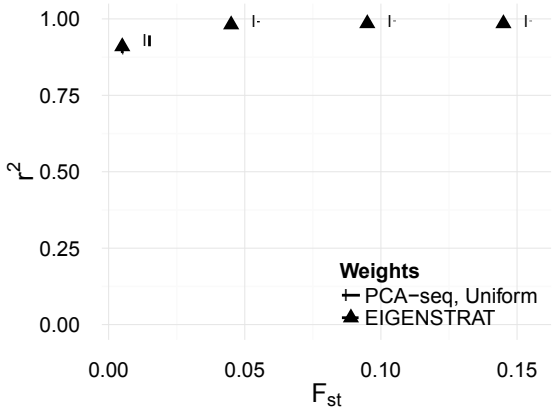
(a) Common, Exact: $l = 10, q = 1$



(b) Common, Approx.: $l = 10, q = 1$

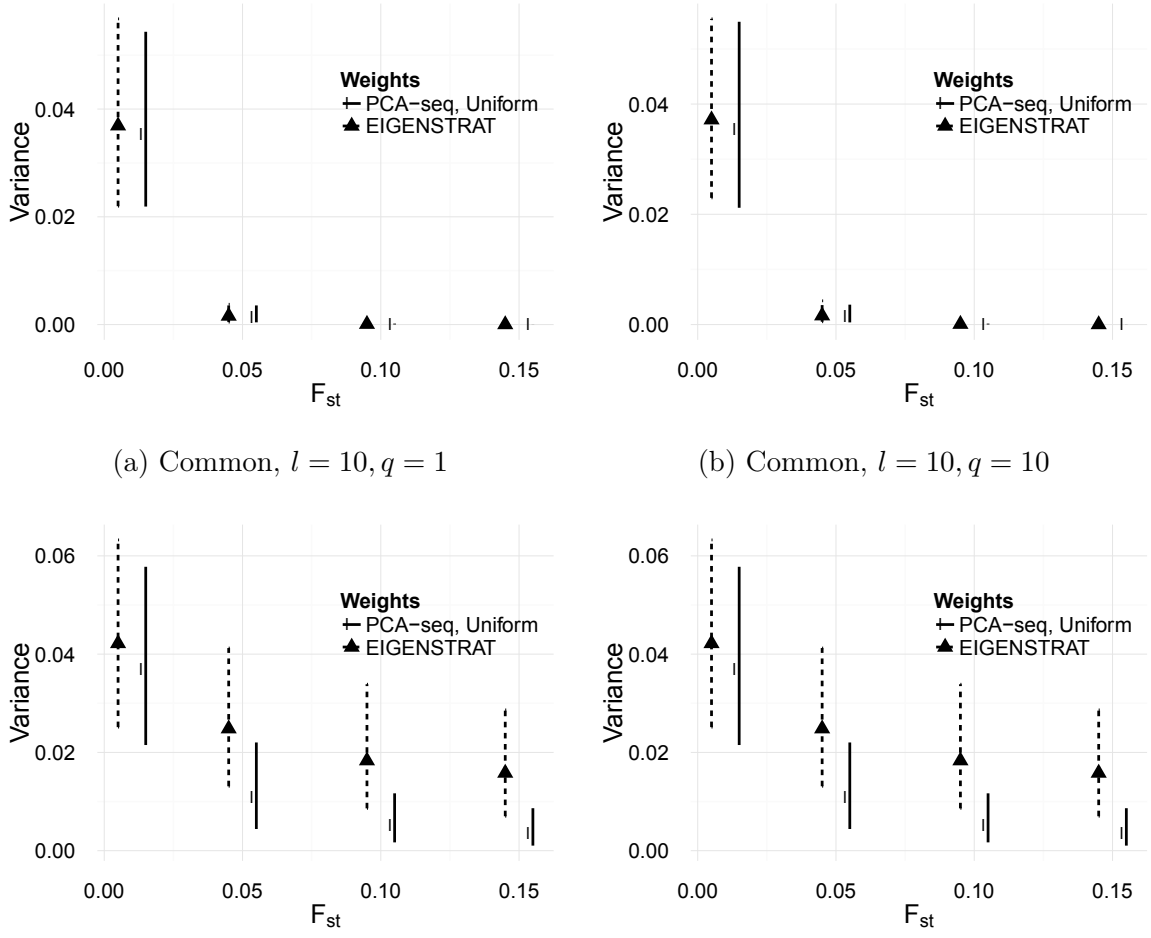


(c) Rare, Exact: $l = 10, q = 1$



(d) Rare, Approx: $l = 10, q = 1$

Figure 3.3: The average and 95% confidence interval for the correlation between the first principal component and the true admixture from 1000 simulation replicates analyzed with EIGENSTRAT and PCA-seq with uniform weights (Exact) and the average and 95% confidence interval for the average of 10 approximate matrix decomposition replicates. Each simulation replicate used simulated data from 10,000 loci where 60% have MAF less than or equal to 0.05 and 200 subjects. l indicates the number of extra dimensions above 10 that were estimated and q is the number of iterations to estimate the $\boldsymbol{Q}$ matrix.



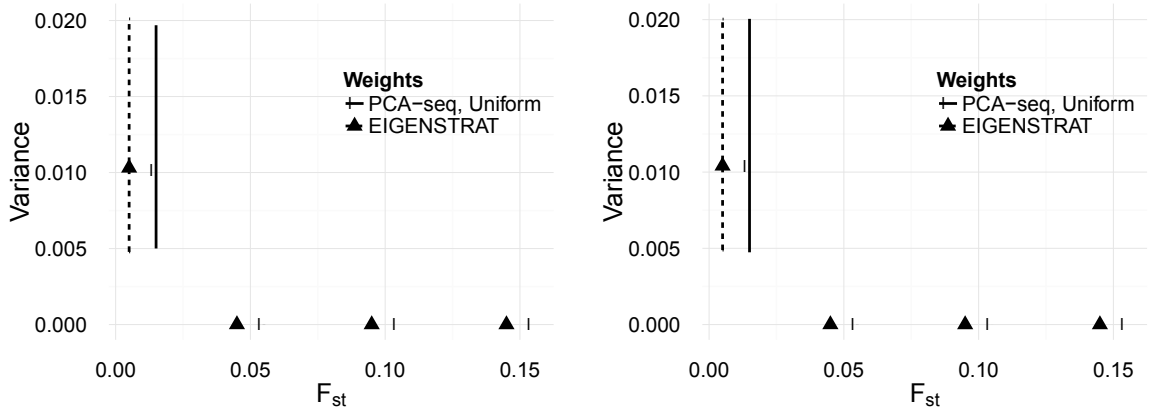
(a) Common, $l = 10, q = 1$

(b) Common, $l = 10, q = 10$

(c) Rare, $l = 10, q = 1$

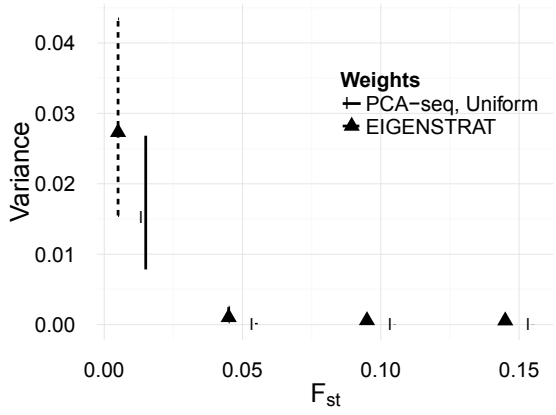
(d) Rare, $l = 10, q = 10$

Figure 3.4: The average and 95% confidence interval for the variance in the correlation between the first principal component and the true admixture from 1000 simulation replicates analyzed with approximate matrix decompositions. The variance was estimated using 10 replicates for each data set. Each simulation replicate used simulated data from 10,000 where 60% of the loci had MAF less than or equal to 0.05 and 200 subjects. l indicates the number of extra dimensions above 10 that were estimated and q is the number of iterations to estimate the $\boldsymbol{Q}$ matrix.

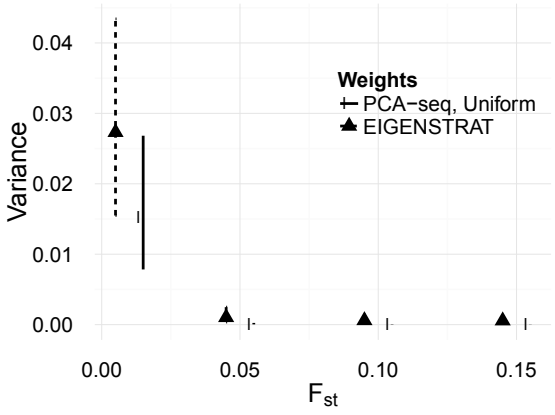


(a) Common, $p = 10, q = 1$

(b) Common, $p = 10, q = 10$



(c) Rare, $p = 10, q = 1$



(d) Rare, $p = 10, q = 10$

Figure 3.5: The average and 95% confidence interval for the correlation between the first principal component and the true admixture from 1000 simulation replicates analyzed with EIGENSTRAT and PCA-seq with uniform weights (Exact) and the average and 95% confidence interval for the average of 10 approximate matrix decomposition replicates. Each simulation replicate used simulated data from 7,000 loci with MAF less than or equal to 0.05 and 200 subjects. p indicates the number of extra dimensions above 10 that were estimated and q is the number of iterations to estimate the $\boldsymbol{Q}$ matrix.

Chapter 4

CHOICE OF WEIGHTS WITH PCA-SEQ

In Chapter 2, we focused on the general utility of using specific weights instead of relying on the default weights implied by using EIGENSTRAT. However, we did not present any specific recommendations for weights, beyond suggesting the use of existing weights found in the literature. Furthermore, we limited our comparisons to the uniform weights, as in a sense, they are the least informative about population structure. In this chapter, we present a method for deriving weights that are informative for population structure using reference data. While this method relies on reference data, it does so only to calculate the weights. Therefore, the population structure inference is still made using the genotype data being analyzed, not the reference data.

This chapter relies heavily on the basics of information theory, so we first present the necessary background. We discuss some of the previous applications of these information theory concepts to genetic data. Then we present our method for deriving informative weights for population structure, along with a discussion of when different weighting schemes are most useful. Finally, we apply these weights to the 1000 Genomes Project data to demonstrate their utility.

4.1 Introduction to Information Theory

Information theory was first developed by Claude Shannon in 1948, as a means of formalizing the notion of how much information a random variable contains. In the landmark paper, "A Mathematical Theory of Communication," he presented the basic concepts of information theory, of which we will make use [29]. Shannon used a notion of informativeness related to how surprising we find an event. He outlined four key properties of his measure of

informativeness, which is information. Let $g_{im} = t$ with $t \in \{0, 1, 2\}$ be the genotype for a subject i at locus m with genotype frequency q_t. Then the information in g_{im} should have the properties:

1. $I(g_{im} = t) \geq 0$
2. $I(g_{im} = t)$ should be zero when $q_t = 0$
3. $I(g_{im} = t)$ should decrease as $q_t \to 0$
4. $I(g_{im}, g_{jm}) = I(g_{im}) + I(g_{jm})$ for independent genotypes g_{im} and g_{jm}.

These properties are intuitive. Property 1 states that observing a genotype cannot give us less information than we had previously. Property 2 states that a genotype we do not observe cannot give us any information. Property 3 states that the more likely a genotype is, and therefore the less surprising it is, the less information we derive from that genotype. Finally, property 4 states that if we observe two independent genotypes jointly, we gain no more information than if we observed the two genotypes separately. These four properties are only satisfied by the function

$$I(g_{im}) = -\log_b(q_t).$$

where b is any base, but is usually taken to be 2 or e. If the information is calculated using base 2, it is the information in bits, and is interpreted as the number of yes-no questions we could answer with the given data. We will use base 2 throughout this chapter, although in some cases we will use the natural logarithm for convenience.

To find the information in a random variable, or rather its distribution, we naturally take the average information. Shannon called this the entropy of the random variable. If we have a locus m with genotypes $t \in \{0, 1, 2\}$ and corresponding genotype frequencies $q_0 \ldots q_2$, then the entropy of the genotype, g_m, at this locus is

$$H(g_m) = \sum_{i=0}^{2} I(g_m = i) q_i.$$

This quantity is bounded from below by zero, and from above by $-\log_b(2)$. That is, a uniformly distributed random variable has the maximum entropy. We can also define conditional entropy, which is naturally the conditional expectation of information, given a second random variable.

Entropy is useful for describing the amount of information inherent in a probability distribution and the associated random variable. However, we may also want to describe the amount of information shared between random variables. This quantity is mutual information, and it is expressed in terms of the entropy. If we have a set of populations, $\mathcal{S}$, then the mutual information between the genotype g_m at a locus m and set of populations is

$$\begin{aligned}\mathcal{I}(\mathcal{S}, g_m) &= H(\mathcal{S}) - H(\mathcal{S}|g_m) \\ &= H(g_m) - H(g_m|\mathcal{S}).\end{aligned}$$

Mutual information is bounded from below by zero and from above by the minimum of $H(g_m)$ and $H(\mathcal{S})$. A mutual information of zero indicates independence. Furthermore, mutual information is equivalent to the Kullback-Leibler distance between the joint distribution of g_m and $\mathcal{S}$ and the product of their marginals.

4.2 Information Theory in Genetics

The information theoretic quantities discussed above have been applied to the analysis of genomic data in several ways. In this section, we present a brief survey of these uses, which have focused on linkage disequilibrium detection, haplotype phasing, clustering of SNPs, and detecting association with phenotypes.

Entropy, mutual information, and other related measures have been used extensively in linkage disequilibrium detection. Mutual information is a popular multi-locus measure of linkage disequilibrium (LD). Nothnagel et al. [30] propose measuring the LD at multiple loci using the normalized entropy difference, which is the difference between the observed entropy and the observed entropy assuming the SNPs are independent, scaled by the observed entropy assuming independence. This measure is equivalent to the G-test or log-likelihood

ratio test statistic and can be expressed in terms of the mutual information between the SNPs. However, this test statistic has several limitations: it does not reach the upper bound of 1 and the size of the test statistic depends on the number SNPs in the haplotype block. Liu and Lin [31] propose an adjusted measure of mutual information which does not have these problems. Liu and Lin take a similar approach as Nothnagel et al., but divide the test statistic above by its maximum value which yields a measure that falls between 0 and 1. They use this measure, along with an entropy-based measure of haplotype diversity to develop an algorithm for selecting tagging SNPs.

Entropy between SNPs has also been used to develop haplotype phasing. Halperin and Karp [32] propose an algorithm that phases genotypes by minimizing the entropy of the resulting haplotypes. Gusev et al. [33] propose a related algorithm that is fast and scalable.

Mutual information has been used to cluster SNPs by their relative similarity. Dawy et al. [34] proposed using a pairwise distance metric related to the pairwise mutual information between two SNPs, called the normalized information distance to find clusters of SNPs due to evolution. Their primary goal was to detect potentially causal clusters of SNPs. In this method, they used population-based controls to calculate the pairwise mutual information between SNPs. They then used the normalized information distance, which is a true metric, to construct a matrix summary of their dataset and applied MDS to this matrix to obtain clusterings of SNPs.

4.3 Informative Weights

We propose deriving informative weights for population structure inference with PCA-seq using reference data that is annotated with population membership. In our method, we estimate the mutual information between each locus and the population membership using the annotated reference data, then apply PCA-seq to the our study data using the mutual information derived from the reference data as weights. There are several methods for estimating mutual information, which we discuss in §4.3.1. Assume we have a reference data set, which consists of a genotype data set $\boldsymbol{G}_{ref}$ and a corresponding set of population

memberships for a set of populations, $\mathcal{S}$, plus a study data set, which consists of just genotype data $\boldsymbol{G}_{study}$. Then for a locus m appearing in $\boldsymbol{G}_{study}$, we calculate the mutual information weights using

$$w_m = \sqrt{\hat{\mathcal{I}}(\boldsymbol{g}_m, \mathcal{S})}$$

where $\boldsymbol{g}_m$ is the corresponding genotype data from the reference data set and $\hat{\mathcal{I}}$ is an estimator of the mutual information. Note that this method requires our reference data set and study data set to have genotype data for the same set of loci.

For the rest of this section, we consider the theoretical mutual information weights when we have discrete population membership labels. Our method could be generalized to continuous definitions of ancestry, but estimation of the mutual information in this case is considerably more complex and treatment of the continuous case is beyond the scope of this dissertation. Therefore, we focus on categorical ancestry, such as population labels.

We know from §4.1 that the mutual information between the genotype and population membership is bounded from above by the minimum of $H(\boldsymbol{g}_m)$ and $H(\mathcal{S})$. If there are three or more populations, $\mathcal{I}(\boldsymbol{g}_m, \mathcal{S})$ is bounded by $H(\boldsymbol{g}_m) = -\log_2(3)$, as there are only three genotype frequencies. If there are only two populations, then $\mathcal{I}(\boldsymbol{g}_m, \mathcal{S})$ is bounded by $H(\boldsymbol{g}_m) = -\log_2(2) = 1$.

We can be more specific about the upper bound on the mutual information, as we know that the genotype is a multinomial distribution with three probabilities. For two equally sized populations, the true mutual information between a locus and the population membership is maximized when each population has a different genotype at the locus. If both populations are equally sized, then such a locus has a MAF of 0.5 if the locus is monomorphic for a different allele in each population. In general, we can derive the maximum theoretical mutual information in this case at each MAF f:

$$\min\left[1, 2f\log_2\left(\frac{2-4f}{1-4f}\right) - \log_2(1-2f) + \frac{1}{2}\log_2(1-4f)\right].$$

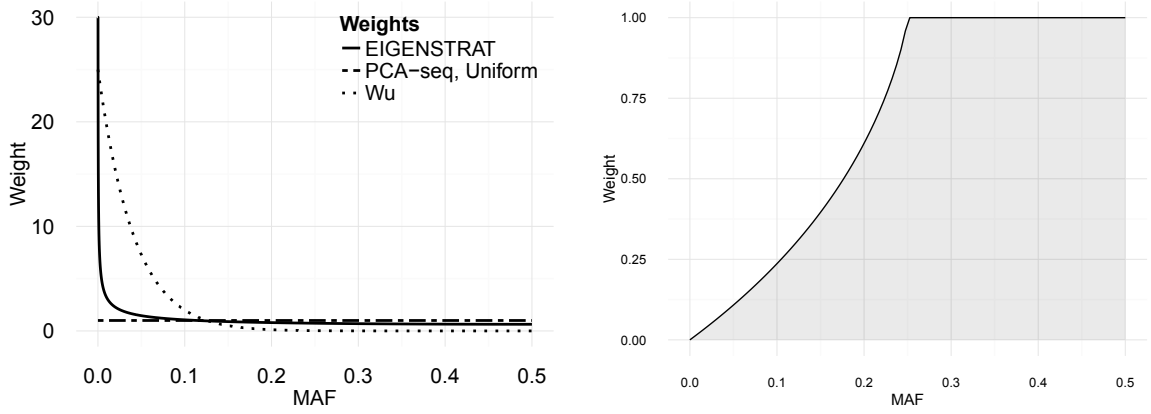
This maximum will be 1 for all MAF ≥ 0.25, as at a MAF of 0.25, one population can be homozygous for the common allele and the other population can be heterozygous for the

minor allele. That is, the maximally informative locus with a MAF of 0.25 is a locus where everyone in one population has a genotype of 0 and everyone in the other population has a genotype of 1.

Figure 4.1 shows a comparison of the EIGENSTRAT weights, uniform weights, and the mutual information weights across a range of MAF in the two population scenario. While the EIGENSTRAT, uniform, and beta distribution weights are constant at a fixed MAF, the mutual information takes a range of values at a fixed MAF. The solid line indicates the maximum value of the mutual information weights, and the grey shading represents the range of values the mutual information weights take at each MAF. The mutual information weights have an inflection point at a MAF of 0.25, as was noted above. For loci with a MAF less than 0.25, loci with higher MAFs have greater maximum mutual information weights. This does not necessarily imply that rare variants are less informative than common variants, as we are not assured to observe fully informative loci. We can imagine a case where every locus with MAF greater than 0.25 is completely uninformative for ancestry, while every locus with MAF less than 0.25 is fully informative.

We can see the utility of weights based on mutual information by considering two loci with the same MAF. Any two loci with the same MAF will receive the same weight under a MAF-based weighting scheme, regardless of their informativeness about population structure. Furthermore, since the variance of a locus is a function of MAF, these two loci will have the same variance under MAF-based weights, even after weighting. If the weights for these to loci are both large, then their variances will also be large. As noted previous in Chapter 3, principal components seeks to maximize the variance of the transformed data of the centered and scaled genotype matrix. The corresponding loadings of these loci will be high, and the principal components will be strongly influenced by these loci.

However, two loci may differ greatly in their informativeness for population structure. If we give every locus with the same MAF the same weight, we are treating informative and uninformative loci at that MAF the same and may be obscuring the population structure. Instead, if we give loci weight based on their informativeness, then informative loci will have



(a) Beta Distribution Weights

(b) Information Weights

Figure 4.1: These figures show the weights by minor allele frequency (MAF) under four different weighting schemes for two equally sized populations. (a) This plot shows three possible weighting schemes using the beta distribution: $\alpha = 0.5, \beta = 0.5$ (EIGENSTRAT); $\alpha = 1, \beta = 1$ (Uniform), $\alpha = 1, \beta = 25$ (Wu). (b) This plot shows the mutual information weights. The solid line indicates the maximum value, and the shading indicates the range of potential values.

much higher variance relative to uninformative loci and informative loci will more strongly influence the principal components.

This has an implication for when various weights will be most useful. If most loci are informative for population structure, then the variability in the data naturally corresponds to population structure, and uniform weights will be useful. But, if most of the variability in the data is not due to population structure, as for example, there are only a few loci that are relatively highly informative for population structure, then informative weights will be most helpful for uncovering population structure with PCA-seq.

4.3.1 Mutual information estimators

There are several possible estimators for entropy and therefore mutual information when the random variables in question are discrete. This problem is easier than the estimation of entropy and mutual information of continuous random variables, to which considerable research effort has been devoted. In this section we focus on three estimators of the entropy of a discrete random variable. In the continuous case, there are direct estimators of mutual information, which avoid estimation of the probability density function and therefore entropy. However, in the simpler discrete case, mutual information estimates are typically based on estimates of the probabilities and entropy. For a random variable X that takes on values $x_1 \dots x_T$, the three entropy estimators we consider are

1. Maximum Likelihood Estimator
$$\hat{H}_{MLE}(X) = -\sum_{t=1}^{T} P(X = x_t) \log_e[P(X = x_t)]$$
2. Miller-Madow Bias Corrected Estimator
$$\hat{H}_{MM}(X) = \hat{H}_{MLE} + \frac{T-1}{2N}$$
3. Jack-knife Estimator
$$\hat{H}_{JK}(X) = N\hat{H}_{MLE} - \frac{N-1}{N}\sum_{j=1}^{N} \hat{H}_{MLE}^{(-j)}.$$

We start by discussing the properties of the MLE estimator, as the properties of the Miller-Madow and Jack-knife estimator can be derived from those of the MLE estimator.

MLE Estimator

The plug-in or maximum likelihood estimator for entropy simply replaces the probabilities with their maximum likelihood estimates:

$$\hat{H}_{MLE}(X) = -\sum_{t=1}^{T} P(X = x_t) \log_e[P(X = x_t)].$$

While appealingly simple, this estimator is negatively biased. Paninski [35] gives upper and lower bounds on the bias of the MLE estimator:

$$-\log_e\left(1+\frac{T-1}{N}\right) \leq Bias(\hat{H}_{MLE}) \leq 0$$

where the lower bound becomes tight as $N/T \to 0$ and the upper bound becomes tight as $N/T \to \infty$. Antos and Kontoyiannis [36] bound the variance of the MLE for all N by

$$\mathrm{Var}(\hat{H}_{MLE}) \leq \left(\frac{\log_e(N)^2}{N}\right)$$

Furthermore, the variance of the MLE decreases at a rate on the order of 1/N as $N \to \infty$, for points on the interior of the simplex.

Given that the MLE estimator is biased, and that no unbiased estimator exists, we are interested in the relative trade-off between bias and variance as the sample size increases. Paninski [35] derives an approximation to the variance to bias ratio when $N >> T$:

$$\frac{\mathrm{Var}}{\mathrm{Bias}^2} \approx \frac{N\log_e(T)^2}{T^2}$$

given that we have $T = 3$ for the genotype and reasonably expect a relatively small number of populations, this variance bound holds for our data. When this ratio is greater than one, the variance dominates. Asymptotically, the three estimators are equivalent, so these results apply to the Miller-Madow and Jack-knife estimators as well. In general, given the relatively small number of populations and the relatively large number of subjects in current studies, we expect this to be true for most genotype datasets.

4.4 Application: 1000 Genomes Data

To mimic analyzing an independent data set with weights derived from the 1000 Genomes Project data, we divided the data from each super-population into two datasets, chosen randomly. This division was done such that the number of subjects from each population in each of the two new data sets is roughly the same. For each super-population, we used one data set to estimate the mutual information between each locus and the populations within

each super-population. We then use these estimates of the mutual information as the weights when we apply PCA-seq to the other half of the data. For each super-population, we applied PCA-seq to all of the loci, without distinguishing between rare and common variants. We considered 3 different weights: the uniform weights plus the MLE and Miller-Madow mutual information estimators. We present the results of these data sets in 3D-plots and in parallel coordinate plots.

Figure 4.2 shows the results from the African super-population with all variants. In this figure, the solid lines represent the average principal component value within each population and the shading represents the minimum and maximum values within each population. Populations that are highly clustered by a principal component will have non-overlapping lines and small shaded areas for that principal component. Populations that are unrelated to a principal component, or that consist of admixed individuals, will have wide shaded areas that overlap with other populations. In general, a set of principal components separates a set of populations well if each population has a unique trajectory across the principal components and the shading is tightly clustered around the mean.

Under EIGENSTRAT, the first four principal components separate the African super-population into five populations. The first and second principal components separate the admixed populations (ASW and ACB) from the African populations. Across the first four principal components, we can see that the two Nigerian populations (Esan in Nigeria (ESN), Yoruba in Ibadan, Nigeria (YRI)) are very tightly clustered together, and the Gambian (Gambians in Western Divisions in the Gambia (GWD)) and Mende (Mende in Sierra Leone (MSL)) populations are also relatively tightly clustered. EIGENSTRAT does not capture this fine-scale structure, and higher-order principal components primarily uncover outliers, leading to extreme principal component values.

Under PCA-seq with uniform weights, the first two principal components uncover continental structure (African versus non-African), as did the first two principal components under EIGENSTRAT. The third principal component from PCA-seq with uniform weights separates the Kenyan population from the other populations. The fourth principal component is

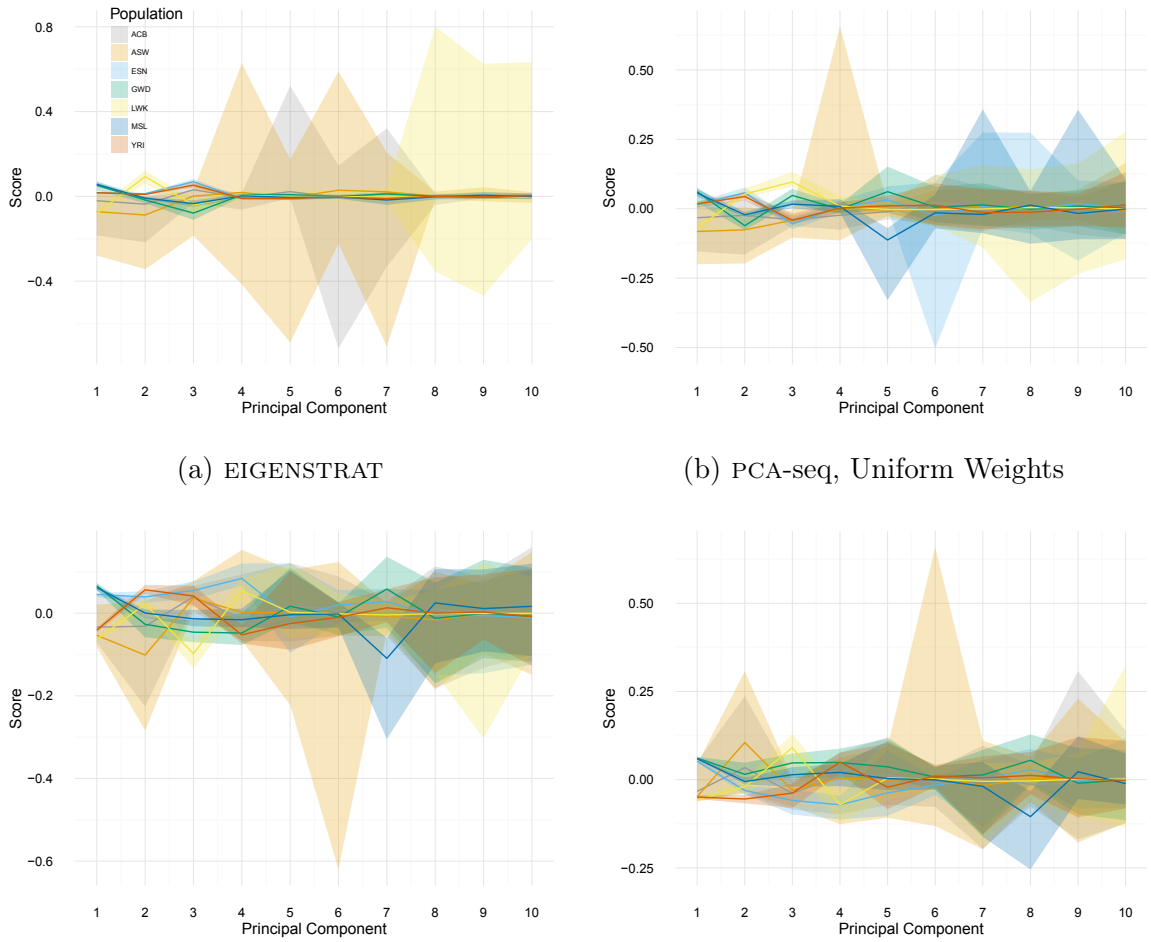
primarily driven by outliers from the admixed African-American population. The fifth principal component provides some separation between the Mende and Gambian populations. In general, PCA-seq with uniform weights shows greater separation between the African populations, although outliers (admixed subjects) can still strongly influence principal components.

When empirical information weights are used, the first two principal components uncover the continental structure. However, the second and third principal components more clearly delineate the within continental structure. In particular, there is much greater separation between the two Nigerian populations (ESN, YRI), as these populations have distinct trajectories across the first five principal components. The separation between the Mende and Gambian populations is less pronounced, although these two populations are still separated, particularly by the seventh principal component. There is still evidence that one admixed subject is an outlier with respect to ancestry, suggesting that this subject has a unique ancestry relative to the rest of the super-population.

The Miller-Madow mutual information weights show similar separation, although generally, these weights perform worse. Given the relatively large sample sizes, in terms of both subjects and loci, the bias correction may be introducing more error into the weights.

4.5 Conclusion

In this chapter, we propose a method for constructing weights that are informative for population structure using the mutual information between ancestry and each locus, as estimated from reference data. This method allows us to better elucidate fine-scale population structure with PCA-seq, although it relies on reference data.



(a) EIGENSTRAT

(b) PCA-seq, Uniform Weights

(c) PCA-seq, Empirical MI Weights

(d) PCA-seq, Miller-Madow MI Weights

Figure 4.2: The mean and range of the first 10 four principal components from EIGENSTRAT and PCA-seq with uniform and mutual information weights applied to the 1000 Genomes Phase 3 African super-population: African Caribbeans from Barbados (ACB), Americans of African Ancestry in the Southwest United States (ASW), Esan in Nigeria (ESN), Gambians in Western Divisions in the Gambia (GWD), Luhya in Webuye, Kenya (LWK), Mende in Sierra Leone (MSL), and Yoruba in Ibadan, Nigeria (YRI). The solid lines represent the average principal component value within each population, while the shading represents the range (minimum to maximum). (MI: Mutual Information)

Chapter 5

CONCLUSION

In conclusion, we have proposed PCA-seq, a method for inferring population structure with human sequence data. This method performs as well as EIGENSTRAT when applied to common variants and performs significantly better than EIGENSTRAT when applied to rare variants. With modification, PCA-seq is computationally feasible when applied to large data sets and can uncover the desired population structure given a reference data set. This method is a significant improvement over existing principal component analysis (PCA)-based methods for population structure, such as EIGENSTRAT, as we have intentionally considered the effects of weighting the genotype data with respect to uncovering population structure with principal components. Furthermore, since PCA-seq can accommodate both rare and common variants, we need not arbitrarily divide genotype data sets into "rare" and "common" variants. This is key, as these classifications are highly data dependent.

There are several avenues of future work to consider, particularly for the informative weights. Future work on PCA-seq should consider whether PCA-seq improves the type I error rates or false discovery rates when using rare variant association testing. We could also consider whether other machine learning methods for uncovering structure in a data set might be more appropriate. For example, independent component analysis is a technique similar to principal component analysis that minimizes the mutual information between the axes. For fast PCA-seq, more work is needed to successfully implement a highly scalable version of the method. Finally, we need to explore the effect of incorrect reference ancestry on the population structure inference when using informative weights.

BIBLIOGRAPHY

[1] National Institutes of Health: National Human Genome Research Institute. The Cost of Sequencing a Human Genome, 2016. URL `https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/`.

[2] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, July 2006.

[3] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[4] Ivan P Gorlov, Olga Y Gorlova, Shamil R Sunyaev, Margaret R Spitz, and Christopher I Amos. Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. *The American Journal of Human Genetics*, 82(1):100–112, January 2008.

[5] R David Hawkins, Gary C Hon, and Bing Ren. Next-generation genomics: an integrative approach. *Nature Publishing Group*, 11(7):476–486, June 2010.

[6] Nathan Tintle, Hugues Aschard, Hu, Inchi, Nora Nock, Haitian Wang, and Elizabeth Pugh. Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 Genomes Project exon sequencing data in unrelated individuals: summary results from Group 7 at Genetic Analysis Workshop 17. *Genetic Epidemiology*, 35(S1):S56–S60, 2011.

[7] Iain Mathieson and Gil McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, 44(3):243–246, February 2012.

[8] Timothy D O'Connor, Adam Kiezun, Michael Bamshad, Stephen S Rich, Joshua D Smith, Emily Turner, NHLBIGO Exome Sequencing Project, ESP Population Genetics, Statistical Analysis Working Group, Suzanne M Leal, and Joshua M Akey. Fine-Scale Patterns of Population Stratification Confound Rare Variant Association Tests. *PLoS One*, 8(7):1–10, 2013.

[9] Qianying Liu, Dan L Nicolae, and Lin S Chen. Marbled inflation from population structure in gene-based association studies with rare variants. *Genetic epidemiology*, 37 (3):286–292, April 2013.

[10] Timothy D O'Connor, Wenqing Fu, NHLBI GO Exome Sequencing Project, ESP Population Genetics and Statistical Analysis Working Group, Emily Turner, Josyf C Mychaleckyj, Benjamin Logsdon, Paul Auer, Christopher S Carlson, Suzanne M Leal, Joshua D Smith, Mark J Rieder, Michael J Bamshad, Deborah A Nickerson, and Joshua M Akey. Rare Variation Facilitates Inferences of Fine-Scale Population Structure in Humans. *Molecular biology and evolution*, November 2014.

[11] Stephen Leslie, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, Ellen C Royrvik, Barry Cunliffe, Daniel J Lawson, Daniel Falush, Colin Freeman, Matti Pirinen, Simon Myers, Mark Robinson, Peter Donnelly, and Walter Bodmer. The fine-scale genetic structure of the British population. *Nature*, 519(7543):309–314, March 2015.

[12] Kevin J Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J Patterson, and Alkes L Price. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics*, 98(3):456–472, March 2016.

[13] B Devlin and K Roeder. Genomic control for association studies. *Biometrics*, 55(4): 997–1004, December 1999.

[14] Luigi Luca Cavalli-Sforza, Paolo Menozzi, and Alberto Piazza. *The history and geography of human genes*. Princeton university press, 1994.

[15] Bruce Rannala and Joanna L Mountain. Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences*, 94(17):9197–9201, 1997.

[16] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, June 2000.

[17] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, September 2009.

[18] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Medical Genetics*, 81(3):559–575, September 2007.

[19] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2 edition, 2009.

[20] Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics*, 8(1):e1002453–16, January 2012.

[21] Yiwei Zhang, Xiaotong Shen, and Wei Pan. Adjusting for Population Stratification in a Fine Scale With Principal Components and Sequencing Data. *Genetic epidemiology*, 37(8):787–801, December 2013.

[22] Bo Eskerod Madsen and Sharon R Browning. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genetics*, 5(2):e1000384, February 2009.

[23] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Medical Genetics*, 89(1):82–93, July 2011.

[24] David J Balding and Richard A Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. In *Human Identification: The Use of DNA Markers*, pages 3–12. Springer, 1995.

[25] Dan Davison. shellfish: Parallel PCA and data processing for genome-wide SNP data, 2009. URL `http://http://www.stats.ox.ac.uk/ davison/software/shellfish/shellfish.php`.

[26] Gad Abraham and Michael Inouye. Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PLOS ONE*, 9(4):e93766–5, April 2014.

[27] N Halko, P G Martinsson, and J A Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2):217–288, January 2011.

[28] Joel A Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.

[29] Claude E Shannon. *The Bell System Technical Journal*, 27:623–656, October 1948.

[30] M Nothnagel, R Fürst, and K Rohde. Entropy as a Measure for Linkage Disequilibrium over Multilocus Haplotype Blocks. *Human Heredity*, 54(4):186–198, May 2003.

[31] Zhenqiu Liu and Shili Lin. Multilocus LD measure and tagging SNP selection with generalized mutual information. *Genetic epidemiology*, 29(4):353–364, 2005.

[32] Eran Halperin and Richard M Karp. The minimum-entropy set cover problem. *Theoretical Computer Science*, 348(2-3):240–250, December 2005.

[33] Alexander Gusev, Ion Mandoiu, and Bogdan Pasaniuc. Highly Scalable Genotype Phasing by Entropy Minimization. *ACM Transactions on Computational Biology and Bioinformatics*, 5(2):252–261, 2008.

[34] Zaher Dawy, Bernhard Goebel, Joachim Hagenauer, Christophe Andreoli, Thomas Meitinger, and Jakob C Mueller. Gene Mapping and Marker Clustering Using Shannon's Mutual Information. *ACM Transactions on Computational Biology and Bioinformatics*, 3(1):47–57, January 2006.

[35] Liam Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, June 2003.

[36] András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, October 2001.

Appendix A

SUPPLEMENTARY MATERIAL

A.1 EIGENSTRAT *and* SDR *Relationship: Full Derivation*

In this section we present a complete derivation of the relationship between the SDR estimator and the EIGENSTRAT estimator. Let $\tilde{\boldsymbol{g}}_i$ be the vector of genotypes for the ith person at all M loci. Then, if we standardize these genotype values by subtracting their means $2\hat{p}_m$ and dividing by $\sqrt{2\hat{p}_m(1-\hat{p}_m)}$, we have a vector of standardized genotypes $\tilde{\boldsymbol{g}}_i$ and the EIGENSTRAT estimator can be written as

$$\tilde{\psi}_{ij} = \frac{1}{M}\tilde{\boldsymbol{g}}_i^{\mathsf{T}}\tilde{\boldsymbol{g}}_j.$$

The Spectral Dimension Reduction estimator has the form:

$$\theta_{ij} = \mathbb{I}(i=j) - \frac{b_{ij}}{\sqrt{t_{ii}t_{jj}}}$$

where $t_{ii} = \sum_{r=1}^{N} b_{ir}$, $b_{ij} = \sqrt{\tilde{\boldsymbol{g}}_i^{\mathsf{T}}\tilde{\boldsymbol{g}}_j}\mathbb{I}(\sqrt{\tilde{\boldsymbol{g}}_i^{\mathsf{T}}\tilde{\boldsymbol{g}}_j} > 0)$, and $\mathbb{I}(\sqrt{\tilde{\boldsymbol{g}}_i^{\mathsf{T}}\tilde{\boldsymbol{g}}_j} > 0) = 1$ if $\tilde{\boldsymbol{g}}_i^{\mathsf{T}}\tilde{\boldsymbol{g}}_j > 0$ and 0 otherwise.

We can rewrite b_{ij} in terms of the EIGENSTRAT estimator:

$$\begin{aligned} b_{ij} &= \sqrt{2M\tilde{\psi}_{ij}} \times \mathbb{I}\left(\sqrt{2M\tilde{\psi}_{ij}} > 0\right) \\ &= \sqrt{2M\tilde{\psi}_{ij}} \times \mathbb{I}\left(\tilde{\psi}_{ij} > 0\right), \end{aligned}$$

noting that since M is the number SNPs and therefore always greater than zero, the indicator will only equal 1 when $\tilde{\psi}_{ij}$ is greater than zero.

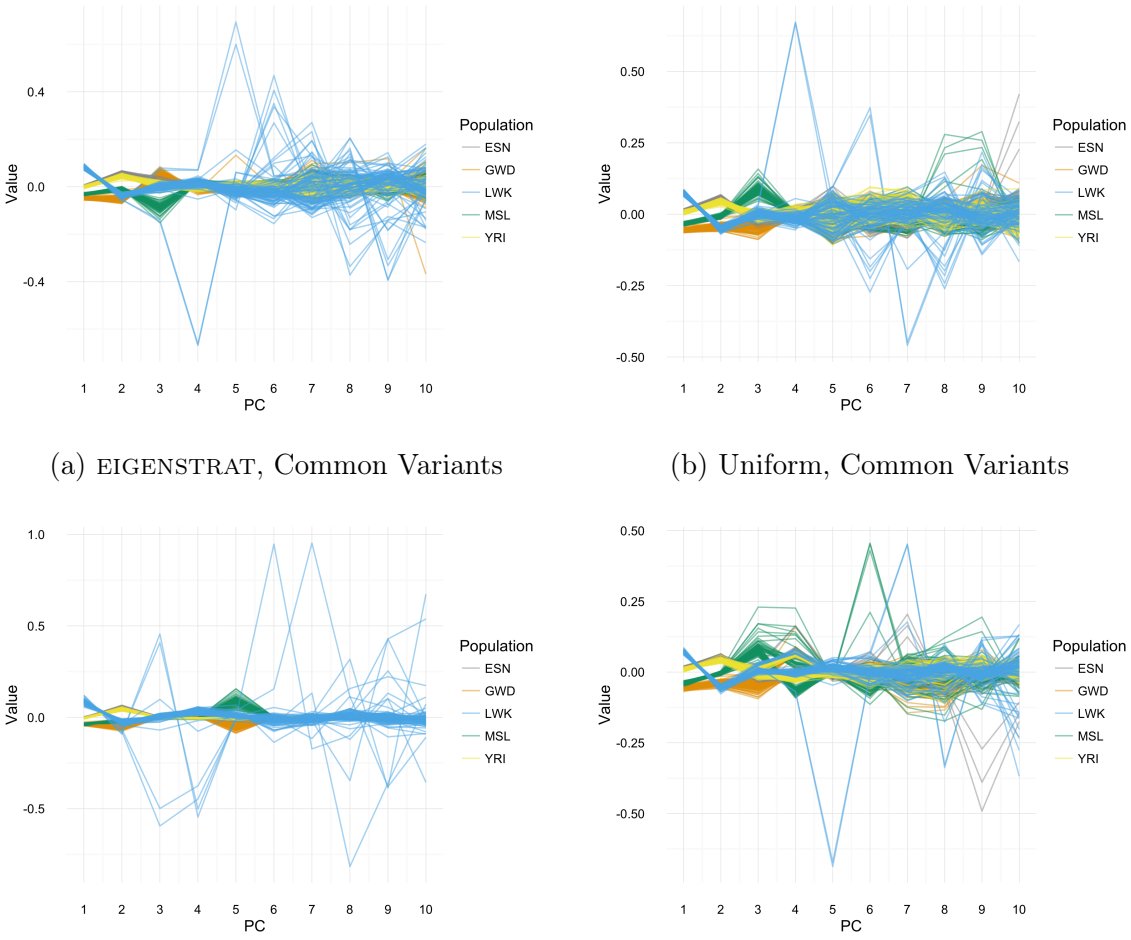
Using this expression, we can write the full SDR estimator in terms of the EIGENSTRAT

estimator.

$$\theta_{ij} = \mathbb{I}(i=j) - \frac{\sqrt{2M\tilde{\psi}_{ij}} \times \mathbb{I}\left(\tilde{\psi}_{ij} > 0\right)}{\sqrt{t_{ii}t_{jj}}}$$
$$= \left(\frac{\mathbb{I}(i=j)}{\tilde{\psi}_{ij}} - \frac{\sqrt{2M}\mathbb{I}(\tilde{\psi}_{ij} > 0)}{\sqrt{t_{ii}t_{jj}\tilde{\psi}_{ij}}} \right) \tilde{\psi}_{ij}$$

While this expression is very complicated, we can treat the first term in the product as a weight that is a function of the genotype data. For extremely rare variants, $\tilde{\psi}_{ij}$ will tend towards infinity. This method will only uncover population structure when applied to rare variants, if this weight sufficiently counteracts this tendency toward infinity.

A.2 PCA-seq *Application to 1000 Genomes Results*



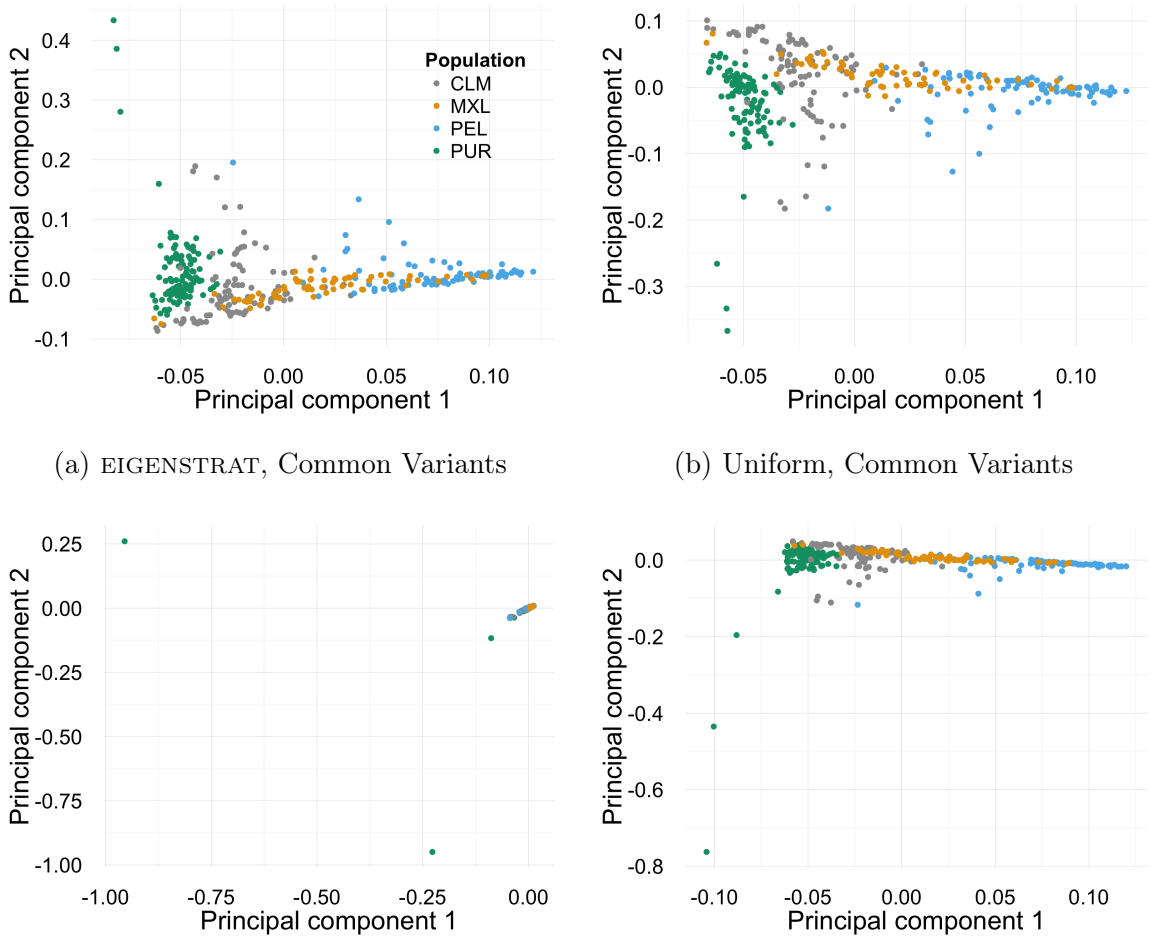
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.1: Top 10 Principal Components from the 1000 Genomes Phase 3 African Ancestry Subpopulations: Esan in Nigeria (ESN), Gambians in Western Divisions in the Gambia (GWD), Luhya in Webuye, Kenya (LWK), Mende in Sierra Leone (MSL), and Yoruba in Ibadan, Nigeria (YRI). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



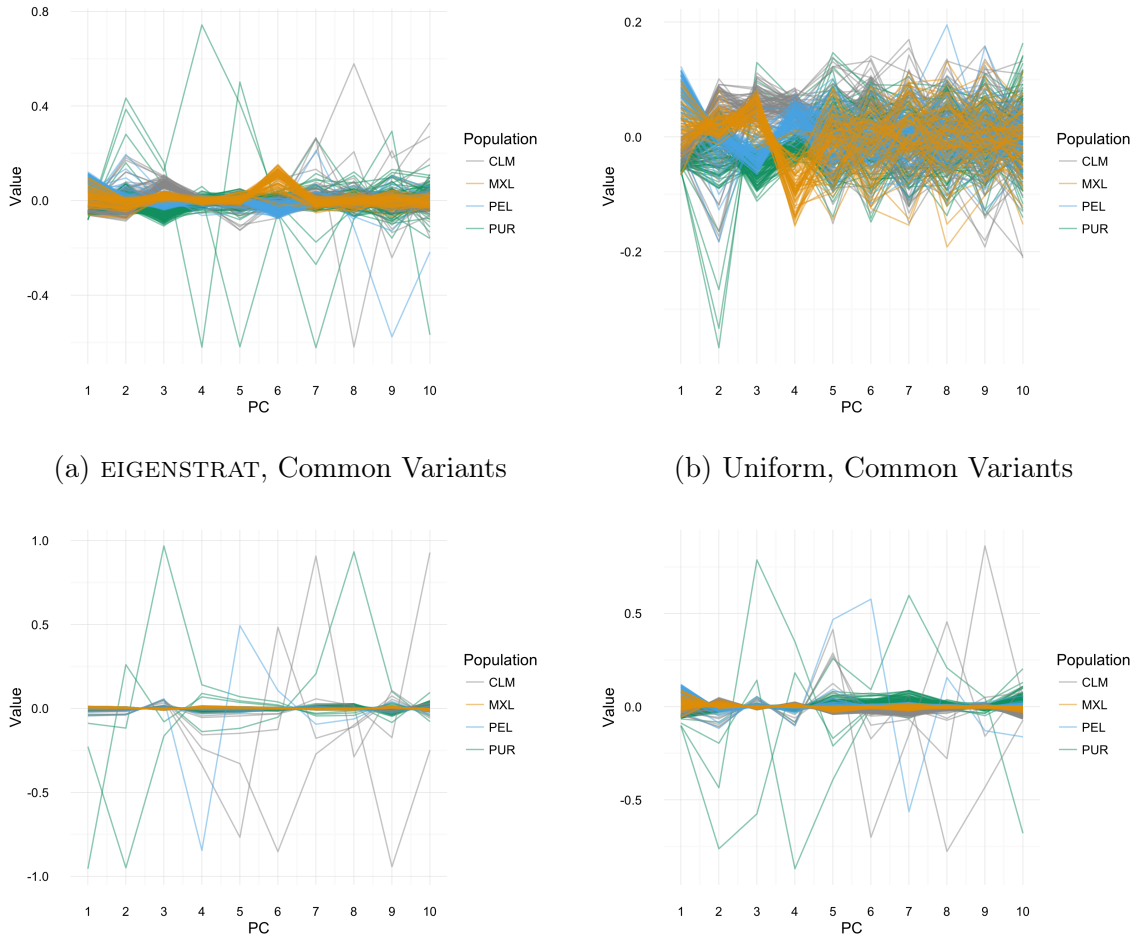
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.2: First and Second Principal Components from the 1000 Genomes Phase 3 American Ancestry Subpopulations: Colombians from Medellin, Colombia (CLM), Mexican Ancestry from Los Angeles, California (MXL), Peruvians from Lima, Peru (PEL), and Puerto Rican (PUR). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



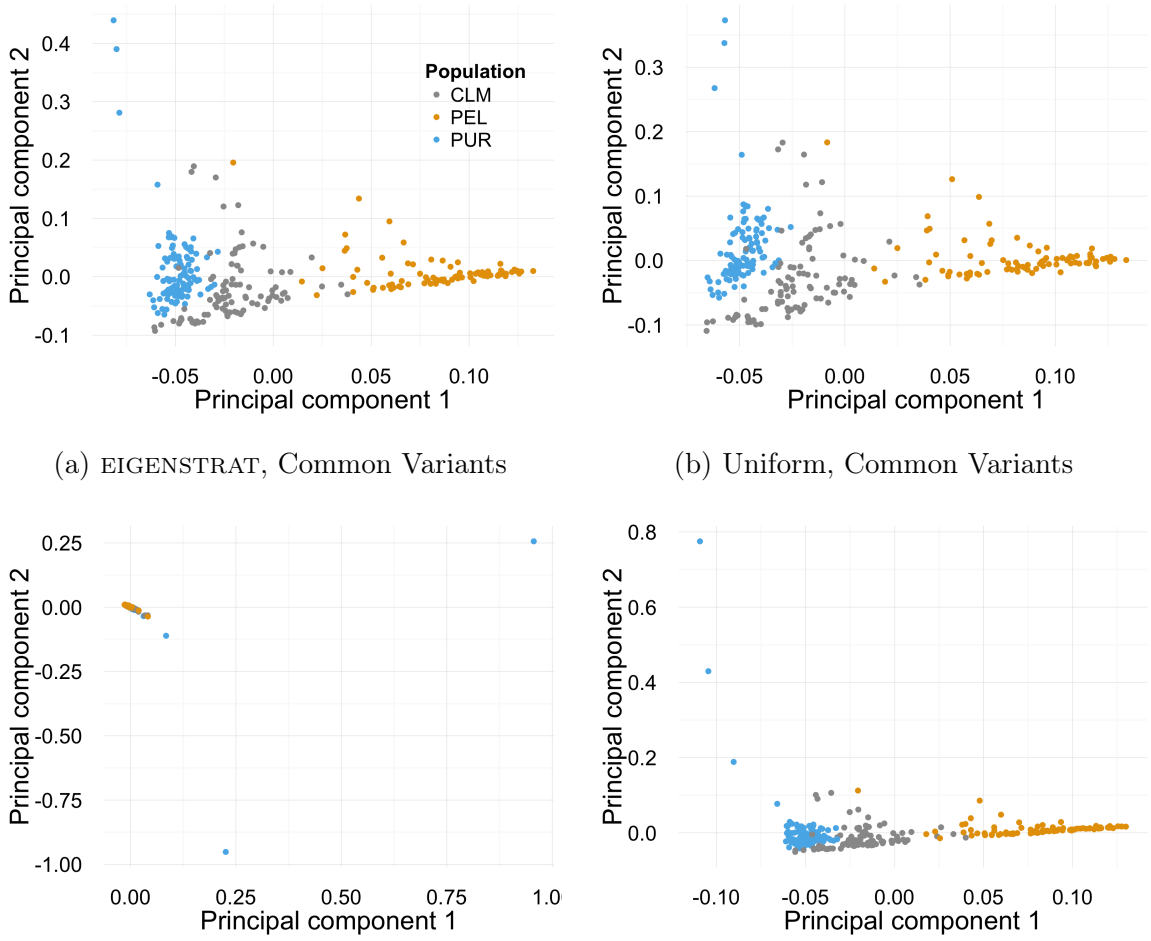
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.3: Top 10 Principal Components from the 1000 Genomes Phase 3 American Ancestry Subpopulations: Colombians from Medellin, Colombia (CLM), Mexican Ancestry from Los Angeles, California (MXL), Peruvians from Lima, Peru (PEL), Puerto Rican (PUR). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



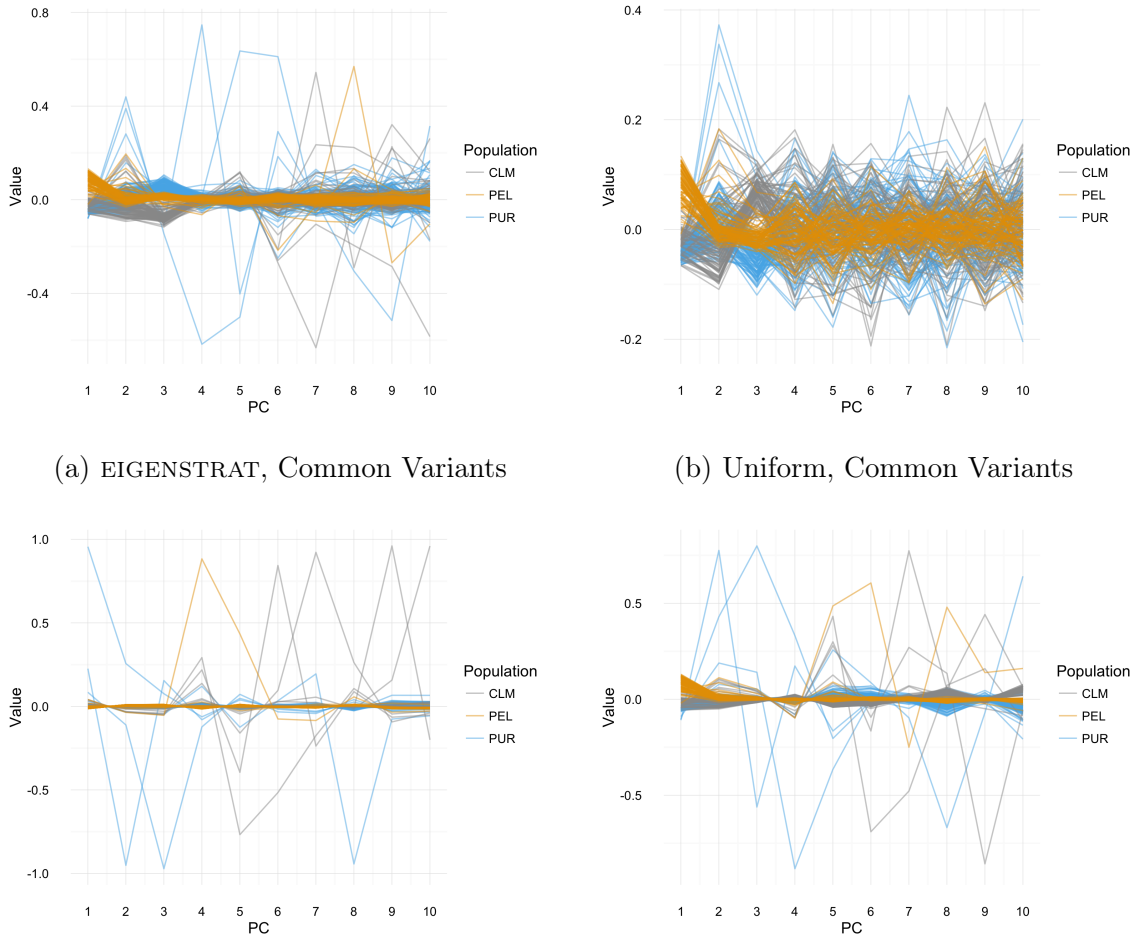
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.4: First and Second Principal Components from the 1000 Genomes Phase 3 American Ancestry Subpopulations: Colombians from Medellin, Colombia (CLM), Peruvians from Lima, Peru (PEL), Puerto Rican (PUR). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



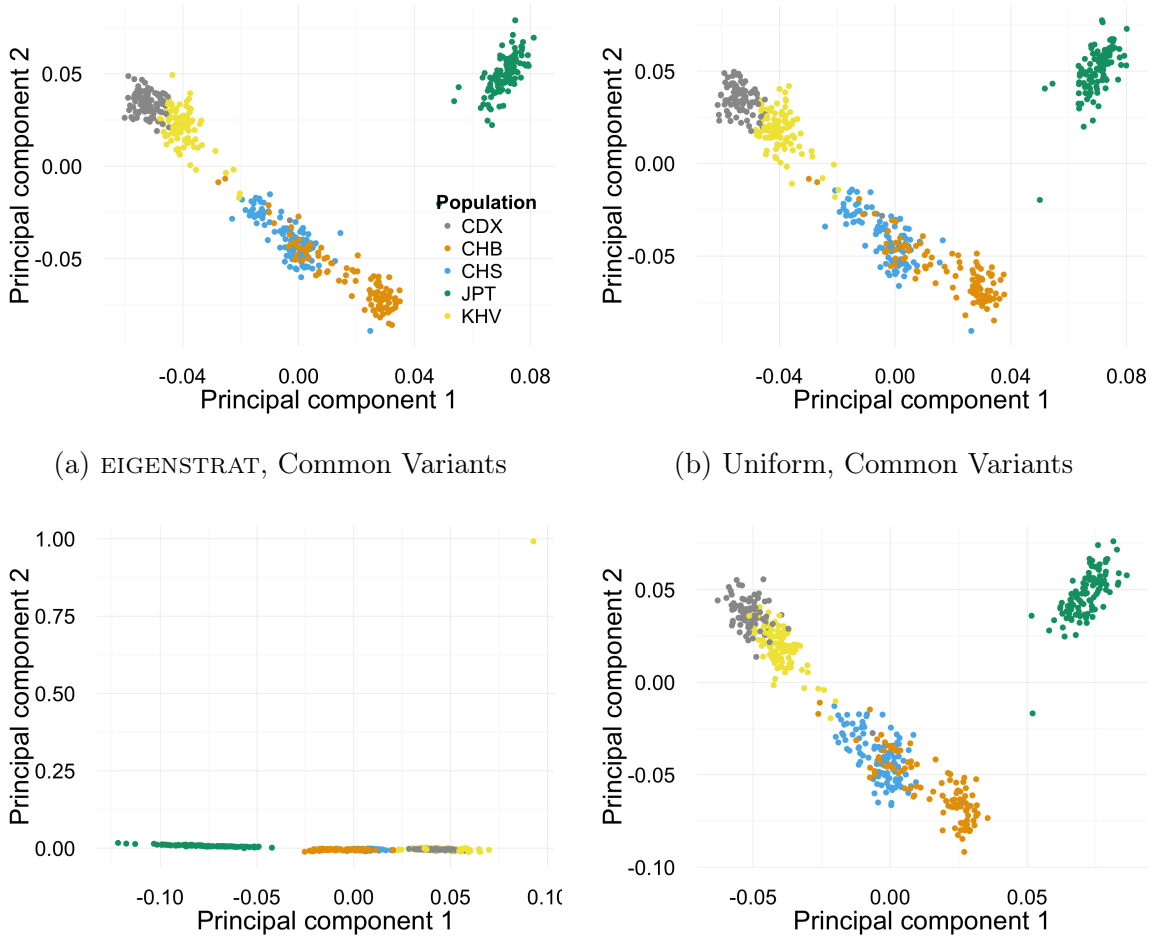
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.5: Top 10 Principal Components from the 1000 Genomes Phase 3 American Ancestry Subpopulations: Colombians from Medellin, Colombia (CLM), Peruvians from Lima, Peru (PEL), Puerto Rican (PUR). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



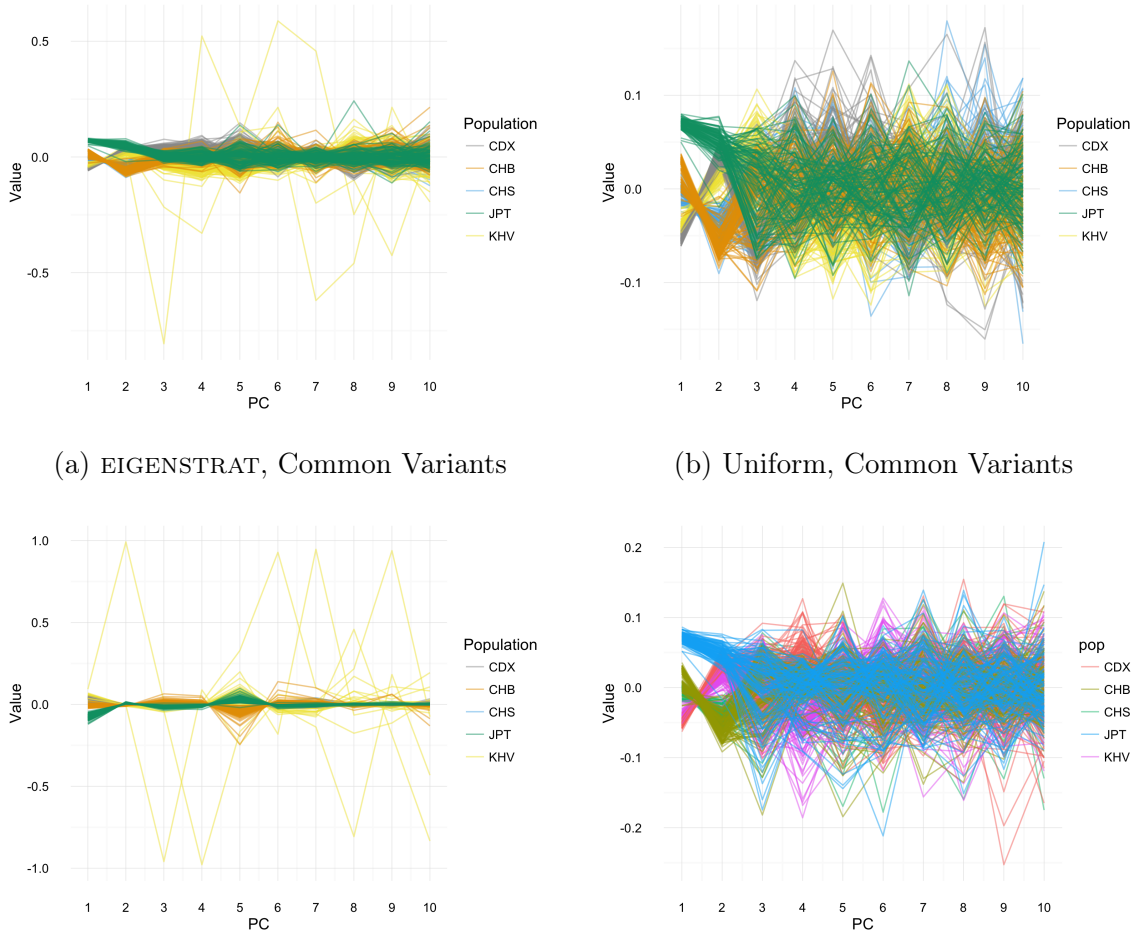
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.6: First and Second Principal Components from the 1000 Genomes Phase 3 East Asian Ancestry Subpopulations: Chinese Dai in Xishuangbanna, China (CDX), Han Chinese in Beijing, China (CHB), Southern Han Chinese (CHS), Japanese in Tokyo, Japan (JPT), and Kihn in Ho Chi Minh City, Vietnam (KHV). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



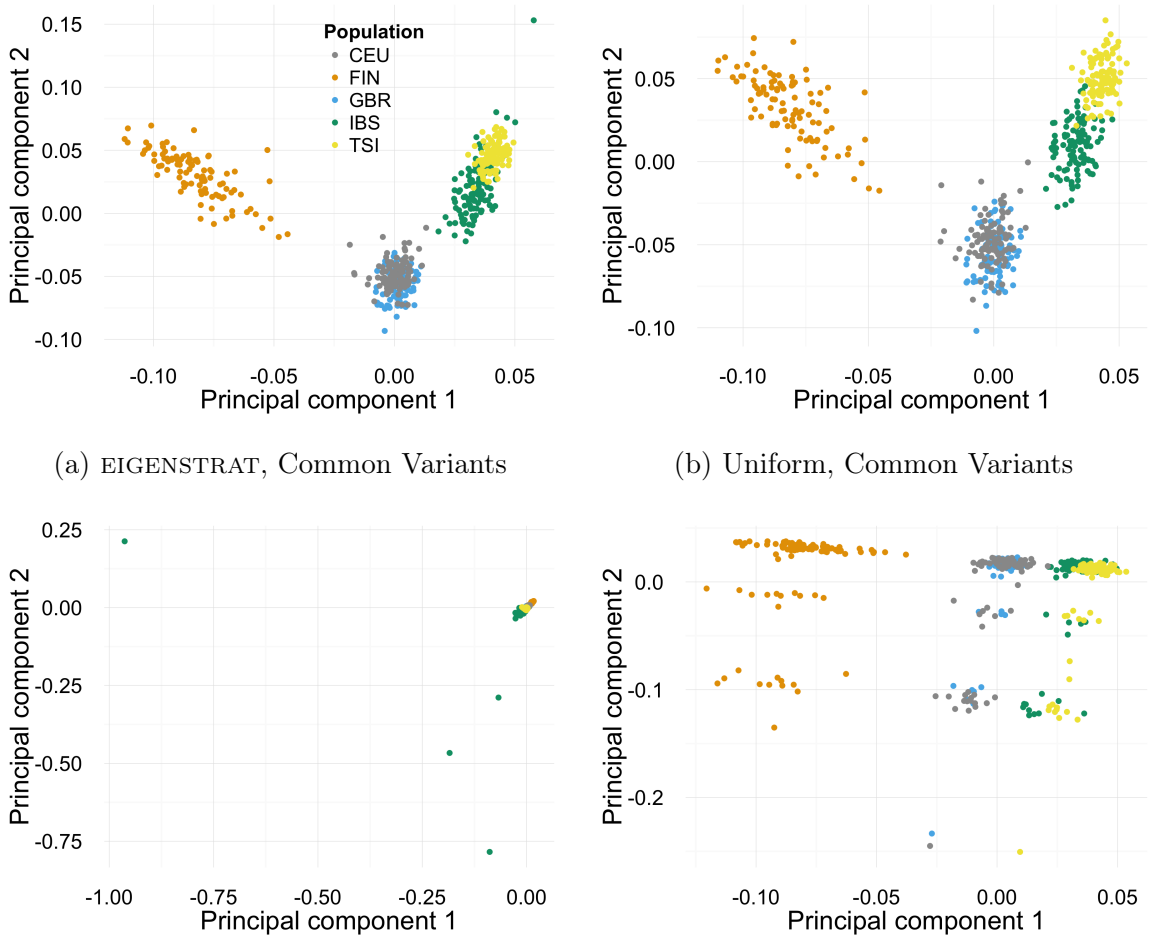
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.7: Top 10 Principal Components from the 1000 Genomes Phase 3 East Asian Ancestry Subpopulations:Chinese Dai in Xishuangbanna, China (CDX), Han Chinese in Beijing, China (CHB), Southern Han Chinese (CHS), Japanese in Tokyo, Japan (JPT), and Kihn in Ho Chi Minh City, Vietnam (KHV). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



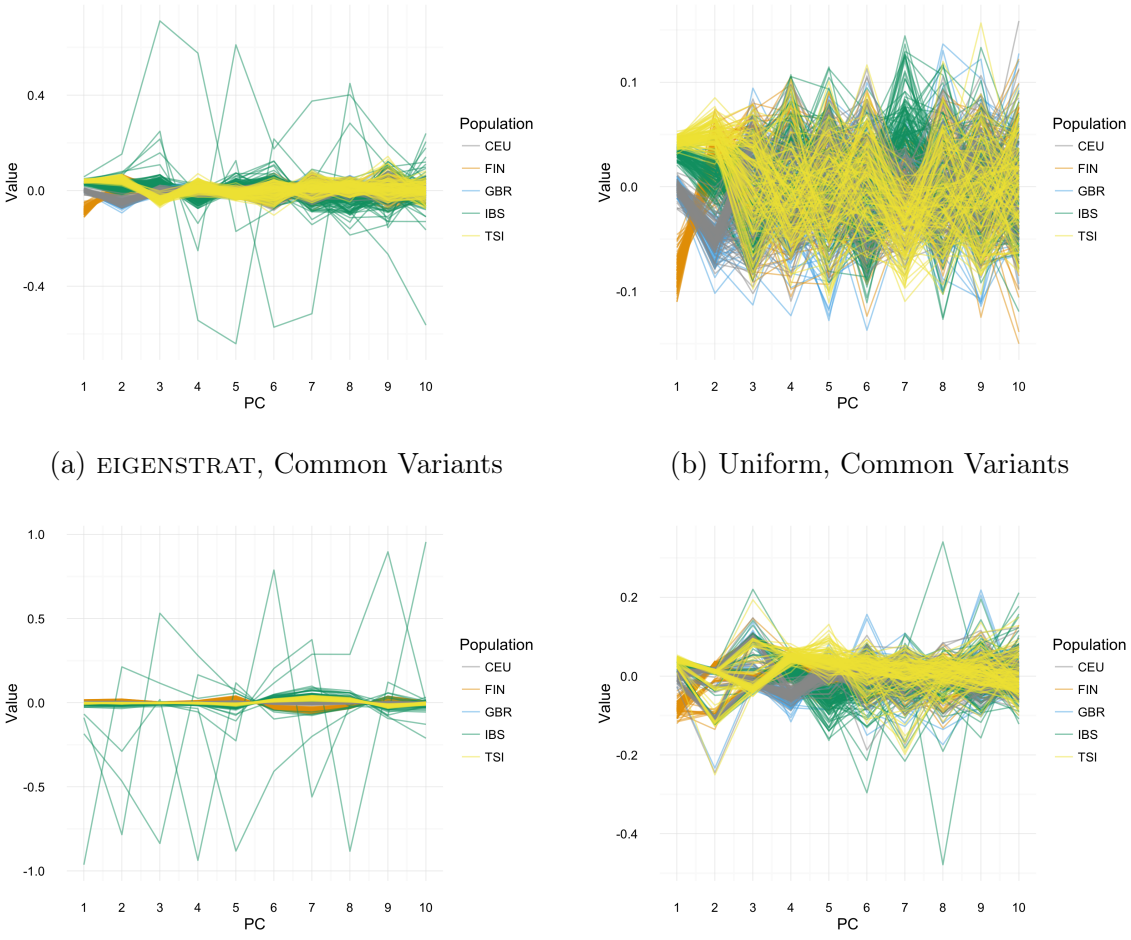
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.8: First and Second Principal Components from the 1000 Genomes Phase 3 European Ancestry Subpopulations: Utah Residents with Northern and Western European Ancestry (CEU), Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian Population in Spain (IBS), and Toscani in Italy (TSI). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



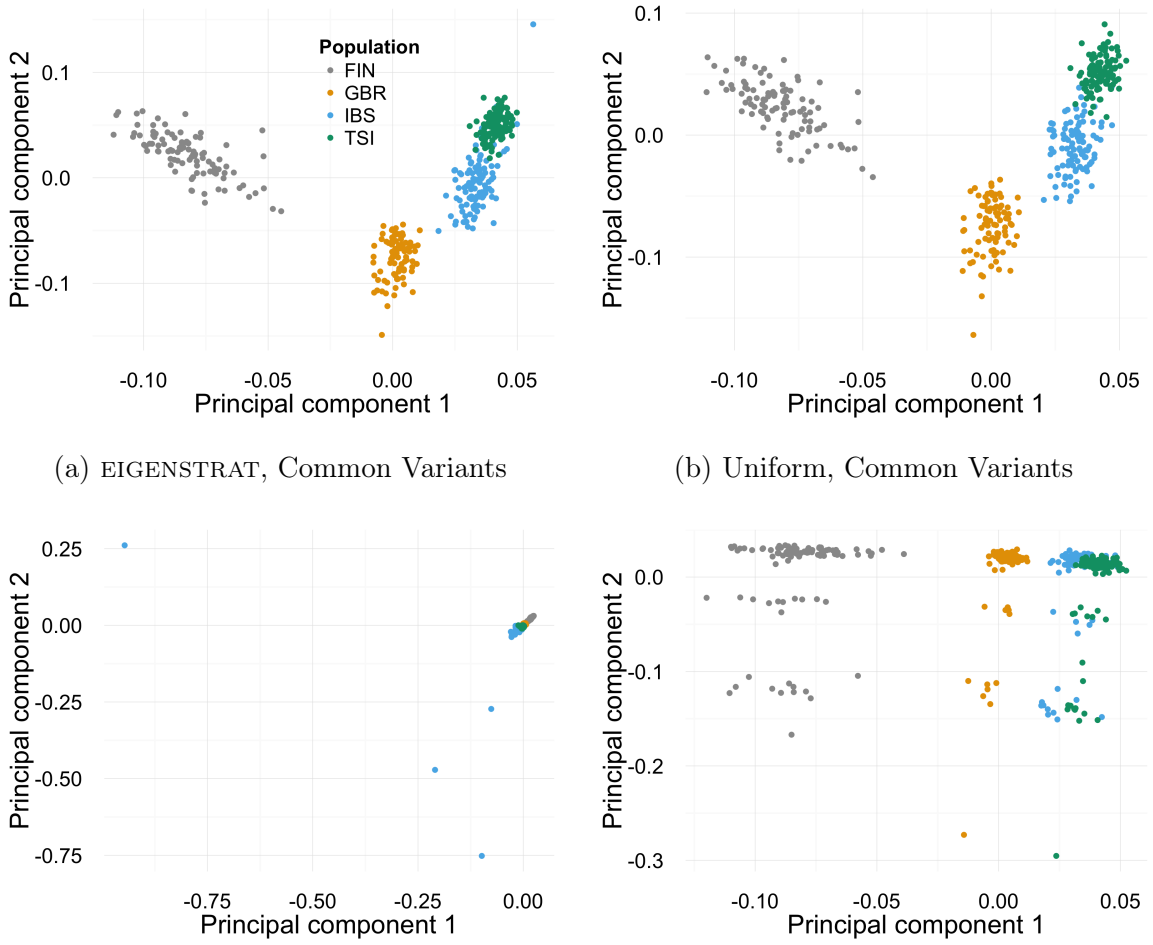
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.9: Top 10 Principal Components from the 1000 Genomes Phase 3 European Ancestry Subpopulations: Utah Residents with Northern and Western European Ancestry (CEU), Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian Population in Spain (IBS), and Toscani in Italy (TSI). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



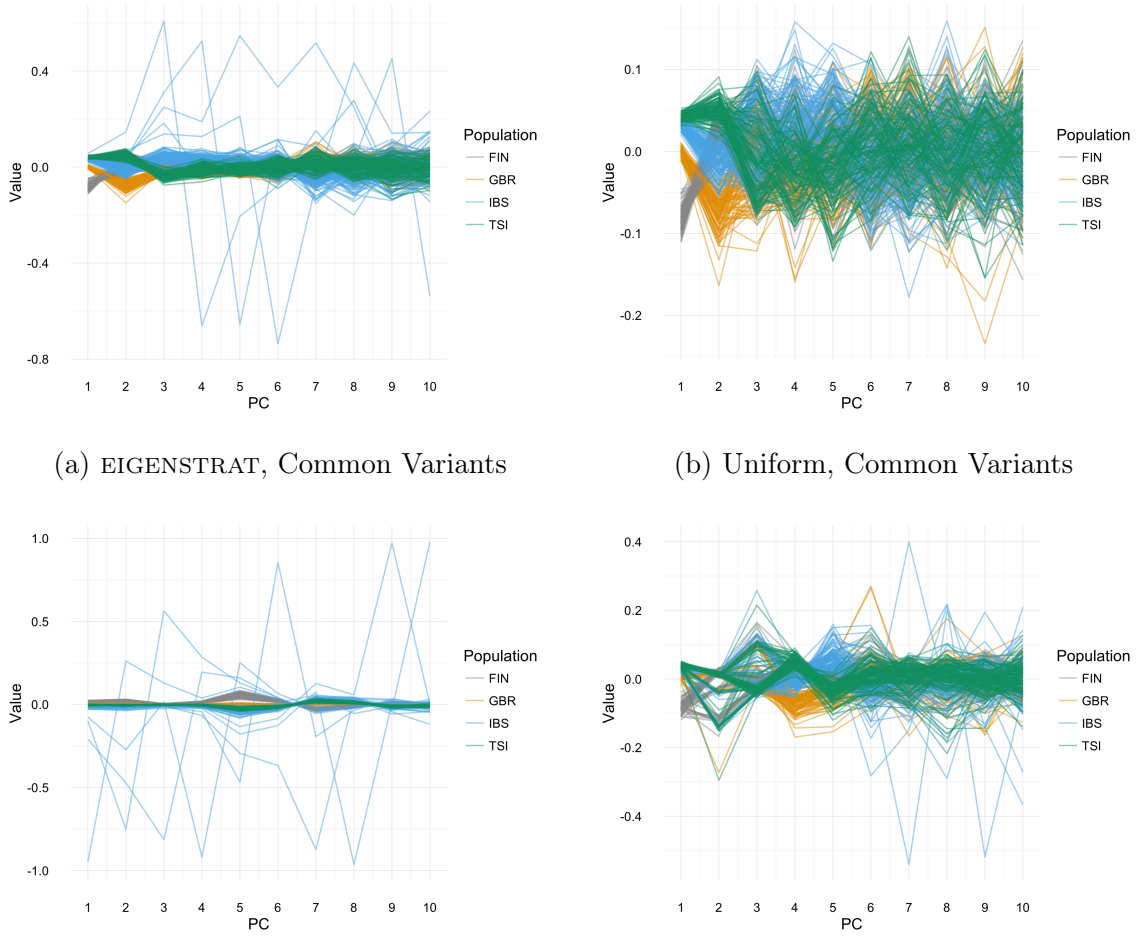
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.10: First and Second Principal Components from the 1000 Genomes Phase 3 European Ancestry Subpopulations:Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian Population in Spain (IBS), and Toscani in Italy (TSI). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



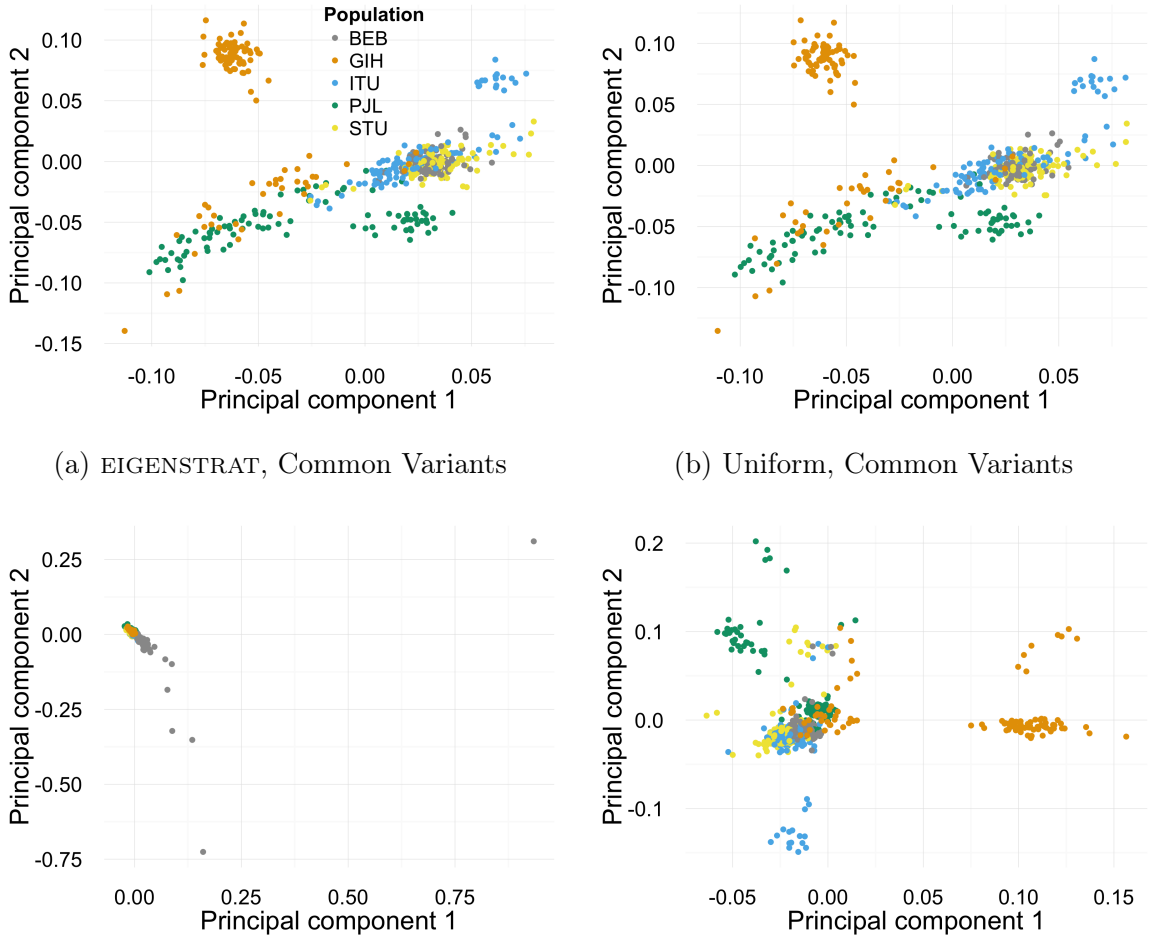
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.11: Top 10 Principal Components from the 1000 Genomes Phase 3 European Ancestry Subpopulations: Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian Population in Spain (IBS), and Toscani in Italy (TSI). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



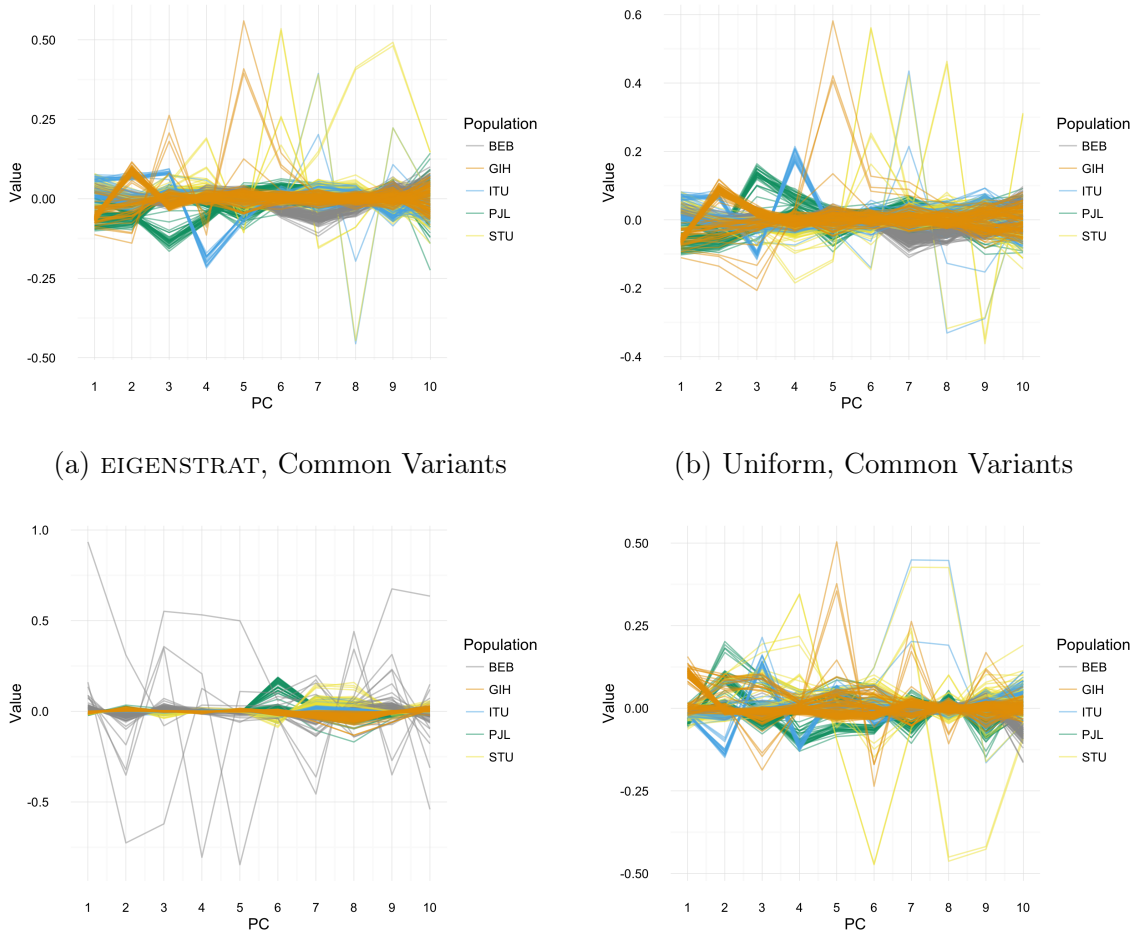
(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.12: First and Second Principal Components from the 1000 Genomes Phase 3 South Asian Ancestry Subpopulations: Bengali from Bangladesh (BEB), Gujarati Indian from Houston, Texas (GIH), Indian Telugu from the United Kingdom (ITU), Punjabi from Lahore, Pakistan (PJL), and Sri Lankan Tamil from the United Kingdom (STU). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.



(a) EIGENSTRAT, Common Variants

(b) Uniform, Common Variants

(c) EIGENSTRAT, Rare Variants

(d) Uniform, Rare Variants

Figure A.13: Top 10 Principal Components from the 1000 Genomes Phase 3 South Asian Ancestry Subpopulations: Bengali from Bangladesh (BEB), Gujarati Indian from Houston, Texas (GIH), Indian Telugu from the United Kingdom (ITU), Punjabi from Lahore, Pakistan (PJL), and Sri Lankan Tamil from the United Kingdom (STU). Figures (a) and (c) are from EIGENSTRAT, while figures (b) and (d) are from PCA-seq with uniform weights. Rare variants are those with MAF less than 0.05, all other variants are common.

A.3 Proofs of Theorems from Chapter 3

In this section, we prove several of the theorems given in Chapter 3. The theorems are restated for convenience.

Theorem A.3.1. *For an $N \times M$ matrix $\tilde{\boldsymbol{G}}$, the principal components of $\tilde{\boldsymbol{G}}$ are given by the eigendecomposition of $\tilde{\boldsymbol{\Phi}} = \tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^{\mathsf{T}}$.*

Proof. Let $\boldsymbol{\eta}_1$ be the loadings from the first principal component. By definition, we wish to find the vector of $\boldsymbol{\eta}_1$ such that we have maximized the variance of the transformed data, subject to the constraint that the squared loadings sum to one. Therefore, we wish to find the $\boldsymbol{\eta}_1$ that maximizes

$$\mathrm{Var}(\boldsymbol{\eta}_1^{\mathsf{T}}\tilde{\boldsymbol{G}}) = \boldsymbol{\eta}_1^{\mathsf{T}}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^{\mathsf{T}}\boldsymbol{\eta}_1,$$

where the variance of of $\boldsymbol{\eta}_1^{\mathsf{T}}\tilde{\boldsymbol{G}}$ is given by the expression on the right if $\tilde{\boldsymbol{G}}$ is centered column-wise. To find the $\boldsymbol{\eta}_1$ that maximizes this expression, subject to the constraint, we will use the Lagrangian and maximize the expression:

$$\boldsymbol{\eta}_1^{\mathsf{T}}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^{\mathsf{T}}\boldsymbol{\eta}_1 - \lambda_1(\boldsymbol{\eta}_1^{\mathsf{T}}\boldsymbol{\eta}_1 - 1).$$

Taking the derivates with respect to $\boldsymbol{\eta}_1$ and λ_1, we have:

$$\frac{\partial}{\partial \boldsymbol{\eta}_1}\left(\boldsymbol{\eta}_1^{\mathsf{T}}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^{\mathsf{T}}\boldsymbol{\eta}_1 - \lambda_1(\boldsymbol{\eta}_1^{\mathsf{T}}\boldsymbol{\eta}_1 - 1)\right) = 2\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^{\mathsf{T}}\boldsymbol{\eta}_1 - \lambda_1\boldsymbol{\eta}_1$$
$$\frac{\partial}{\partial \lambda_1}\left(\boldsymbol{\eta}_1\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^{\mathsf{T}}\boldsymbol{\eta}_1 - \lambda_1(\boldsymbol{\eta}_1^{\mathsf{T}}\boldsymbol{\eta}_1 - 1)\right) = -(\boldsymbol{\eta}_1^{\mathsf{T}}\boldsymbol{\eta}_1 - 1)$$

Setting the above expressions equal to zero and rearranging slightly, we have the two equations:

$$\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^{\mathsf{T}}\boldsymbol{\eta}_1 = \lambda_1\boldsymbol{\eta}_1$$
$$\boldsymbol{\eta}_1^{\mathsf{T}}\boldsymbol{\eta}_1 = 1,$$

where we have divided both sides of the first equation by 2 and absorbed a factor of $\frac{1}{2}$ into λ_1. The $\boldsymbol{\eta}_1$ that satisfies the above equations satisfies both the original constraint, and meets

the definition of an eigenvector with eigenvalue λ_1 for the matrix $\tilde{\mathbf{\Phi}} = \tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}$. Therefore, the first principal component of $\tilde{\boldsymbol{G}}$ is an eigenvector of its empirical covariance matrix.

To find the second eigenvector, we need to repeat the same process with the additional constraint that $\boldsymbol{\eta}_2$ is orthogonal to $\boldsymbol{\eta}_1$. Therefore, we again use the Lagrangian:

$$\boldsymbol{\eta}_2^\mathsf{T}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}\boldsymbol{\eta}_2 - \lambda_2(\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_2 - 1) - \lambda_2'(\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_1 - 0).$$

Taking the derivatives with respect to $\boldsymbol{\eta}_2$, λ_2 and λ_2' we have:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}_2}\left(\boldsymbol{\eta}_2^\mathsf{T}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}\boldsymbol{\eta}_2 - \lambda_2(\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_2 - 1) - \lambda_2'(\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_1 - 0)\right) &= 2\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}\boldsymbol{\eta}_2 - \lambda_2\boldsymbol{\eta}_2 - \lambda_2'\boldsymbol{\eta}_1 \\
\frac{\partial}{\partial \lambda_2}\left(\boldsymbol{\eta}_2^\mathsf{T}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}\boldsymbol{\eta}_2 - \lambda_2(\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_2 - 1) - \lambda_2'(\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_1 - 0)\right) &= -(\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_2 - 1) \\
\frac{\partial}{\partial \lambda_2'}\left(\boldsymbol{\eta}_2^\mathsf{T}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}\boldsymbol{\eta}_2 - \lambda_2(\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_2 - 1) - \lambda_2'(\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_1 - 0)\right) &= -(\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_1 - 0)
\end{aligned}$$

Setting each of the equations equal to zero, and rearranging slightly, we have

$$\begin{aligned}
2\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}\boldsymbol{\eta}_2 - \lambda_2\boldsymbol{\eta}_2 - \lambda_2'\boldsymbol{\eta}_1 &= 0 \\
\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_2 &= 1 \\
\boldsymbol{\eta}_2^\mathsf{T}\boldsymbol{\eta}_1 &= 0
\end{aligned}$$

The second two equations have yielded the constraints. If we take the first equation and multiply it by $\boldsymbol{\eta}_1^\mathsf{T}$, we have:

$$2\boldsymbol{\eta}_1^\mathsf{T}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}\boldsymbol{\eta}_2 - \lambda_2\boldsymbol{\eta}_1^\mathsf{T}\boldsymbol{\eta}_2 - \lambda_2'\boldsymbol{\eta}_1^\mathsf{T}\boldsymbol{\eta}_1 = 0$$

From the constraints, we know the second term is zero, and the third term is λ_2'. Furthermore, we know that the project of the original data $\tilde{\boldsymbol{G}}$ under each principal component is uncorrelated and therefore has covariance zero. This implies that

$$\mathrm{Cov}(\boldsymbol{\eta}_1^\mathsf{T}\tilde{\boldsymbol{G}}, \boldsymbol{\eta}_2^\mathsf{T}\tilde{\boldsymbol{G}}) = \boldsymbol{\eta}_1^\mathsf{T}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^\mathsf{T}\boldsymbol{\eta}_2 = 0.$$

Therefore, the previous equation becomes

$$\lambda_2' = 0.$$

Substituting this back into the original equation derived from the differentiating the Lagrangian and rearranging, we have

$$\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^{\mathsf{T}}\boldsymbol{\eta}_2 = \lambda_2\boldsymbol{\eta}_2$$
$$\boldsymbol{\eta}_2^{\mathsf{T}}\boldsymbol{\eta}_2 = 1$$
$$\boldsymbol{\eta}_2^{\mathsf{T}}\boldsymbol{\eta}_1 = 0$$

Therefore, we have demonstrated that the second principal component is an eigenvector of $\tilde{\boldsymbol{\Phi}}$. The rest of the principal components can be derived in a similar manner. □

Theorem A.3.2. *For a square matrix $\tilde{\boldsymbol{\Phi}}$, the eigendecomposition of $\tilde{\boldsymbol{\Phi}}$ is equivalent to the spectral decomposition of $\tilde{\boldsymbol{\Phi}}$.*

Proof. Let the spectral decomposition of an N by N matrix $\tilde{\boldsymbol{\Phi}}$ be $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}$, where $\boldsymbol{U}$ is an orthogonal matrix and $\boldsymbol{D}$ is a diagonal matrix. Let $\boldsymbol{u}_1 \dots \boldsymbol{u}_N$ be the column vectors of $\boldsymbol{U}$. Then

$$\begin{aligned}\boldsymbol{\Phi}\boldsymbol{u}_j &= \left(\sum_{i=1}^{N} d_i\boldsymbol{u}_i\boldsymbol{u}_i^{\mathsf{T}}\right)\boldsymbol{u}_j \\ &= \sum_{i=1}^{N} d_i\boldsymbol{u}_i\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_j \\ &= d_j\boldsymbol{u}_j\boldsymbol{u}_j^{\mathsf{T}}\boldsymbol{u}_j \\ &= d_j\boldsymbol{u}_j,\end{aligned}$$

since the column vectors of $\boldsymbol{U}$ are orthogonal, implying that $\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_j = 1$ for $i \neq j$. As $\tilde{\boldsymbol{\Phi}}\boldsymbol{u}_j = d_j\boldsymbol{u}_j$ for all j and the columns of $\boldsymbol{U}$ are orthogonal, the columns of $\boldsymbol{U}$ are the set of eigenvectors for $\tilde{\boldsymbol{\Phi}}$ and the diagonal elements of $\boldsymbol{D}$ are the eigenvalues of $\tilde{\boldsymbol{\Phi}}$. □

Theorem A.3.3. *For a square matrix $\tilde{\boldsymbol{\Phi}}$, the spectral decomposition of $\tilde{\boldsymbol{\Phi}}$ is given by the* SVD *of $\tilde{\boldsymbol{\Phi}}$.*

Proof. Let the spectral decomposition of an N by N matrix $\tilde{\boldsymbol{\Phi}}$ be $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}$, where $\boldsymbol{U}$ is an orthogonal matrix and $\boldsymbol{D}$ is a diagonal matrix. Let □

Theorem A.3.4. *Given a matrix $\tilde{\boldsymbol{G}}$ with M rows and N columns, the matrix $\tilde{\boldsymbol{G}}^{\mathsf{T}}\tilde{\boldsymbol{G}}$ is symmetric.*

Proof. If $\tilde{\boldsymbol{G}}$ has dimensions M by N, then $\tilde{\boldsymbol{G}}^{\mathsf{T}}$ has dimensions N by M, and $\tilde{\boldsymbol{G}}^{\mathsf{T}}X$ has dimensions N by N. Furthermore,

$$\begin{aligned}(\tilde{\boldsymbol{G}}^{\mathsf{T}}\tilde{\boldsymbol{G}})^{\mathsf{T}} &= \tilde{\boldsymbol{G}}^{\mathsf{T}}(\tilde{\boldsymbol{G}}^{\mathsf{T}})^{\mathsf{T}} \\ &= \tilde{\boldsymbol{G}}^{\mathsf{T}}\tilde{\boldsymbol{G}}.\end{aligned}$$

□

Table A.1: Asymptotic Time Complexity of Mathematical Operations

Operation	Asymptotic Time Complexity
Matrix Multiplication: $[N \times M][M \times P]$	$\mathrm{O}(NMP)$
QR-decomposition: $[N \times M]$	$\mathrm{O}(NM \min\{N, M\})$
SVD: $[N \times M]$	$\mathrm{O}(NM \min\{N, M\})$
Eigendecomposition: $[N \times M]$	$\mathrm{O}(NM \min\{N, M\})$
Column-wise ℓ_2-norm standardization: $[N \times M]$	$\mathrm{O}(NM)$
Random multivariate normal matrix: $[N \times M]$	$\mathrm{O}(NM)$

A.4 Asymptotic Time Complexity Derivations

For convenience, we repeat the table of asymptotic complexities shown in Chapter 3.

A.4.1 EIGENSTRAT

The exact EIGENSTRAT algorithm consists of two steps, construct the GRM and then take the eigendecomposition of the GRM. Constructing the GRM requires a single matrix multiplication, which as complexity $\mathrm{O}(N^2M)$. The eigendecomposition has complexity $\mathrm{O}(N^3)$, so the total complexity of EIGENSTRAT is $\mathrm{O}(N^3 + N^2M)$.

A.4.2 `flashpca`

Most steps in `flashpca` have complexities that are directly computed from the complexity of the corresponding mathematical operation, provided that the dimension of the matrix is known. The complexity of steps 1 through 4 and 9-15 are all easily calculated in this manner, provided the dimension of $\boldsymbol{Q}$ is known. Therefore, we need only workout the dimension of $\boldsymbol{Q}$ and address the complexity of the iteration in steps 5 through 8.

Step 6 consists of multiplying the GRM, an $N \times N$ matrix, by $\boldsymbol{A}^{(c-1)}$, an $N \times (k + l)$

Table A.2: Asymptotic Complexity of the `flashpca` Algorithm

Step	Asymptotic Time Complexity
1	$\mathrm{O}(N^2M)$
2	$\mathrm{O}(M(k+l))$
3	$\mathrm{O}(NM(k+l))$
4	$\mathrm{O}(NM)$
5-8	$\mathrm{O}(C(N^2(k+l)+NM))$
9	$\mathrm{O}(N^2M)$
10	$\mathrm{O}(NM(k+l))$
11	$\mathrm{O}(M(k+l)^2)$
12	$\mathrm{O}((k+l)^3)$
13	$\mathrm{O}(N(k+l)^2)$
14	$\mathrm{O}(N)$
15	$\mathrm{O}(N(k+l)^2)$
Total	$\mathrm{O}(N^2M+NM+M)$

matrix. Therefore, step 6 has complexity $\mathrm{O}(N^2(k+l))$. Step 7, which is the column-wise normalization of an $N\times(k+l)$ matrix, has complexity $\mathrm{O}(N(k+l))$. Therefore, since these two steps are repeated C times, steps 5–8 have a total complexity of $\mathrm{O}(C(N^2(k+l)+N(k+l)))$.

In step 9, the QR-decomposition of $\boldsymbol{A}^{(C)}$, which is an $N\times(k+l)$ matrix yields a $\boldsymbol{Q}$ matrix with dimensions $N\times N$. Therefore, we have the complexities give in Table 3.2 and which are repeated in the table below for reference.

Combining these, we have a total complexity of

$$O(2N^2M + M(k+l) + 2NM(k+l) + NM + CN^2(k+l) + CNM \\ + M(k+l)^2 + (k+l)^3 + N + 2N(k+l)^2).$$

Writing this as a polynomial in N, so we can compare to the complexity of EIGENSTRAT, we have

$$O(N^2[2M + C(k+l)] + N[M\{2(k+l) + C + 1\} + 2(k+l)^2 + 1] + M[(k+l)^2 + (k+l)]).$$

Since k, l and C will be very small relative to M and small relative to N, we can ignore these constants, to find

$$O(N^2M + NM + M).$$

A.4.3 `FastPCA`

As with `flashpca`, most steps in `FastPCA` have asymptotic time complexities that are easily derived. The asymptotic time complexity of steps 1-2 and 8-10 are easily calculated in this manner, provided we have the dimensions of $\boldsymbol{A}$. Note that step 7, the construction of $\boldsymbol{A}$ does not necessarily add any more time to the computation, as the subsequent steps can be performed without explicitly forming $\boldsymbol{A}$.

Step 4 consists of multiplying an $N \times M$ matrix by an $M \times (k+l)$ matrix, which has complexity $\mathrm{O}(NM(k+l))$, and step 5 is the multiplication of an $M \times N$ matrix by an $N \times (k+l)$ matrix, which has complexity $\mathrm{O}(NM(k+l))$. Therefore, steps 3-6 have complexity $\mathrm{O}(2CNM(k+l))$.

Matrix $\boldsymbol{A}$ is an $M \times (C+1)(k+l)$, so step 8 has computational complexity $\mathrm{O}(M(C+1)^2(k+l)^2)$. Matrix $\boldsymbol{U}$ from the SVD of $\boldsymbol{A}$, is an $M \times M$ matrix, but only the first $(C+1)(k+l)$ columns of $\boldsymbol{U}$ are non-zero, since $\boldsymbol{U}$ forms an orthonormal basis of column vectors. Therefore, in step 9, $\boldsymbol{U}$ can be truncated and the asymptotic cost of the multiplication is $O(NM(C+1)(k+l))$. Finally, in step 10, since $\boldsymbol{B}$ is an $(C+1)(k+l) \times N$ matrix, this

Table A.3: Asymptotic Complexity of the `FastPCA` Algorithm

Step	Asymptotic Time Complexity
1	$\mathrm{O}(N[k+l])$
2	$\mathrm{O}(NM[k+l])$
3–6	$\mathrm{O}(2CNM[k+l])$
7	—
8	$\mathrm{O}(M[C+1]^2[k+l]^2)$
9	$\mathrm{O}(NM[C+1][k+l])$
10	$\mathrm{O}(N[C+1][k+l]^2)$
Total	$\mathrm{O}(NM+M)$

step has asymptotic complexity $\mathrm{O}(N(C+1)^2(k+l)^2)$, assuming $(C+1)(k+l)$ is less than N, which is a reasonable assumption for a large sequencing study.

Combining these asymptotic time complexities, we have that `FastPCA` has overall time complexity of

$$O(N(k+l) + NM(k+l) + 2CNM(k+l) + M(C+1)^2(k+l)^2 + M(C+1)(k+l)).$$

Rearranging this to create a polynomial expression in N, we have

$$O(N[M\{2C(k+l) + (k+l)\} + (k+l)] + M[(C+1)^2(k+l)^2 + (C+1)(k+l)]).$$

Again, assuming that k, l, and C are very small relative to M and small relative to N, we have that the asymptotic time complexity of `FastPCA` is

$$O(NM+M)$$

Table A.4: Asymptotic Complexity of the General SSVD Algorithm

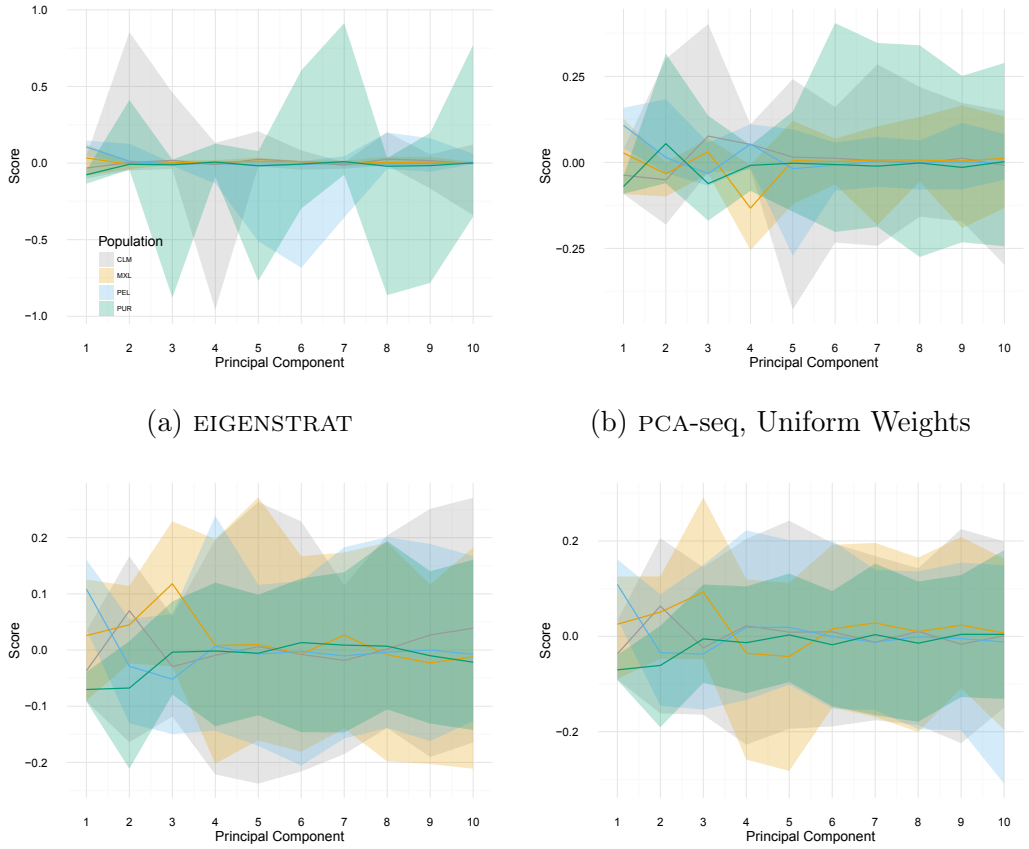
Step	Asymptotic Time Complexity
1	$\mathrm{O}(M[k+l])$
2	$\mathrm{O}(NM[k+l])$
3	$\mathrm{O}(M[k+l]^2)$
4–9	$\mathrm{O}(2CNM[k+l]+C[N+M][k+l])$
10	$\mathrm{O}(NM[k+l])$
11	$\mathrm{O}(N[k+l]^2)$
Total	$\mathrm{O}(NM+M+N)$

A.4.4 Fast PCA-*seq*

Most steps in fast PCA-seq have asymptotic time complexities that are easily derived. Combining these asymptotic time complexities, as we did for the previous estimators, we have that fast PCA-seq has overall time complexity of

$$O(NM+M+N).$$

A.5 PCA-seq with Mutual Information Weights: 1000 Genome Project Application



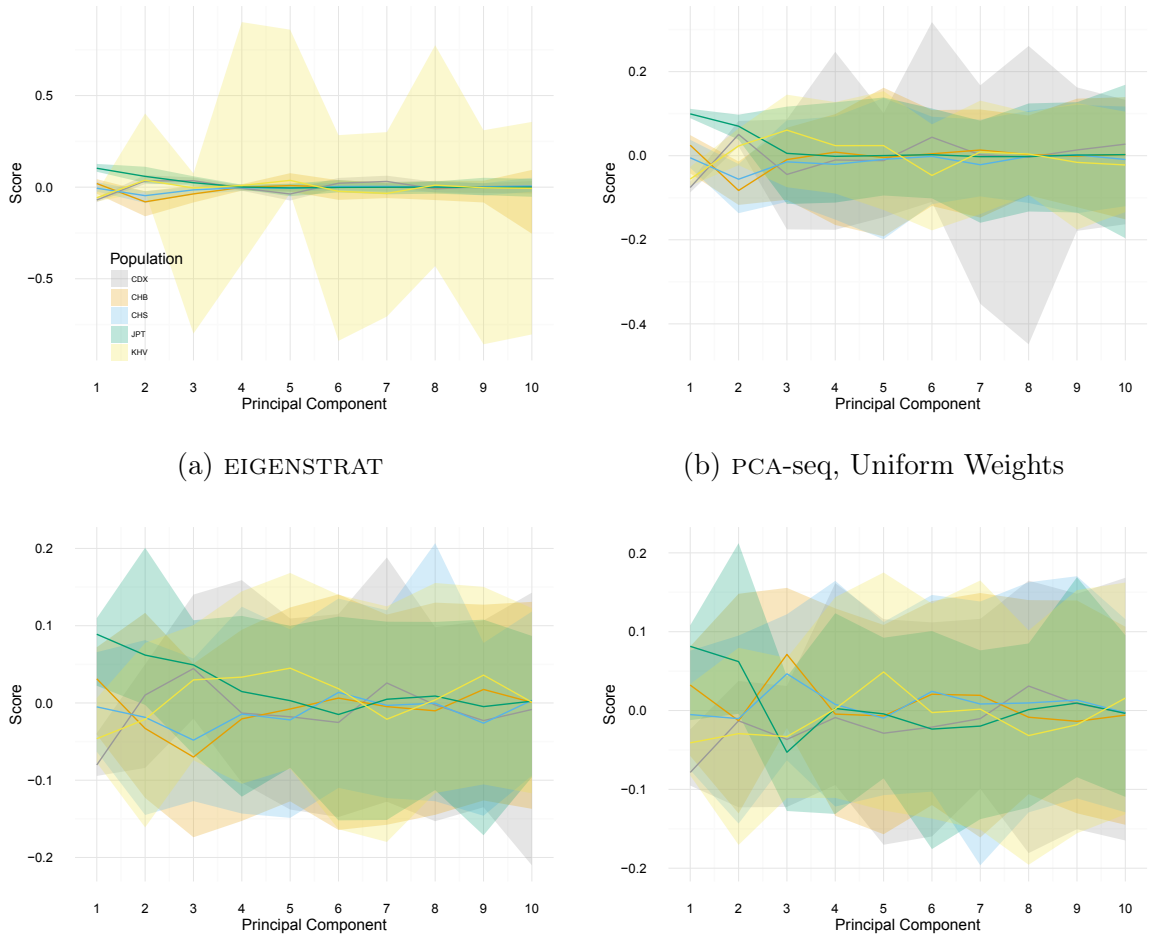
(a) EIGENSTRAT

(b) PCA-seq, Uniform Weights

(c) PCA-seq, Empirical MI Weights

(d) PCA-seq, Miller-Madow MI Weights

Figure A.14: The mean and range of the first 10 four principal components from EIGENSTRAT and PCA-seq with uniform and mutual information weights applied to the 1000 Genomes Phase 3 American super-population: Colombians from Medellin, Colombia (CLM), Mexican Ancestry from Los Angeles, California (MXL), Peruvians from Lima, Peru (PEL), Puerto Rican (PUR). The solid lines represent the average principal component value within each population, while the shading represents the range (minimum to maximum). (MI: Mutual Information)



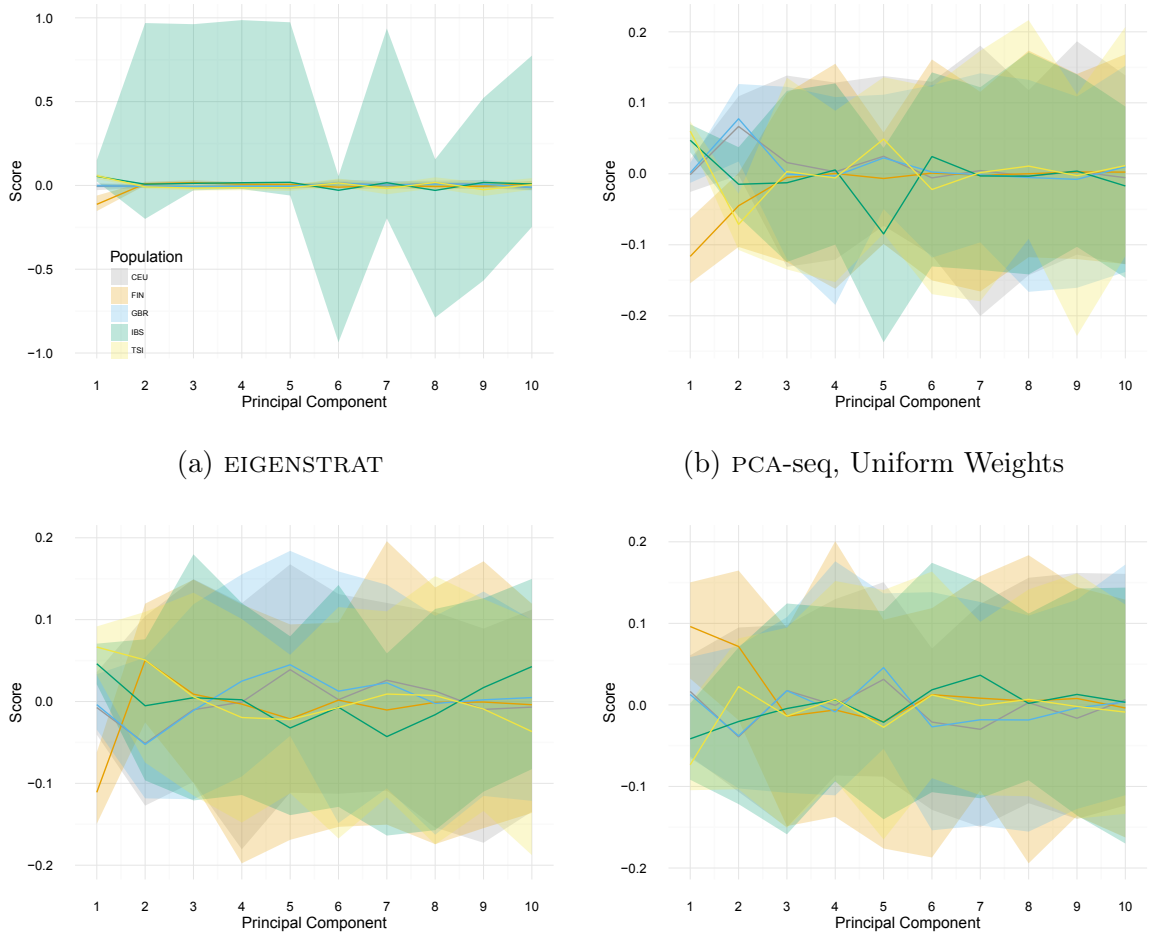
(a) EIGENSTRAT

(b) PCA-seq, Uniform Weights

(c) PCA-seq, Empirical MI Weights

(d) PCA-seq, Miller-Madow MI Weights

Figure A.15: The mean and range of the first 10 four principal components from EIGENSTRAT and PCA-seq with uniform and mutual information weights applied to the 1000 Genomes Phase 3 East Asian super-population: Chinese Dai in Xishuangbanna, China (CDX), Han Chinese in Beijing, China (CHB), Southern Han Chinese (CHS), Japanese in Tokyo, Japan (JPT), Kihn in Ho Chi Minh City, Vietnam (KHV). The solid lines represent the average principal component value within each population, while the shading represents the range (minimum to maximum). (MI: Mutual Information)



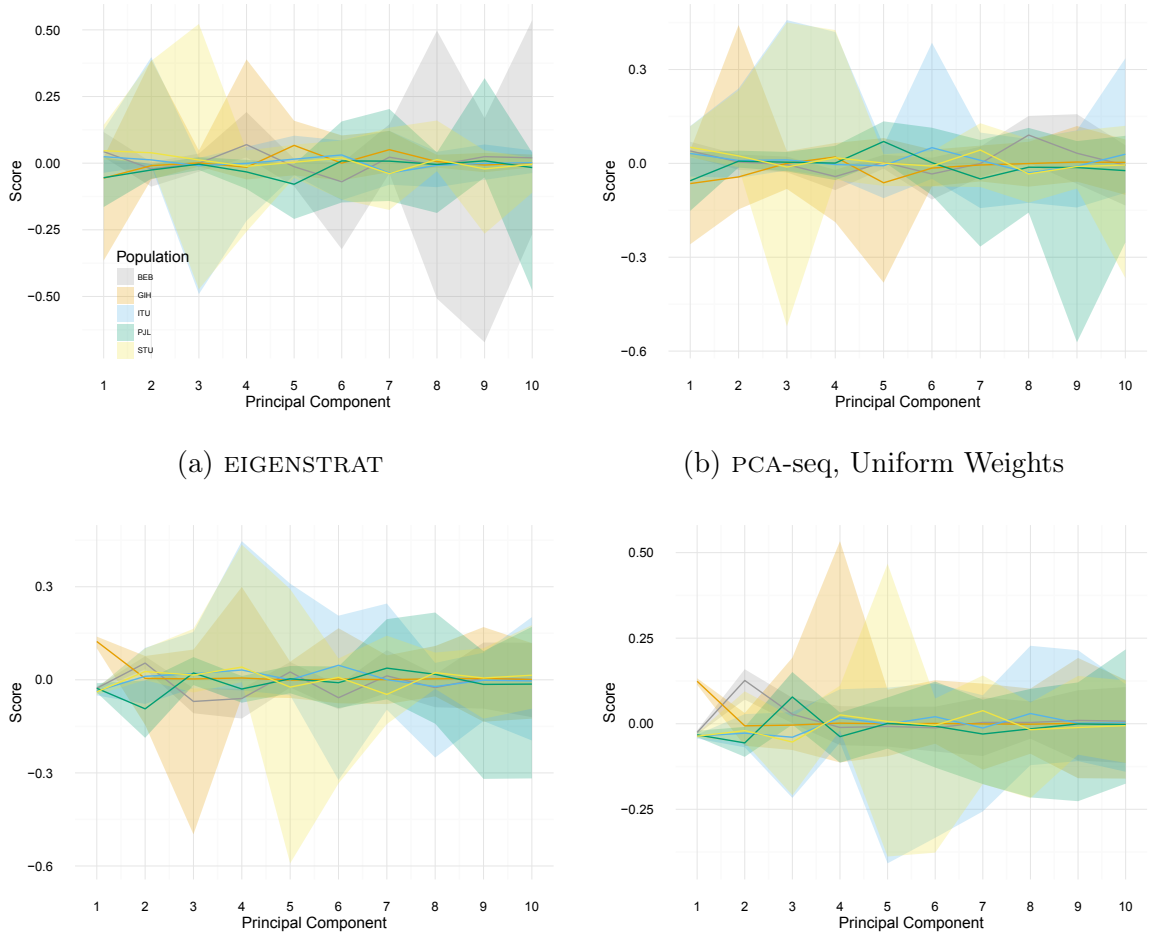
(a) EIGENSTRAT

(b) PCA-seq, Uniform Weights

(c) PCA-seq, Empirical MI Weights

(d) PCA-seq, Miller-Madow MI Weights

Figure A.16: The mean and range of the first 10 four principal components from EIGENSTRAT and PCA-seq with uniform and mutual information weights applied to the 1000 Genomes Phase 3 European super-population: Utah Residents with Northern and Western European Ancestry (CEU), Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian Population in Spain (IBS), Toscani in Italy (TSI). The solid lines represent the average principal component value within each population, while the shading represents the range (minimum to maximum). (MI: Mutual Information)



(a) EIGENSTRAT

(b) PCA-seq, Uniform Weights

(c) PCA-seq, Empirical MI Weights

(d) PCA-seq, Miller-Madow MI Weights

Figure A.17: The mean and range of the first 10 four principal components from EIGENSTRAT and PCA-seq with uniform and mutual information weights applied to the 1000 Genomes Phase 3 South Asian super-population: Bengali from Bangladesh (BEB), Gujarati Indian from Houston, Texas (GIH), Indian Telugu from the United Kingdom (ITU), Punjabi from Lahore, Pakistan (PJL), Sri Lankan Tamil from the United Kingdom (STU). The solid lines represent the average principal component value within each population, while the shading represents the range (minimum to maximum). (MI: Mutual Information)