Sparse Principal Component Analysis for Identifying Ancestry-Informative Markers in Genome Wide Association Studies

Seokho Lee[1], Michael P. Epstein[2], Richard Duncan[2] and Xihong Lin[3,*]

[1]Department of Statistics, Hankuk University of Foreign Studies, Yongin, Korea

[2]Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia

[3]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts

email: xlin@hsph.harvard.edu

SUMMARY: Genomewide association studies (GWAS) routinely apply principal component analysis (PCA) to infer population structure within a sample to correct for confounding due to ancestry. GWAS implementation of PCA uses tens of thousands of SNPs to infer structure, despite the fact that only a small fraction of such SNPs provides useful information on ancestry. The identification of this reduced set of ancestry-informative markers (AIMs) from a GWAS has practical value; for example, researchers can genotype the AIM set to correct for potential confounding due to ancestry in follow-up studies that utilize custom SNP or sequencing technology. We propose a novel technique to identify AIMs from genomewide SNP data using sparse principal component analysis (sparse PCA). The procedure uses penalized regression methods to identify those SNPs in a genomewide panel that significantly contribute to the principal components while encouraging SNPs that provide negligible loadings to vanish from the analysis. We found that sparse PCA leads to negligible loss of ancestry information compared to traditional PCA analysis of genomewide SNP data. We further demonstrate the value of sparse PCA for AIM selection using real data from the International HapMap Project and a genomewide study of Inflammatory Bowel Disease. We have implemented our approach in open-source R software for public use.

KEY WORDS: Ancestry-informative markers; Genome-wide association studies; Population stratification; Principal component analysis; Variable selection

Introduction

Results from genomewide association studies (GWAS) are vulnerable to spurious associations between single-nucleotide polymorphisms (SNPs) and phenotype that arise from confounding due to population stratification. Such confounding originates from the union of SNP allele-frequency heterogeneity to phenotype heterogeneity across latent subpopulations within a test sample. GWAS studies typically correct for such confounding by using the project's genomewide SNP data to infer underlying population structure and subsequently adjust for such structure in tests of SNP-phenotype association. While a variety of statistical methods exist for determining population structure from genetic data [Li and Yu, 2008; Lee et al., 2009; Li et al., 2010], the most common approach for inferring ancestry is to conduct a principal component analysis (PCA) of genomewide SNP data [Chen et al., 2003; Zhang et al., 2003; Price et al., 2006; Patterson et al., 2006]. Construction of the principal components is straightforward and computationally efficient and the top components with the largest eigenvalues reliably capture differences in genetic variation due to subpopulation status. Researchers can use the principal components to correct for confounding due to population stratification in a variety of ways, including treating the components as covariates in regression models [Price et al., 2006; Patterson et al., 2006], using the components to construct a stratification score [Epstein et al., 2007], or combining the components into a measure of ancestry for fine matching of samples [Luca et al., 2008].

PCA typically uses tens of thousands of independent SNPs from the GWAS panel to infer population structure within a sample. However, a small subset of SNPs that show marked differences in allele frequencies among different populations may be sufficient to provide information on ancestry. The identification of this small subset of ancestry-informative markers (AIMs) from a GWAS has substantial value for follow-up genetic studies of regions identified from the initial study. For example, a SNP-based validation study likely

will use a custom chip that only genotypes a limited number of SNPs. Consequently, the replication study will require the inclusion of a small set of AIMs (determined, perhaps, from the genotyped SNPs in the initial GWAS) to ensure the replication results are robust to confounding due to population stratification. Similarly, a follow-up study of rare variation that employs sequencing of targeted regions likely will genotype a small set of AIMs to ensure that association tests of rare variation [Li and Leal, 2008; Price et al., 2010] are unbiased.

Existing studies that have identified AIMs typically utilize samples with known population structure (like the International HapMap project) and select as AIMs those SNPs that show the most between-population variation in allele frequencies [Smith et al., 2004; Lao et al., 2006; Tian et al., 2006; Kosoy et al., 2008; Price et al., 2008; Halder et al., 2008]. This process can be quite laborious and further requires apriori knowledge of the populations in the test sample. This latter issue will be problematic if researchers have collected samples from a population whose ancestry is poorly catalogued in existing reference databases of genetic variation. To identify AIMs for a validation study of such samples, researchers will need to identify SNPs most correlated with ancestry in the initial GWAS. In this context, these existing methods for AIM selection are not applicable. Paschou and colleagues [Paschou et al., 2007; 2008; 2010; Drineas et al., 2010] previously proposed a method for selecting AIMs without the need for prior knowledge of ancestry, by calculating a score for each SNP using the sum of squared loadings of leading principal components inferred from the sample and selecting AIMs using the SNPs with the largest scores.

In this article, we propose a different statistical method for selecting AIMs that avoids the painstaking screening of individual SNPs for AIM selection and further does not require explicit knowledge of the underlying populations in a sample. We propose a process for AIM selection using sparse principal component analysis [Jolliffe, 2004; Zou et al., 2006; Shen and Huang, 2008; Witten et al., 2009] that considers each SNP's contribution (or loading) to the

principal components derived from genomewide marker data. Unlike standard PCA methods that yield a multitude of SNPs with non-zero contributions to the principal components, sparse PCA methods produce a limited number of influential SNPs by imposing a penalty function during optimization, which encourages SNPs with modest or negligible loadings to vanish. The process is easy to implement and comprehensive in that it can allow for all genomewide SNP data for AIM selection. We provide R code implementing our approach on our website (see Web Resources).

We organize the rest of the paper as follows. First, we describe traditional PCA and show how the process can be reformulated as an alternating regression problem. We then introduce sparse PCA by modifying the alternating-regression form of traditional PCA to incorporate a penalty term during optimization that encourages SNPs with negligible loadings to vanish. We describe different penalized methods for use in sparse PCA, including lasso [Tibshirani, 1996] and adaptive lasso [Zou, 2006]. We then illustrate our sparse PCA approach for selecting AIMs using genomewide SNP data from the International HapMap Project [The International HapMap Consortium, 2005] and a GWAS of Inflammatory Bowel Disease (IBD). Using the IBD dataset, we show our approach provides misclassification rates for predicting ancestral groups based on AIMs comparable to Paschou's approach [Paschou et al., 2007; 2008]. We further confirm its ability to identify influential markers using simulated data. Our approach is implemented in user-friendly R software for public use.

Materials and Methods

Reformulation of Standard PCA

We assume a sample of n subjects genotyped for a total of d SNPs across the genome. We let z_{ij} denote the number of copies of a reference allele that subject i $(i = 1, ..., n)$ possesses

at SNP j ($j = 1, ..., d$). Thus, z_{ij} will take values of 0, 1, or 2. We center the genotype entries for each SNP j by replacing z_{ij} with $x_{ij} = (z_{ij} - \bar{z}_j)/\sqrt{n}$, where $\bar{z}_j = \sum_{i=1}^{n} z_{ij}/n$ denotes the mean genotype value for SNP j.

We summarize the sample genotype information by a matrix X with n rows and d columns that has (i, j)th entry x_{ij}. Application of PCA to X produces both principal components, which denote the axes of genetic variation inferred from the genomewide SNP data, as well as scores that provide the coordinates of each subject along the principal components. We can perform PCA of the genotype matrix X in a variety of ways, such as using singular value decomposition (SVD) or the power method [Jolliffe, 2004]. For this article, we propose to use an alternating-regression algorithm for PCA since the framework facilitates subsequent efficient implementation of sparse PCA for AIM selection.

The alternating-regression algorithm for PCA begins by identifying the first principal component (PC1) and the scores of the subjects along the first principal component (PCscore1). We denote PC1 by $v = (v_1, \cdots, v_d)^T$, where v_j denotes SNP j's contribution to the first principal component, and denote PCscore1 by $a = (a_1, \cdots, a_n)^T$. To estimate v and a, we use an iterative process to minimize the sum of squares

$$\sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - a_i v_j)^2 \tag{1}$$

Assuming an initial value for a, we first estimate v_j for each j as the slope of a linear regression model with no intercept using x_{ij}'s ($i = 1, \cdots, n$) as the response and a_i's as the independent variable. This gives $v_j = \sum_{i=1}^{n} x_{ij} a_i / \sum_{i=1}^{n} a_i^2$ for $j = 1, \cdots, d$. Then, fixing v and normalizing the v vector to have unit length, we estimate each a_i as the slope from a linear regression model with no intercept using x_{ij}'s as the response and v_j's ($j = 1, \cdots, d$) as the independent variable. This gives $a_i = \sum_{j=1}^{d} x_{ij} v_j$ for $i = 1, \cdots, n$. We repeat this iterative procedure fixing a to estimate v and vice versa until convergence. Once we identify PC1 (v) and PCscore1 (a), we replace x_{ij} by $x_{ij} - a_i v_j$ and use the alternative-regression approach in

(1) to identify the second principal component and its scores. We can continue this process until we estimate the desired number of principal components. This alternating-regression algorithm gives the same PCs and PC scores as standard eigenvector based calculations (Shen and Huang, 2008)

Sparse Principal Component Analysis with Lasso

To perform sparse PCA, we modify the alternating-regression algorithm for standard PCA in (1) to allow for a penalty term on the principal component $v = (v_1, \cdots, v_d)^T$. The penalty term encourages SNPs with negligible influence on an axis of genetic variation to have zero loadings. In selecting the penalty term, we initially consider the lasso or L_1 penalty [Tibshirani, 1996]. We refer to sparse PCA using lasso as L-PCA throughout this paper.

Estimation of principal component $v = (v_1, \cdots, v_d)^T$ and corresponding score $a = (a_1, \cdots, a_n)^T$ using L-PCA proceeds by minimizing the following modified criterion from (1)

$$\sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ij} - a_i v_j)^2 + 2\lambda\sum_{j=1}^{d}|v_j|. \tag{2}$$

where λ is a positive penalty parameter that controls complexity of the derived principal component. The value of the parameter can vary between $\lambda = 0$ (which implies standard PCA that does not penalize the SNP loadings in v) and $\lambda = \infty$ (which forces each SNP to make zero contribution to the principal component). In section , we describe a data-driven procedure for selecting an appropriate value for λ.

Assuming λ is known, we minimize (2) in a similar iterative fashion as the minimization of (1). For known v, we estimate each a_i using the same regression model described in the previous section. Once we obtain a, we then estimate each v_j as a solution of lasso regression where x_{ij} is a response variable, a_i $(i = 1, \cdots, n)$ is a independent variable, and the penalty on the loading v_j is $2\lambda|v_j|$. Let $\text{sgn}(\cdot)$ denote the sign function and $\{\cdot\}_+$ denote a truncation function that returns its argument if it is nonnegative or 0 if it is negative. We can show

that lasso regression yields the following soft-threshold solution for v_j [Tibshirani, 1996]

$$v_j = \frac{1}{\sum_{i=1}^{n} a_i^2} \cdot \text{sgn}\left(\sum_{i=1}^{n} x_{ij} a_i\right) \left\{ \left| \sum_{i=1}^{n} x_{ij} a_i \right| - \lambda \right\}_+ . \tag{3}$$

If $\lambda = 0$, then the solution of v_j in (3) is the same as the estimate from the minimization (1) from traditional PCA. However, for $\lambda > 0$, equation (3) shows that the estimate of v_j is shrunk to 0 if the magnitude of $\sum_{i=1}^{n} x_{ij} a_i$ is smaller than the penalty parameter. Consequently, that particular SNP will not contribute to the loading of the principal component. We iteratively solve for a and v using ordinary and lasso regression, respectively, until convergence.

Sparse Principal Component Analysis with Adaptive Lasso

Using L-PCA, the estimated non-zero SNP loadings of v are shrunk by a constant penalty term λ. Rather than applying a constant shrinkage to non-zero SNP loadings, it might be advantageous to adaptively vary the shrinkage for each SNP such that SNPs with large loadings are penalized less than SNPs with modest loadings. Such a strategy could identify more accurate principal components and scores than those identified by L-PCA since the effects of SNPs with large loadings are not diluted in the analysis. Therefore, we investigate a modified sparse PCA procedure using a penalty function derived from adaptive lasso regression [Zou, 2006] that varies the penalty applied to each loading within the algorithm. We refer to sparse PCA using adaptive lasso as AL-PCA in this paper.

Following Zou [2006], we minimize the following objective function in AL-PCA

$$\sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - a_i v_j)^2 + 2\lambda \sum_{j=1}^{d} |v_j| / |\hat{v}_j|, \tag{4}$$

For this minimization, the penalty function depends on $\hat{v}_j$, which denotes the estimate of the loading for SNP j using traditional PCA assuming no penalty. We can perform minimization for AL-PCA using an algorithm similar to that used in the minimization for L-PCA in (2). Assuming known v, we estimate a using the same regression method described in the previous

two sections. Once we obtain a, we then estimate each v_j as a solution of adaptive lasso regression where x_{ij} is a response variable, a_i $(i = 1, \cdots, n)$ is a independent variable, and the penalty on the loading v_j is $2\lambda|v_j|/|\hat{v}_j|$. We can show that the adaptive lasso regression yields the following soft-threshold solution for v_j

$$v_j = \frac{1}{\sum_{i=1}^{n} a_i^2} \cdot \text{sgn}\left(\sum_{i=1}^{n} x_{ij} a_i\right) \left\{ \left| \sum_{i=1}^{n} x_{ij} a_i \right| - \frac{\lambda}{|\hat{v}_j|} \right\}_+ . \tag{5}$$

The solution reveals that the SNP loading v_j is now shrunk by a SNP-specific threshold value $\lambda/|\hat{v}_j|$. If SNP j has a large loading on the principal component then its associated penalty will be smaller than if the loading is more modest. Consequently, using AL-PCA, the loadings of SNPs with large impact on genetic variation should be similar to the loadings from standard PCA whereas SNPs with smaller impact on variation will have their loadings shrunk closer to 0. Adaptive lasso regression often yields sparser parameter estimates than traditional lasso and less biased estimates for large parameter values [Zou, 2006]. Therefore, in identifying AIMs that explain population variation, AL-PCA will likely discover a smaller set of markers than L-PCA with larger loadings.

Selection of the Penalty Parameter λ for L-PCA and AL-PCA

We propose a likelihood-based procedure to select the penalty λ for L-PCA and AL-PCA using the genotypes and principal-component scores from the sample. For a specific value of λ, we model the likelihood as $L = \prod_{i=1}^{n} P[x_i, a_i] = \prod_{i=1}^{n} P[x_i|a_i] P[a_i]$, where $x_i = (x_{i1} \cdots, x_{id})^T$ denote the vector of standardized genotypes for subject i. We assume $x_i|a_i \sim N(a_i v, \sigma^2 I_d)$, where v is a vector of the SNP loadings assuming a specific λ and I_d is a $d \times d$ identity matrix. Using the likelihood L, we calculate the Bayesian Information Criterion (BIC) of the observed data as $BIC(\lambda) = -2\log(\widehat{L}) + \log(\text{number of data points}) \times \{\text{number of parameters}\}$, where we evaluate $\log(\widehat{L})$ using the maximum likelihood estimates

of σ^2 and σ_a^2. Specifically, assuming $a_i \sim N(0, \sigma_a^2)$, we define the BIC as

$$BIC(\lambda) = nd \log(\widehat{\sigma^2}) + n \log(\widehat{\sigma_a^2}) + \log(nd) \times (n + |\mathcal{V}|) \tag{6}$$

where $\widehat{\sigma^2} = \sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - a_i v_j)^2/(nd)$, $\widehat{\sigma_a^2} = \sum_{i=1}^{n} a_i^2/n$, and $|\mathcal{V}|$ is the number of nonzero entries of v. We choose as λ the value that minimizes BIC. Since BIC is analytically intractable, we employ a grid search to the find the minimum value.

Even though BIC gives us an automatic and data-driven way to select the penalty parameter λ, we may desire to regulate the complexity to force a specific number of SNPs to have non-zero loadings. For example, if a study has a budget to genotype p AIMs, we may want to choose the value of λ that yields exactly p SNPs with non-zero loadings. To conduct this task, we note that p decreases as λ increases in value. Consequently, if we identify more than (less than) p SNPs with non-zero loadings using the BIC-based λ, then we can increase (decrease) the value of λ until we reach our target of p SNPs. This strategy is straightforward to implement in statistical software.

Evaluation of Additional Principal Components and Construction of AIM Set

Once we identify the first principal component v and scores a using sparse PCA, we next apply the algorithm to infer the second principal component. First, we remove the effects of the first principal component from the genotype data by replacing x_{ij} with $x_{ij} - \hat{a}_i \hat{v}_j$, where $\hat{a}$ and $\hat{v}$ are derived from standard PCA without the penalty term. We remove the effects of the first principal component using standard PCA rather than sparse PCA because we wish to identify AIMs that summarize variation along the original principal components of the data. Removing the effects of the first principal component derived from sparse PCA will alter the formation of the second principal component from that anticipated using standard PCA. This could lead to inaccurate selection of AIMs.

After removing the effects of the first principal component, we apply sparse PCA to the modified entries $x_{ij} - \hat{a}_i \hat{v}_j$ to obtain the second principal component and corresponding scores.

We can then derive additional principal components using the same sequential process. To choose the total number of principal components k to consider for AIM selection, we estimate the number of components with non-zero eigenvalues using procedures commonly used in GWAS studies to infer significant axes of ancestry. Such procedures can include using Tracy-Widom statistics [Patterson et al., 2006; Luca et al., 2008] or a "scree plot" of the ordered eigenvalues.

From the sparse PCA analysis, we will identify small sets of SNPs with non-zero loadings for each of the k principal components. Some SNPs will have non-zero loadings across multiple principal components, whereas others might contribute to only one component. The unique set of SNPs with non-zero loadings across the principal components can be considered the initial set of AIMs to be considered for follow-up studies. Within this AIM set, we can prioritize SNPs further for genotyping by ranking those SNPs with non-zero loadings across multiple principal components above those that have non-zero loadings in only one component.

Algorithm for Sparse PCA

Here we describe the whole algorithm.

(1) Construct the $n \times d$ matrix X of scaled and centered SNP genotypes.

(2) Set the initial value of a. We propose to set the initial value of a using Xv, where v is the first right-singular vector of X (equivalently the first principal component score from PCA of $X^T X$).

(3) Find principal component v using soft-thresholding (3) or (5) and then normalize it.

(4) Compute principal component score $a = Xv$.

(5) Repeat 3–4 until convergence. This step may be conducted multiple times on grids of λ for penalty parameter selection.

(6) $X \leftarrow X - \hat{a}\hat{v}^T$ for the next principal component. Here $\hat{a}$ and $\hat{v}$ are derived from standard PCA without penalty.

(7) Repeat 2–6 k times for obtaining k principal components.

Computing time of sparse PCA algorithm generally increases linearly with the number of SNPs and the sample size.

Results

Application of Sparse PCA to HapMap Data

We applied our sparse PCA procedure for AIM selection to genomewide SNP data from the International HapMap Project [The International HapMap Consortium, 2005]. The dataset consists of 90 subjects of European ancestry (Utah residents with ancestry from northern and western Europe; CEU) from 30 parent-offspring trios, 90 subjects of African ancestry (Yoruba in Ibadan, Nigeria; YRI) from 30 parent-offspring trios, and 90 unrelated subjects of Asian ancestry (45 Han Chinese in Beijing, China; CHB and 45 Japanese in Tokyo, Japan; JPT). For illustrating sparse PCA, we use unrelated subjects only and so exclude the CEU and YRI offspring from analysis. Consequently, our sample for sparse PCA analysis consists of 210 subjects (60 CEU, 60 YRI, 45 CHB, 45 JPT).

The International HapMap Project genotyped the sample for 3,976,554 autosomal SNPs across the genome. Prior to application of sparse PCA, we first applied quality-control procedures to exclude problematic SNPs from the analysis. We removed SNPs with a missing genotype rate greater than 5%, with a minor allele frequency less than 0.01, or with a Hardy Weinberg P-value less than 0.005. From the remaining 2,217,538 SNPs, we further pruned the set to identify a subset of markers that were in linkage equilibrium. Such pruning is required to avoid identifying principal components that are due to SNP-SNP correlation arising from linkage disequilibrium rather than the ancestry that we are interested in detecting. For the

pruning, we first extracted a set of 207,972 SNPs with a missing genotype rate less than 1% that were separated by at least 10Kb. Next, we applied the H-clust [Rinaldo et al., 2005] procedure to select a further subset of SNPs with squared correlation less than 0.04, resulting in our final set of 24,395 independent SNPs for AIM selection. We assumed genotype data in the HapMap dataset were missing at random and imputed a missing SNP genotype using a random number from a trinomial distribution assuming probabilities of the three possible genotypes equal to sample frequencies in HapMap.

Application of PCA to the 24,395 SNPs in the HapMap dataset revealed the top two principal components were sufficient for explaining the genetic variability within the sample due to ancestry. Therefore, we applied the L-PCA and AL-PCA procedures using the BIC criterion in (6) to the top two principal components to identify AIMs. Using a personal desktop computer with Intel® Core™ i5-650 3.20GHz Processor, these L-PCA and AL-PCA analyses required only 206.32 and 123.26 seconds of computation time, respectively. From the original set of 24,395 SNPs, L-PCA identified 10,594 SNPs with non-zero loadings for at least one of the two principal components (7,609 SNPs had non-zero loadings for the first principal component while 5,009 SNPs had non-zero loadings for the second component). AL-PCA yielded similar results to L-PCA with the former procedure identifying 10,432 SNPs with non-zero loadings along at least one of the two components (7,501 SNPs with non-zero loadings for the first component, 4,888 with non-zero loadings for the second component). We present scatterplots of the individual scores from traditional PCA, L-PCA, and AL-PCA in Figure 1. The results show that the scores from traditional PCA and the two sparse PCA methods are almost identical and can clearly separate the three distinct ethnic groups. These results suggest sparse PCA methods provide similar inference on population structure compared to traditional PCA even though the former methods use less than 50% of the SNPs that the latter method uses for constructing principal components.

[Figure 1 about here.]

We provide scatterplots of the SNP loadings of the two significant principal components derived from traditional PCA and the sparse PCA procedures in Figure 2. As shown in the top set of panels that compares the SNP loadings from L-PCA to traditional PCA, we observe that L-PCA encourages SNPs with negligible contributions in standard PCA to have zero loadings under the sparse procedure. In the bottom set of panels, we compare the SNP loadings from AL-PCA to traditional PCA and observe similar findings although we note that the loadings under AL-PCA bend away from zero more than the loadings from L-PCA. This is consistent with the fact that AL-PCA shrinks SNPs with larger loadings less than SNPs with smaller loadings whereas L-PCA shrinks every non-zero loading by an equal amount based on the penalty parameter.

[Figure 2 about here.]

We next use sparse PCA to identify a small set of 200 AIMs for each PC from the HapMap dataset. Rather than applying sparse PCA with a penalty λ derived using BIC criteria, we instead apply the technique varying the value of λ until we obtain 200 SNPs with non-zero loadings along each of the two significant dimensions. Using a personal desktop computer with Intel® Core™ i5-650 3.20GHz Processor, these L-PCA and AL-PCA analyses required only 119.87 and 77.74 seconds of computation time, respectively. We show the resulting principal-component scores derived from the 200 AIMs for each PC based on L-PCA and AL-PCA in Figure 3. The plot shows clear separation of the three subpopulations, which demonstrates that the set of AIMs selected as the top 200 AIMs for each of the first two PCs are able to account for genetic variation due to population subgroups. We also note that the AIMs selected by our sparse PCA procedures are not identical to those SNPs with the largest loadings using standard PCA. As an example, for the first principal component, the top 200 AIMs selected by L-PCA overlaps in only 49 of the SNPs with the top 200 loadings from

standard PCA. For AL-PCA, the number of overlapping SNPs changes to 98. For the second principal component, the top 200 AIMs selected by L-PCA and AL-PCA overlap with 117 and 116 of the 200 SNPs with the largest loadings from standard PCA, respectively.

[Figure 3 about here.]

We next repeated our sparse PCA analyses but included 152 reported AIMs distinguishing European, African, and East Asian populations [Smith et al., 2004] with the other 24,395 SNPs from HapMap. We applied sparse PCA but manually varied the penalty function to obtain PC-wisely 200 SNPs with non-zero loadings across the two principal components. L-PCA and AL-PCA identified 137 and 136, respectively, of the 152 AIMs listed by Smith et al. [2004]. Therefore, our efficient sparse PCA can successfully identify the majority of AIMs identified by previous laborious methods.

Application of Sparse PCA to IBD Data

We applied our sparse PCA procedure to determine AIMs in a genetic study of inflammatory bowel disease (IBD) data published by Price et al. [2008]. The IBD dataset consists of 912 European American controls from the New York Health Project and U. S. Inflammatory Bowel Disease Consortium that are genotyped for the 317,503 SNPs on the Illumina HumanHap300 panel. Subjects reported their ancestry by indicating one or more of the following: "Scandinavian", "Northern European", "Central European", "Eastern European", "Southern European", "East Mediterranean", or "Ashkenazi Jewish". Following Price et al. [2008], we simplified this classification into four exclusive categories: NW (North Western European), SE (South Eastern European), AJ (Ashkenazi Jewish) and NR (Not reported other than being "European"). Applying the same quality-control and pruning techniques as described in the previous section, we obtained a reduced set of 32,772 independent SNPs under consideration for AIM selection. Using EIGENSTRAT [Price et al., 2006], we applied traditional PCA to the SNP data and identified 17 subjects who were outliers along the

principal components. We removed these outliers from further analyses, such that our final sample consisted of 895 samples composed of 106 NWs, 8 SEs, 389 AJs and 392 NRs.

We randomly divided our sample into two data sets with equal proportions of each ethnic category present in each data set, using the first data set as the training set and the second data set as the validation set. Using the training set, we applied sparse PCA varying the penalty parameter λ to identify 150 SNPs with non-zero loadings along each of the first two principal components. In Figure 4, we plot the sparse PCA scores using L-PCA (top left panel) and AL-PCA (bottom left panel) for the individuals used in the training set. Similar to Price et al. [2008], we observe that the AJ group in the training set is clearly distinguished from the other ethnic groups, indicating that the first sparse principal component accounts for genetic difference between the Ashkenazi Jewish population and other European populations.

We next examined whether the set of AIMs identified as 150 AIMs on the first two PCs by L-PCA/AL-PCA in the training set accurately differentiated the populations in the validation dataset. To do this, we applied traditional PCA to these AIMs detected using L-PCA and AL-PCA to the validation dataset. In Figure 4, we show the scores for the first two principal components for L-PCA (top right panel) and AL-PCA (bottom right panel) calculated using the validation set. We observe similar findings to the training set with the first principal component clearly distinguishing subjects of Ashkenazi Jewish ancestry from the other European populations. These results suggest our proposed sparse PCA approach can identify small sets of AIMs that can distinguish different ancestral groups in a European sample.

[Figure 4 about here.]

As shown in Figure 4, only the first PC is useful for predicting ancestry, we next compared the predictive ability of the top 150 AIMs selected by L-PCA, AL-PCA and the approach of Paschou and colleagues [Paschou et al., 2007; 2008; 2010; Drineas et al., 2010] along the

first principal component for assigning ancestry. We divided the samples into training and validation datasets of equal size. We applied linear discriminant analysis (LDA) [Hastie et al., 2009] using the AIMS selected in the training set to classify the subjects into two populations: Ashkenazi Jewish and all other European groups. We then applied the estimated classification rule to the validation data to group subjects into the two populations and compared their predicted memberships with the observed memberships. As shown in Figure 4, the two groups are well separated by the first principal component. To obtain robust results, we applied the proposed sparse PCA methods to 50 different random splits and present the average of misclassification rates in Table 1, The results show that sparse PCA methods give misclassification rates comparable to Paschou's method, while standard PCA using all the pruned markers has a smaller misclassification error rate. Furthermore, we provide misclassification rates based on application of standard PCA to the original pruned set of 32,772 SNPs and note that the misclassification rates for this procedure is similar to those using sparse PCA even though the latter method uses only 0.5 percent of the available marker data.

[Table 1 about here.]

Simulations

We performed additional simulations to examine the ability of sparse PCA to correctly differentiate AIMs from random markers within a sample. We based our simulations on findings by Akey et al. [2002], who identified SNPs showing substantial allele-frequency differences in a mixed sample of African Americans, East Asians, and European Americans. Within the sample, the authors constructed the fixation index F_{ST} for each SNP with larger values of the index indicating markers that showed increased population differentation. We used this information to construct appropriate panels of AIMs in our simulated datasets.

We generated datasets comprised of either $n = 300$ or $n = 600$ subjects consisting of

equal proportions of African-American, East Asian, and European-American subjects. For each dataset, we generated a set of d SNPs with 100 consisting of AIMs and the remaining $d - 100$ being random markers that have no allele-frequency differences between different populations. We varied the value of d between 250 and 6000. We considered two different sets of 100 AIMs for our simulations; the first set consisted of the 100 SNPs from Akey et al. [2002] with large F_{ST} values varying between 0.24-0.54 (Simulation 1), while the second set had smaller F_{ST} values ranging between 0.07-0.10(Simulation 2). Using either set, we simulated a subject's genotype at an AIM as the sum of two draws from a Bernoulli distribution with success probability equal to the minor allele frequency of the SNP in the subject's population listed in Akey et al. [2002]. To generate a genotype at a random marker that shows no difference among populations, we first assumed a SNP with an allele frequency chosen from a uniform distribution with range 0.05 and 0.95. We then generated each subject's genotype at this SNP as the sum of two draws from a Bernoulli distribution with success probability equal to the allele frequency. For a dataset generated under a given design, we applied sparse PCA and identified those markers that were AIMs across the first 2 principal components (the first 2 components are sufficient for distinguishing 3 discrete populations). We recorded the number of SNPs that were correctly identified as AIMs (true positives) as well as the number that were incorrectly inferred as AIMs (false positives). We analyzed 1000 replicate datasets under each model.

Table 2 shows the average number of true AIMs and false AIMs identified under Simulation 1 where the 100 AIMs show substantial differences in allele frequency among the 3 populations. Across different sample sizes and different number of SNPs considered in the model, we found that sparse PCA was able to correctly identify all AIMs in every replicate. Meanwhile, the average number of random SNPs incorrectly identified as AIMs across replicates was generally small with no more than an average of 6 SNPs incorrectly assigned across the

different models. As expected, we observed that AL-PCA generally identified a sparser model than L-PCA.

[Table 2 about here.]

Table 3 reports the average number of true AIMs and false AIMs inferred in Simulation 2 where the 100 AIMs demonstrate less differentation across populations than the AIMs in Simulation 1. For $n = 600$, we observed that sparse PCA procedures were able to identify nearly all true AIMs while incorrectly inferring only a few false AIMs from the random markers. Interestingly, we observed that AL-PCA identified a smaller model than L-PCA when the number of SNPs d in the model was small but identified a larger model than L-PCA when d increased in size. While our expectation is that AL-PCA often finds a smaller model than L-PCA, we note that most results demonstrating this general finding are based on situations where the number of predictors d is smaller than the sample size n and the signal contained in d is strong. Less is known about the relative size of the AL-PCA model compared to the L-PCA model when d is larger than n and the signal in d is modest. We observed similar findings for $n = 300$ although the average number of true AIMs identified by both sparse PCA procedures was generally smaller; particularly when the number of SNPs d in the model increased in size.

[Table 3 about here.]

Discussion

Follow-up studies of candidate genomic regions using SNP or resequencing data require genotyping a small set of ancestry-informative markers (AIMs) to ensure significant findings are valid and not due to confounding due to population stratification. In this article, we propose a sparse PCA procedure for identifying AIMs from genomewide SNP data applicable to a sample of arbitrary size and unknown ancestry, which makes the procedure inherently more

flexible than existing methods for AIM selection. In particular, our approach is appropriate for selecting AIMs from a GWAS sample to genotype in a replication sample of similar population origin. Using both real and simulated data, we show that our approach can correctly identify AIMs within a sample and, in studies where populations are known, detect similar sets of AIMs to those identified using standard procedures. We have implemented our approach in open-source R code (see Web Resources) for public use.

We note that researchers implementing our sparse PCA procedure to select AIMs must first perform some preprocessing of the sample data before analysis. In particular, for a GWAS SNP panel, researchers will need to prune the initial set of SNPs to yield a smaller set of SNPs that are in approximately linkage equilibrium. This process is necessary to ensure that principal components derived from the SNP data are due to ancestry and not due to linkage disequilibrium among markers. Methods for SNP pruning are available in software packages like H-clust [Rinaldo et al., 2005] and PLINK Purcell et al. [2007]. It should be noted that the initial stage of pruning out correlated SNPs could yield many different sets of uncorrelated SNPs, and hence selection of ancestry markers is not unique. Furthermore, it is important to remove subjects who are outliers along the principal-component scores using appropriate criteria Price et al. [2006]; Luca et al. [2008] since such subjects can strongly influence the dimension-reduction procedure and thereby influence the choice of AIMs.

We focus in this paper on the use of spare PCA to identify AIMs and show that the method can identify a small set of markers that capture genetic ancestry as efficiently as genomewide SNP data can. We also can apply sparse PCA directly to genomewide marker data for the purpose of population-stratification adjustment in GWAS studies and whole-genome/whole-exome sequencing studies. Specifically, sparse PCA is a more stable algorithm when the number of dimensions of data (in this context, the number of markers) is greater than the sample size. Consequently, there may be value in using sparse PCA rather than traditional

PCA methods to correct for population stratification in genetic studies of complex traits. We will explore this potential application in a subsequent manuscript.

Web Resources

The URLs for data presented herein are as follows:

R code for AIM selection, http://www.hsph.harvard.edu/∼xlin/software.html

Acknowledgements

This work was supported by National Institutes of Health grants R37CA076404 and P01CA134294 for X.L. and S.L. and HG003618 for M.P.E and R.D.. And it was also supported for S.L. by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0011608).

References

Akey, J. M., Zhang, G., Zhang, K., Jin, L., and Shriver, M. D. (2002). Interrogating a high-density snp map for signatures of natural selection. *Genome Res*, 12:1805–1814.

Chen, H.-S., Zhu, X., Zhao, H., and Zhang, S. (2003). Qualitative semiparametric test to detect genetic association in case-control design under structured population. *Ann Hum Genet*, 67:250–264.

Drineas, P., Lewis, J., and Paschou, P. (2010). Inferring geographic coordinates of origin for europeans using small panels of ancestry informative markers. *PLoS One*, 5:e11892. doi:10.1371 / journal.pone.0011892.

Epstein, M. P., Allen, A. S., and Satten, G. A. (2007). A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet*, 80:921–930.

Halder, I., Shriver, M., Thomas, M., Fernandez, J. R., and Frudakis, T. (2008). A panel

of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Human Mutation*, 29:648–658.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Element of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition.

Jolliffe, I. T. (2004). *Principal component analysis*. Springer, second edition.

Kosoy, R., Nassir, R., Tian, C., White, P. A., Butler, L. M., Silva, G., Kittles, R., Alarcon-Riquelme, M. E., Gregersen, P. K., Belmont, J. W., De La Vega, F. M., and Seldin, M. F. (2008). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in america. *Hum Mutat*, 30:69–78.

Lao, O., van Duijn, K., Kersbergen, P., de Knijff, P., and Kayser, M. (2006). Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *Am J Hum Genet*, 78:680–690.

Lee, A. B., Luca, D., Klei, L., Devlin, B., and Roeder, K. (2009). Discovering genetic ancestry using spectral graph theory. *Genet Epidemiol*, 34:51–59.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases:application to analysis of sequence data. *Am J Hum Genet*, 83:311–321.

Li, M., Reilly, M. P., Rader, D. J., and Wang, L. (2010). Correcting population stratification in genetic association studies using a phylogenetic approach. *Bioinformatics*, 26:798–806.

Li, Q. and Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet Epidemiol*, 32:215–226.

Luca, D., Ringsquist, S., Klei, L., Lee, A. B., Gieger, C., Wichmann, H. E., Schreiber, S., Krawczak, M., Lu, Y., Styche, A., Devlin, B., Roeder, K., and Trucco, M. (2008). On the use of genetic control samples for genome-wide association studies: genetic matching

highlights causal variants. *Am J Hum Genet*, 82:453–463.

Paschou, P., Drineas, P., Lewis, J., Nievergelt, C. M., Nickerson, D. A., Smith, J. D., Ridker, P. M., Chasman, D. I., Krauss, R. M., and Ziv, E. (2008). Tracing sub-structure in the european american population with pca-informative markers. *PLoS Genet*, 4:e1000114, doi:10.1371 / journal.pgen.1000114.

Paschou, P., Lewis, J., Javed, A., and Drineas, P. (2010). Ancestry informative markers for fine-scale individual assignment to worldwide populations. *J Med Genet*, 47:835–847.

Paschou, P., Ziv, E., Burchard, E. G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. W., and Drineas, P. (2007). Pca-correlated snps for structure identification in world-wide human populations. *PLoS Genet*, page e160. doi:10.1371 / journal.pgen.0030160.

Patterson, N. J., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet*, 2:e190, doi:10.1371 / journal.pgen.0020190.

Price, A. L., Butler, J., Patterson, N., Capelli, C., Pascali, V. L., Scarnicci, F., Ruiz-Linares, A., Groop, L., Saetta, A. A., Korkolopoulou, P., Seligsohn, U., Waliszewska, A., Schirmer, C., Ardlie, K., Ramos, A., Nemesh, J., Arbeitman, L., Goldstein, D. B., Reich, D., and Hirschhorn, J. N. (2008). Discerning the ancestry of european americans in genetic association studies. *PLoS Genet*, 4(1):e236. doi:10.1371/journal.pgen.0030236.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L., and Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet*, 86:832–838.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38:904–909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). Plink: a toolset

for whole-genome association and population-based linkage analysis. *Am J Hum Genet*, 81:559–575.

Rinaldo, A., Bacanu, S., Devlin, B., Sonpar, V., Wasserman, L., and Roeder, K. (2005). Characterization of multilocus linkage disequilibrium. *Genet Epidemiol*, 28:193–206.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J Multi Anal*, 99:1015–1034.

Smith, M. W., Patterson, N., Lautenberger, J. A., Truelove, A. L., McDonald, G. J., Waliszewska, A., Kessing, B. D., Malasky, M. J., Scafe, C., Le, E., De Jager, P. L., Mignault, A. A., Yi, Z., de Thé, G., Essex, M., Sankalé, J., Moore, J. H., Poku, K., Phair, J. P., Goedert, J. J., Vlahov, D., Williams, S. M., Tishkoff, S. A., Winkler, C. A., De La Vega, F. M., Woodage, T., Sninsky, J. J., Hafler, D. A., Altshuler, D., Gilbert, D. A., O'Brien, S. J., and Reich, D. (2004). A high-density admixture map for disease gene discovery in african american. *Am J Hum Genet*, 74:1001–1013.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 432:1299–1320.

Tian, C., Hinds, D. A., Shigeta, R., Kittles, R., Ballinger, D. G., and Seldin, M. F. (2006). A genomewide single-nucleotide-polymorphism panel with high ancestry information for african american admixture mapping. *Am J Hum Genet*, 79:640–649.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J Royal Statist Soc B*, 58:267–288.

Witten, D. J., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with application to sparse principal components and canonical correlation analysis. *Biostat*, 10:515–534.

Zhang, S. L., Zhu, X. F., and Zao, H. Y. (2003). On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals.

Genet Epidemiol, 24:44–56.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J Am Statist Assoc*, 101:1418–1429.

Zou, H., Hastie, T. J., and Tibshirani, R. J. (2006). Sparse principal component analysis. *J Comput Graph Statist*, 15:265–286.

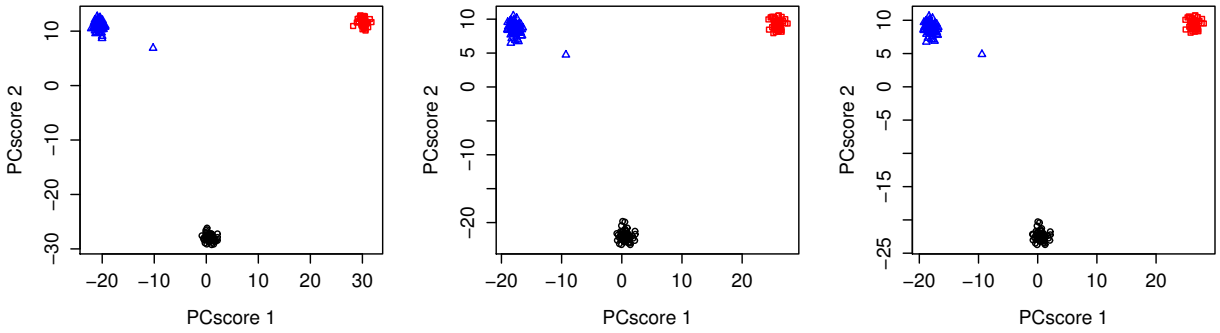

Figure 1. Scatterplots of the first two principal component scores. Scatterplots of the first two principal component scores from traditional PCA (left), L-PCA (middle) and AL-PCA (right). Black circles are European, red rectangles are African and blue triangles are Asian.

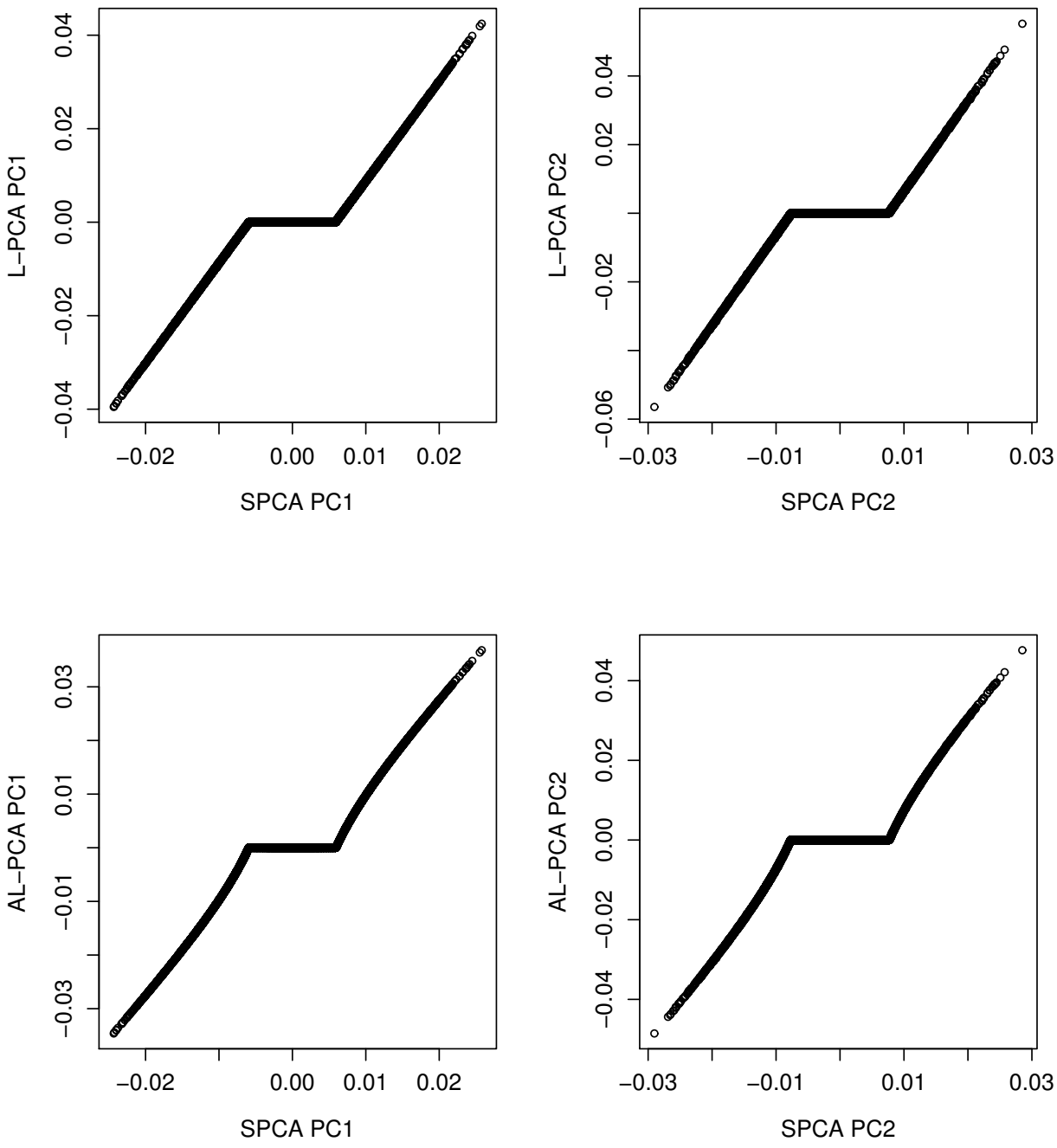

Figure 2. Scatterpots of sparse PC versus standard PC. Scatterplots of sparse principal component and traditional principal component are presented. Top left panel is scatterplot for the first principal component and top right is for the second principal component using L-PCA. Bottom panels are from AL-PCA.

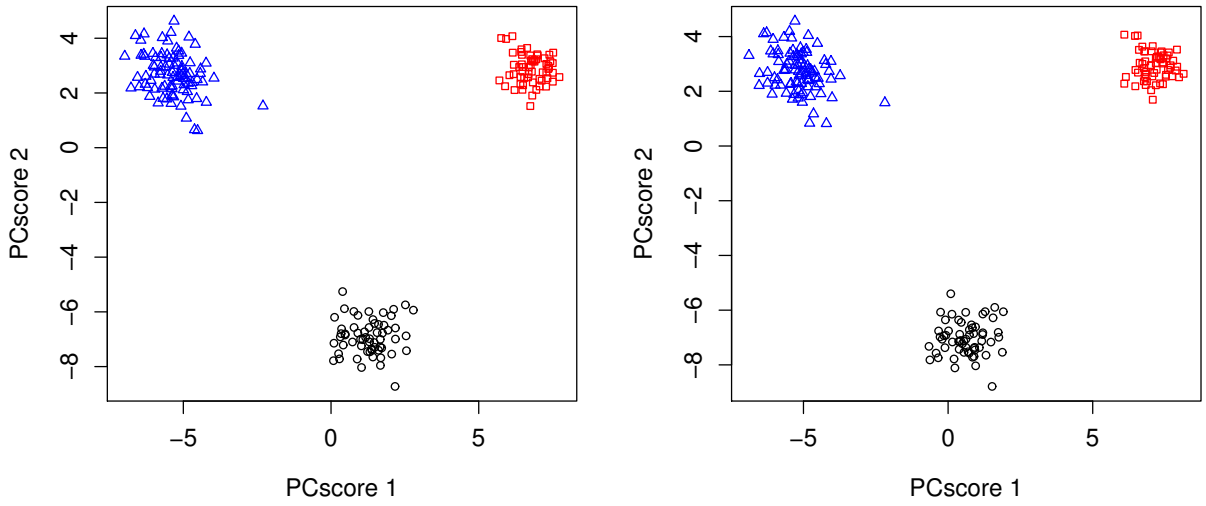

Figure 3. Scatterplots of the first two principal component scores with 200 nonzero SNP loadings. Scatterplots of the first two principal component scores are drawn when we restrict each principal component to have only 200 nonzero SNP loadings. Left panel is for L-PCA and right is for AL-PCA. Black circles are Caucasian, red rectangles are African and blue triangles are Asian.

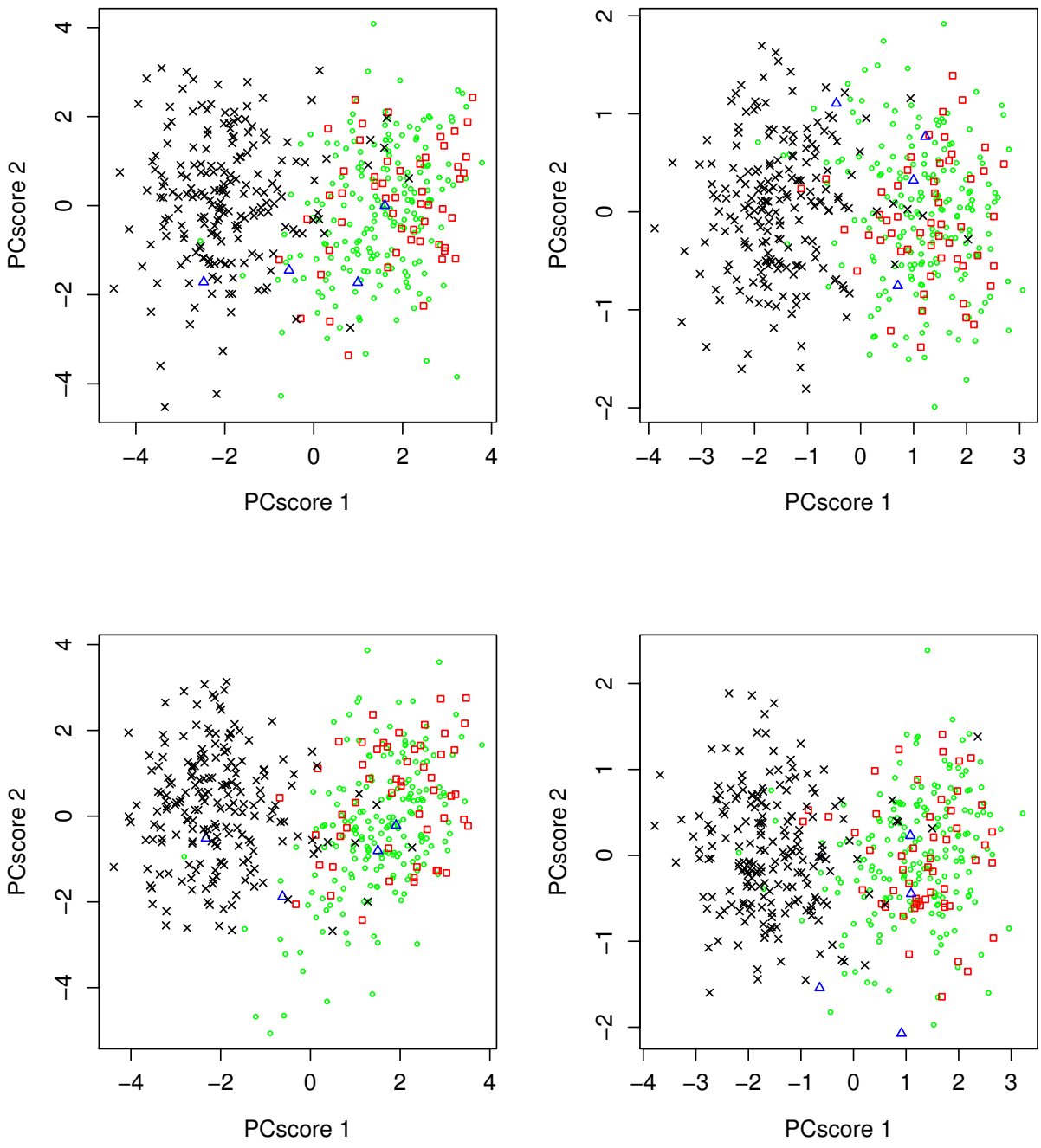

Figure 4. Scatterplots of principal component scores using sparse PCs from training set. The top-left panel is PC plots for training set and the top-right panel for validation set using L-PCA. Bottom panels are using AL-PCA. Black cross is for AJ, red rectangle for NW, blue triangle for SE and green circle for NR.

Table 1
Results of the analysis of the IBD data *Average misclassification rates (in percentage) over 50 different splits using L-PCA, AL-PCA, standard PCA (SPCA), and Paschou's method are provided for training and validation datasets.*

	L-PCA	AL-PCA	Paschou	SPCA
Training set	5.7092	5.2841	5.0515	4.9217
Validation set	5.0313	4.9777	4.9821	4.2545

Table 2
Simulation 1 *Average (SD) number of true AIMs and false AIMs identified by sparse PCA*

n	d	Method	True AIMs (SD)	False AIMs (SD)
300	250	Lasso	100.00 (0.00)	5.08 (2.96)
		Adaptive Lasso	100.00 (0.00)	2.85 (2.22)
	500	Lasso	100.00 (0.00)	5.18 (3.22)
		Adaptive Lasso	100.00 (0.00)	3.26 (2.41)
	750	Lasso	100.00 (0.00)	4.80 (3.00)
		Adaptive Lasso	100.00 (0.00)	3.26 (2.42)
	1500	Lasso	100.00 (0.00)	4.52 (2.96)
		Adaptive Lasso	100.00 (0.00)	3.44 (2.44)
	3000	Lasso	100.00 (0.00)	4.43 (2.83)
		Adaptive Lasso	100.00 (0.00)	3.60 (2.56)
	6000	Lasso	100.00 (0.00)	4.25 (2.71)
		Adaptive Lasso	100.00 (0.00)	3.82 (2.65)
600	250	Lasso	100.00 (0.00)	5.94 (3.20)
		Adaptive Lasso	100.00 (0.00)	3.01 (2.14)
	500	Lasso	100.00 (0.00)	6.28 (3.27)
		Adaptive Lasso	100.00 (0.00)	3.45 (2.45)
	750	Lasso	100.00 (0.00)	6.42 (3.23)
		Adaptive Lasso	100.00 (0.00)	3.49 (2.52)
	1500	Lasso	100.00 (0.00)	6.07 (3.34)
		Adaptive Lasso	100.00 (0.00)	3.45 (2.53)
	3000	Lasso	100.00 (0.00)	5.81 (3.13)
		Adaptive Lasso	100.00 (0.00)	3.58 (2.59)
	6000	Lasso	100.00 (0.00)	5.62 (3.17)
		Adaptive Lasso	100.00 (0.00)	3.58 (2.63)

Table 3
Simulation 2 *Average (SD) number of true AIMs and false AIMs identified by sparse PCA*

n	d	Method	True AIMs (SD)	False AIMs (SD)
300	250	Lasso	99.46 (0.92)	3.16 (2.22)
		Adaptive Lasso	99.44 (0.90)	2.86 (2.13)
	500	Lasso	98.71 (1.67)	4.15 (2.83)
		Adaptive Lasso	98.61 (1.64)	3.89 (2.76)
	750	Lasso	98.09 (1.99)	4.51 (2.96)
		Adaptive Lasso	97.98 (1.99)	4.34 (2.94)
	1500	Lasso	96.40 (3.19)	5.05 (3.33)
		Adaptive Lasso	96.43 (2.92)	5.25 (3.26)
	3000	Lasso	93.96 (4.52)	5.40 (3.51)
		Adaptive Lasso	92.95 (5.02)	6.50 (4.02)
	6000	Lasso	86.53 (11.61)	6.92 (4.31)
		Adaptive Lasso	76.87 (12.30)	7.56 (4.40)
600	250	Lasso	100.00 (0.00)	3.87 (2.51)
		Adaptive Lasso	100.00 (0.00)	3.07 (2.08)
	500	Lasso	100.00 (0.00)	4.62 (2.91)
		Adaptive Lasso	100.00 (0.00)	4.05 (2.66)
	750	Lasso	100.00 (0.00)	4.90 (3.12)
		Adaptive Lasso	100.00 (0.00)	4.61 (2.78)
	1500	Lasso	100.00 (0.04)	5.09 (3.24)
		Adaptive Lasso	100.00 (0.03)	5.00 (3.21)
	3000	Lasso	99.99 (0.11)	5.08 (3.17)
		Adaptive Lasso	99.98 (0.14)	5.61 (3.50)
	6000	Lasso	99.97 (0.19)	5.24 (3.22)
		Adaptive Lasso	99.92 (0.30)	6.40 (3.69)