Cluster analysis and PCA for modeling population structure

Ph.D Thesis proposal

Diana Luca
Department of Statistics, Carnegie Mellon University

July 27, 2007

Abstract

Case-control studies for association are widely used for finding genetic variants causally associated with phenotypes. Unfortunately, population structure can induce false positives. For instance, if cases and controls have different genetic backgrounds, differences in frequencies of distinct forms of variants might be due to differences in ancestral population of origin. Traditional approaches to control for the effects of population stratification include eigen-analysis, cluster analysis and matching based on genetic markers, are employed to improve the modeling of structure. Our approach goes further in that we show how to systematically obtain optimal matching and how to determine outlying subjects that cannot be successfully matched to others in the available registry. Simulations and an application to real data show improved results applying the new method.

1 Introduction

Among the resources being amassed for whole genome association (WGA) are "control databases", more precisely databases of samples from the population that are not necessarily screened to be disease free. Often these samples will have been characterized genetically by a large-scale genotyping array, possibly as a result of a WGA study. An open question is how to use these samples in a cost- and power-effective manner in various settings, especially when a subsample of the control database is desired for a new genetic study. Here we address one critical component of study design, how to select samples matched for ancestry. Suppose, for example, a smaller set of cases and a larger set of controls have been characterized for

a set of genetic markers, but now new genotyping is to be performed as part of a follow-up study.

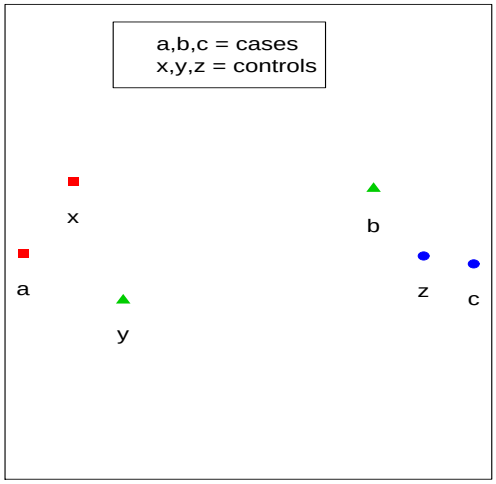
As platforms for genetic association analysis become standardized, numerous sources of pre-genotyped control subjects have become available. Typically many more controls are available than cases. For example, the data set analyzed in this paper has genetic information for 462 cases, which are Americans of European descent with type 1 diabetes, and 2159 controls of which 1429 are from north and north-east Germany and 730 are from southern Germany.

Matching based on non-genetic variables (Lee 2006 [11]) as well as SNP panels (Hinds et al. 2004 [8]) has been used successfully in genetic association studies previously. Our approach goes further in that we show how to systematically obtain optimal matching using a panel of genetic markers and how to determine outlying subjects that cannot be successfully matched to others in the available registry. The matching is based on genetic similarities derived from principal component analysis (PCA) and multidimensional scaling with an approach similar to that taken by Price et al (2006) [14].

The best known form of matching is *matched pairs*; however, assuming the criterion for matching are sufficient to remove the effects of unmeasured confounding, an alternative to matched pairs known as *full matching* is optimal (Rosenbaum 1995 [16]). Figure 1 illustrates 3 cases (a,b,c) and 3 controls (x,y,z) which appear to fall into two distinct clusters. Matched pairs would create 3 strata, (a,x), (c,z), and (b,y). Clearly the pair (b,y) does not define a homogeneous strata. Alternatively, full matching minimizes the total distance between subjects within strata with the constraint being that each strata includes a single case and one or more controls, or a single control and one or more cases. For this example the full matching solution forms two strata by clustering (a,x,y) and (b,c,z).

PCA is highly sensitive to outlying observations. A few points lying far from the majority of the data can determine several principal axes of the representation. Indeed, outliers can obscure the discovery of axes that potentially separate the data into distinct types. For this reason, Price et al. remove subjects that have highly unusual measures on any of the major eigenvectors. Likewise, with matching criterion it is necessary to determine which pairs or strata span an unusual distance. For example, in Figure 1, the matched-pair (b,y) should be removed from the analysis, because this pair is inconsistent with the model of homogeneity within strata. In general, if the controls are more numerous than the cases and they span a

Figure 1: Full matching versus pair matching



larger range of ethnicities than the cases, it should be possible to find one or more controls similar to each case. If the situation is reversed, then some cases will have to be removed from the analysis to account for the effects of structure. In this work we formalize the notion of outlying subjects and propose a method for removing them from the analysis so as to discover the key axes that describe the population structure.

2 Literature review

Case-control studies rely on the unrealistic assumption of population homogeneity. In the face of population heterogeneity, spurious associations can arise. As a response to this problem two approaches to controlling structure have arisen: genomic control, which corrects for minor stratification using an estimate of the inflation factor[5] ; and structured association, which clusters subjects into more homogeneous subsets prior to analysis [15] . Both of these approaches have shortcomings when applied to large samples with huge panels of SNPs. The former exhibits diminishing power because the effect of stratification increases with sample size. The latter does not scale well and becomes computationally intractable. Hence, a third traditional approach for population classification, based on eigenanalysis, has recently been updated for association testing [13]. This method combines principal compo-

nents with modern statistics (Tracy-Widom theory) to test for population structure. The application of PCA to genetic data has become a standard tool. Cavalli-Sforza [4] show that principal components displayed in two dimensions reflect the geographical distribution of populations. For populations that are geographically close, they found that genetic and geographic distances are often highly correlated. Zhang et al. (2003) [19] propose a semi-parametric test for association (SPTA) to control for population stratification through a set of genomic markers by first deriving a genetic background variable and then modeling the relationship between trait values, genotypic scores at the candidate marker, and genetic background variables through a semiparametric model. The genetic background variable is defined for each sampled individual using PCA on a set of independent markers.

The most challenging problem with population stratification occurs when some candidate SNPs are highly differentiated, but the majority of SNPs have similar allele frequencies across populations. This situation arises when a SNP is under strong population specific selection. The Campbell et al. height/lactase data [3] provides an exceptional challenge in that the data include only a modest number of loci for calculating the PC. These include 111 missense and noncoding SNPs and 67 ancestry informative marker SNPs. The lactase variant, LCT, is known to exhibit extreme differentiation across the European continent. Moreover, the allele frequency gradient matches the heighth gradient in Europe, maximizing the opportunity for confounding. Ignoring the inherent structure in these data, one obtains a significant association between LCT and height (p-value = 0.0037). The subtle structure in these data proved too challenging for GC, Structure or eigenstrat to correct [14]. Not surprisingly, the subjects are also difficult to match using PCA. It appears that the 178 available markers cannot successfully differentiate the subjects along the European cline which is necessary for reliable correction of stratification. Presumably with a larger, more informative sample of SNPs, matching would remove the spurious association between height and LCT.

Lee [11], studies matching based on stratum-delineating variables such as race, ethnicity, nationality, ancestry and birthplace, followed by the genomic controlling stage. The author shows that using crude matching, some power is lost but the type I error is correct.

Hinds et al.[8], propose genetic ancestry matching prior to DNA pooling. Data is analyzed using *structure* [15], a model-based method for identifying subpopulations. Results indicate that relatively simple matching can control for population stratification, even for a phenotype with a very large ancestry effect.

Another way to control for population substructure is based on propensity scores [16]. For genetic association studies this quantity is obtained by modeling the odds of disease given a panel of genetic markers. Cases and controls can be clustered into a handful of strata based on having similar scores. The data can then be tested for association, conditional on these strata [6]. This uses both case/control status and the panel of genetic markers to stratify the subjects. In contrast, PCA uses only the structure apparent in the genetic SNP panel to stratify subjects. A propensity score approach achieved partial success with Campbell et al. data. The scores were used to define five strata. Next the data were analyzed using the Mantel-Haenzel test and the stratified logistic test with resulting p-values of 0.039 and 0.44, respectively. Although these tests usually perform similarly, the former failed to correct for the spurious association even with the benefit of the propensity strata. The difference is likely due to the notable lack of balance in the strata. The first strata includes 4 tall and 78 short individuals, while the last strata includes 71 tall and 3 short individuals. It is not difficult to imagine a scenario in which the extreme strata contain no cases (or no controls). In practice this method fails to scale to large panels of SNPs. Thus we focus on investigating the limits of the matched-strata and eigenanalysis approaches.

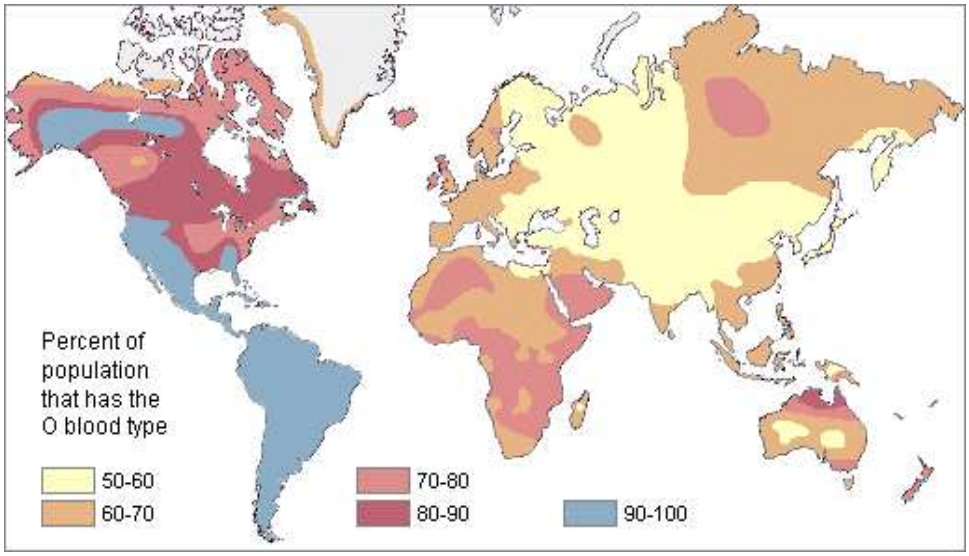
3 Methods

3.1 Background

People have pairs of alleles at each genetic marker, one inherited from each parent. In this project the genetic markers consist of Single Nucleotide Polymorphisms (SNPs), which have two forms. An individual's alleles can differ or be identical. They produce variations in inherited traits such as blood type and eye color. For example, the O blood type is very common around the world (Figure 2). About 63% of humans share it. Type O is particularly high in frequency among the indigenous populations of Central and South America, where it approaches 100%. It also is relatively high among Australian Aborigines and in Western Europe (especially in populations with Celtic ancestors). The lowest frequency of O is found in Eastern Europe and Central Asia, where B is common.

The minor allele is the form that is less frequently observed in a population. Human populations vary, so an allele that is common in one geographical or ethnic group may be much rarer in another. For instance the lactase variant is present in about 80% of

Figure 2: Distribution of the O blood type in native populations around the world. Both clinal and discontinuous distributions exist, suggesting a complicated evolutionary history for humanity (from Wikipedia).



northwestern european, but only in 20% of southeastern europeans. This variation is called a gradient or a cline.

Variables coding for SNP genotypes are defined in Table 1. The additive model, used in this study, counts the number of minor alleles within a pair.

Genotype	Dominant	Additive	Recessive
AA	0	0	0
Aa	1	1	0
aa	1	2	1

Table 1: Coding Genotypes

3.2 Model for Population Stratification

The mean of allele frequencies from a set of populations is assumed to be the allele frequency of an ancestral population. Individual populations have each diverged from the ancestral population over time. Suppose within a subpopulation C people are genetically i.i.d., and allele a is drawn with probability p_c in this subpopulation. If X is counting allele a, then

$$X \sim Binomial(2, p_c).$$

Let P be the random variable which varies across subpopulations, with p the realized value in subpopulation C. The distribution of X can be described using the following simple hierarchical model:

$$X|p \sim Binomial(2, p)$$

$$p \sim Beta(\alpha_1, \alpha_2)$$

$$\alpha = \alpha_1 + \alpha_2 = \frac{1}{F_{st}} - 1.$$

F_{st} defined above is called Fixation index and is a measure of population differentiation. Building on model 1, we can generalize the distribution of X for K populations. Assume that we have the minor allele frequencies of an ancestral population *p.loci* (in our simulations *p.loci* is uniform between .05 and .5) at L loci. From this ancestral population K subpopulations have been formed. Let N be the vector containing the population size $(N_1, N_2, \ldots, N_K)$. Knowing F_{st}, for each marker l we can define

$$\alpha_{1,l} = p.loci_l \times (\frac{1}{F_{st}} - 1)$$

$$\alpha_{2,l} = (1 - p.loci_l) \times (\frac{1}{F_{st}} - 1).$$

For each population k, $k = 1, \ldots, K$, define

$$p_{kl} = Beta(1, \alpha_{1,l}, \alpha_{2,l})$$

Then

$$X_{n_k,l} = Binom(2, p_{k,l}), n_k = 1, \ldots, N_k. \quad (1)$$

When used in simulation studies this is often called the Balding-Nichols model (1995).

Figure 3 ilustrates how the first two principal components show the population structure. In the first plot (Figure 3 top) there are 3 populations, each with 100 subjects, $N = (100, 100, 100)$, $L = 1000$ and $F_{st} = 0.03$.

To simulate a cline (or a gradient), it is enough to order p_{kl}, so that $p_{1l} \leq \ldots \leq p_{Kl}$ for each l. The first principal component in Figure 3 (botom) shows the gradient structure of a population, formed by 9 subpopulations, each with 30 subjects, L and F_{st} as above.

3.3 Multidimensional Scaling

Multidimensional scaling (MDS) constructs a configuration of n points in Euclidean space using information about the distances between the n objects. We apply this technique to reduce the dimensionality of the data so that the Euclidean distances between subjects in the reduced space is as close as possible to Euclidean distances in the original space.

Let X be a $n \times p$ matrix, derived from n subjects and p markers, with each column having mean 0 and standard deviation 1. A singular value decomposition gives

$$X = U\Gamma V^T,$$

where Γ is a diagonal matrix with singular values $\gamma_1, \gamma_2, \ldots, \gamma_p$ as diagonal entries. Note that

$$S = \frac{1}{n}X^T X = \frac{1}{n}V\Gamma^2 V^T = \frac{1}{n}V\Lambda V^T,$$

is the sample correlation matrix of the markers. Λ is also a diagonal matrix with singular values $\lambda_1, \lambda_2, \ldots, \lambda_p$ as diagonal entries, where $\lambda_i = \gamma_i^2$, $i = 1, \ldots, p$.
The columns of V represent the principal components of markers. The eigenvalues of $X^T X$ are proportional to the variances of the principal components. The matrix $U\Gamma$ then containes the principal components scores, which are the coordinates of the subjects in the space of principal components.

Usually in genetic data the number of subjects is less than the number of tested markers ($n < p$), so we can assume that the rank of S is n-1 (we lose a dimension by centering the columns). Suppose that the first $n-1$ eigenvalues are non-zero and distinct. Then $\lambda_1, \lambda_2 \ldots, \lambda_{n-1}$ are also the non-zero eigenvalues of $K = XX^t = U\Lambda U^T$. Note that K represents the centered inner product matrix, $K_{ij} = x_i^t x_j$.

Consider a l-dimensional representation $\Phi_l(i) = \phi_1(i) \ldots \phi_l(i)$ of subject i, where $l << min(n, p)$. Let D be the matrix of Euclidean distances between subjects $d_{ij}^2 = ||x_i - x_j||^2$. Define $\hat{d}_{ij}^2 = ||\Phi_l(i) - \Phi_l(j)||^2$ as the entries of the matrix of Euclidean distances between subjects in the l-dimensional space. Measure the discrepancy between D and $\hat{D}$ according

to $\delta = \sum_{i,j}(d_{ij}^2 - \hat{d}_{ij}^2)$. Then δ is minimized over all configurations in $\mathbf{R}^l$ when X is projected onto its principal coordinates in l dimensions, i.e. when

$$\phi_l(i) = [\gamma_1 u_1(i), \ldots, \gamma_l u_l(i)]^T.$$

(Mardia et al, Theorem 14.4.1).

3.4 Hypothesis Test for Population Structure

Patterson et al.(2006) propose a formal significance test for population structure. The method is based on a theoretical result by Johnstone (2006) on the eigenvalue distribution of a null sample covariance matrix. We can test the null hypothesis of identity covariance matrix, which is equivalent to population homogeneity, against an alternative hypothesis, that the covariance matrix has some general value, equivalent to population heterogeneity.

The sample covariance matrix S follows a Wishart distribution. Assume that $min\{n, p\} \to \infty$, and that $\hat{l}_1$ is the largest eigenvalue of S. After appropriate centering and scaling, the largest eigenvalue follows a distribution discovered by Tracy and Widom, called the TW law:

$$P\{n\hat{l}_1 \leq \mu_{np} + \sigma_{np} | H_0\} \to TW(s).$$

The centering and scaling parameters depend on both n and p,

$$\mu_{np} = (\sqrt{p-1} + \sqrt{n})^2,$$

$$\sigma_{np} = (\sqrt{p-1} + \sqrt{n})(\frac{1}{\sqrt{p-1}} + \frac{1}{\sqrt{n}})^{1/3}.$$

The test for population structure will be applied iteratively. If we find the first l eigenvalues $\lambda_1, \ldots, \lambda_l$ of S to be significant, we test $\lambda_k, \ldots, \lambda_{n-1}$ as though S were an $(n-l-1) \times (n-l-1)$ Wishart matrix. If an eigenvalue is not significant, the smaller eigenvalues will not be significant either.

3.5 Distribution of distances for null data

Let n and l be fixed. Project X onto its principal coordinates in the first l dimensions. To measure the distance between individuals in the l-dimensional space, we use the Euclidean distance

$$g(i, i') = \left\{ \sum_k \lambda_k (u_{ik} - u_{i'k})^2 \right\}^{1/2}$$

To create a distance that is stable for any large value of p we re-norm the λ's to have the T-W distribution. So

$$g^*(i,i') = \left\{ \sum_k \left(\left(\frac{\lambda_k - \mu_{np}}{\sigma_{np}} \right) - a \right) (u_{ik} - u_{i'k})^2 \right\}^{1/2},$$

where a is chosen so that none of the normed eigenvalues is negative. Now we have the distribution of distances of null data. These values should be independent of p, but will depend on n the number of subjects and l, the number of chosen dimensions.

3.6 Full matching and pair matching

These two matching methods construct matched sets or strata when there are many observed covariates (in our case markers) x. We would like to compare case and control groups with similar distribution of x, even if matched individuals have different values of x. The form of an optimal stratification is always the same and it is called full matching (Rosenbaum 1995). It is a matched sample in which each matched set contains either one case and one or more controls or one control subject and one or more cases. This method minimizes the sum of distances between cases and the matched controls. Pair matching is not optimal. When the treated and control groups have different distributions of x, there are regions of x values with many cases and few controls and other regions with many controls and few cases. Forcing every case to have the same number of controls leads to suboptimal matches.

3.7 Clustering and rescaling

Cases are paired to controls to minimize the distance between individuals within strata using either the full matching or pair matching criterion. Individuals, inevitably, will not match perfectly. We determine a rule for stopping the pairing process when the distance between case and control is greater than expected, assuming the case and control are from a homogeneous population.

We need to adjust the data so that they are scaled for the null hypothesis and yet the axes reflect variation in the full data set. Applying the test for population structure, we first determine l the number of dimensions to be used. Next we use Wards algorithm to form hierarchical groups of mutually exclusive subsets based on the first l principal components. Cluster membership is assessed by calculating the total sum of squared deviations from the mean of a cluster. The criterion for fusion is that it should produce the smallest possible

increase in the error sum of squares. We need a stopping rule for choosing the number of clusters (k). Start with $k = 2$ and apply the test for population structure on each of the clusters. Homogeneous clusters are kept unchanged and Wards algorithm is applied only on the heterogeneous clusters. Repeat this process, increasing k until all the clusters are homogeneous. Next, we calculate the scaled distances as follows:

Let $S_k \subset \{1, 2, \ldots, n\}$ be the indices of subjects in the k'th subset. Let r_k be the *number* of subjects in the k'th cluster. For scaling the k'th subset we work with subject $i \in S_k$ using eigenvector values $(u_{i1}, \ldots, u_{il})$. For $j = 1, \ldots, l$, rescale the u_{ij}'s. By construction, $\sum_i u_{ij}^2 = 1$ and $\bar{u}_j = 0$. Let

$$\bar{u}_{jk} = \sum_{i \in S_k} u_{ij}/r_k.$$

A traditional sums of squares decomposition leads to

$$1 = \sum_i u_{ij}^2 = \sum_k \sum_{i \in S_k} (u_{ij} - \bar{u}_{jk})^2 + \sum_k r_k \bar{u}_{jk}^2,$$

i.e., SSTot = SSE+SSModel. For homogeneous data, the sums of squares attributable to the model (SSModel) would be near zero. Hence, to remove the effect of clusters of various means, we subtract this effect and rescale the data to unity. Define

$$c_j^2 = \sum_k \sum_{i \in S_k} (u_{ij} - \bar{u}_{jk})^2,$$

and rescale the data such that

$$u_{ij}^* = \frac{u_{ij}^*}{c_j}.$$

Notice that we scale differentially in each of the j dimensions to stretch and shrink accordingly to get the right scaling for homogeneous data.

The reason for the rescaling exercise is to stretch the data out so that the distances between elements in a cluster are as they would be if the mean of each cluster were 0. Note that

$$\sum_k \sum_{i \in S_k} (u_{ij} - \bar{u}_{jk})^2 / c_j^2 = 1,$$

while in the unscaled data

$$\sum_k \sum_{i \in S_k} u_{ij}^2 = 1.$$

Now we can find the distances between subjects as we have done for the homogeneous population, using the u_{ij}^* instead of u_{ij}. Match rescaled data using full matching and measure

the distances between cases and controls. A case (control) will be declared an outlier if the minimum distance to a control (case) exceeds the 99th quartile of the null distribution of distances.

After eliminating the outliers, we match the kept individuals using either full matching or pair matching and then perform conditional logistic analysis.

4 Simulations and Results

For each panel of reference SNPs generated we considered three approaches to correct for the effects of structure: (i) use PCA to infer the primary axes of variation and regress out these effects (ref); (ii) use PCA to determine full matching strata and analyze using conditional logistic regression; and (iii) use PCA to determine matched-pairs and analyze using conditional logistic regression. Although we compare the size and power of the PCA, full matching and pair matching methods of analysis, the methods are not competitors. The matching methods are designed to limit analysis to strata that are chosen to control for the effects of structure. The PCA method is not designed to select among available controls in the design of a study.

Our first battery of simulations is based on SNPs sampled from two subpopulations, with 200 individuals per subpopulation. Allele frequencies for the subpopulations were generated using the Balding-Nichols model (1), with allele frequencies varying uniformly between 0.05 and 0.5. To correct for structure L reference SNPs were available. Of these SNPs, 99% had $F_{st} = 0.01$ and 1% had $F_{st} = 0.1$. The result is a panel of SNPs that has a minor amount of information about the substructure such as one might anticipate among a sample from a single continental sample. Null candidate SNPs of three types were considered: (i) strongly differentiation SNPS, like the SNP in the lactase gene (LCT) with $F_{st} = 0.15$; (ii) moderately differentiated SNPs with $F_{st} = 0.03$; and undifferentiated, with $F_{st} = 0.01$.

Case status was assigned to 80 and 20 of the individuals from subpopulations 1 and 2, respectively. The remaining individuals were assigned control status. For the matched-pairs analysis we paired each case to the closest control until we obtained 100 matched pairs. For the other two methods of analyses we analyzed all 200 cases and controls. For power calculations, causal SNPs with relative risk $R = 2$ were generated using the approach described in [14].

Our second battery of simulations is based on nine subpopulations distributed along a

gradient designed to simulate a cline such as would be observed across Europe. The 100 cases are distributed with 2, 4, 6, 7, 9, 12, 15, 20 and 25 subjects in populations 1-9, respectively. The 300 controls are distributed randomly across the 9 subpopulations.

Our third battery of simulations is also based on nine subpopulations distributed along a gradient. This time 50 cases are distributed randomly in populations 6-9 (Figure 5). The 300 controls are distributed randomly across the 9 subpopulations.

Ten panels of reference SNPs were generated for each scenario to be investigated. For each of these panels, we simulated 1000 independent causal SNPs and 1000 independent null candidate for each type of null SNP investigated. Notice that for both scenarios of structure (pair of subpopulation and gradient) there are a sufficient number of controls to form homogeneous pair matches, provided the PC provide sufficient information to successfully differentiate the subpopulations.

In the first two scenarios, case-control ratio can be detected by PC. Table 3 illustrates that with a sufficient panel of SNPs the effects of substructure can be removed using any of the three methods, even the effects of highly differentiated SNPs such as LCT. We note that the gradient model is much easier to correct than the pair of subpopulations.

It is interesting to note that matched-pairs can correct for the effects of substructure with considerably less information than PCA or full matching (Table 3). At the same time, the matched-pairs analysis has somewhat less power to detect true positives (Table 4). Both of these results can be explained by the fact that matched-pairs design includes only 100 pairs (100 cases and 100 controls), while the other methods include 100 cases and 300 controls. Thus the other methods have greater power to detect association in the data, hence the greater power.

In the third scenario, the PC are not sufficiently informative to remove the effects of substructure (Table 5). Removing the outliers results in greater power for pair matching and full matching to detect this spurious structure.

4.1 Application

An Affy 500K platform genome scan applied to 462 cases with type 1 diabetes was compared with the POPGEN, SHIP and KORA panel of 2159 controls of which 1429 are from north and north-east Germany and 730 are from southern Germany. In this application, although most of the subjects are of European descent, both cases and controls exhibit complex population

heterogeneity.

In the initial phase of analysis, 27 dimensions were required to explain the significant axes of variation. Many of these axes are required to explain outliers. After removing 56 controls, only 7 important axes of variation remained (these subjects were more than 6 standard deviations from expectation in at least one of the 27 dimensions of the eigenvector space). Next we computed the distance between each case and the nearest control and vice versa based on 7 principal axes using equation (2). The resulting distribution of distances indicated that 24 cases and 19 controls could not be matched to a case/control with similar ancestry (individuals with minimum distances greater than 0.08 were removed). Repeating this process of finding the significant eigenvalues and the corresponding mimimum distances between cases and controls in the corresponding axes, we subsequently removed an additional 19 cases and 64 controls. Excluding these outliers, only 2 significant eigenvalues remain.

At this point we believe we have removed enough outliers to obtain reliable estimates of the principal eigenvectors. We performed a cluster analysis to isolate homogeneous strata. The data are clustered into 23 strata each with 20 or more elements. Using these strata, we rescale the data as described in the Methods. Based on our simualtions, observations with rescaled distances exceeding 0.065 are outliers. Using this criterion, an additional 7 cases and 46 controls are removed from the dataset. The resulting distances in the full matching set are consistent with expectations for cases and controls matched with homogeneous strata (Figure 4). Judging from the fact that the two principal axes separate the two German control samples (Figure 5), it appears that these dimensions explain important gradations in the European continent.

Finally, using the full matching algorithm, cases and controls were stratified based on their genetic ancestry into 36 levels. Due to the differing ancestry of the cases and controls, full matching strata have unequal representation. Most strata contain a single case and 1 or 2 controls; however, for some strata, many cases are matched to a single control (e.g., a single control matched to 34 cases) and vice versa (e.g., 169 controls matched to a single case).

Conditional logistic regression was performed on the stratified data. With matching, known variants still exhibit significant results. Table 2 counts the number of p-values smaller than 0.01 and 0.001. By using matching rather than a regression approach to remove confounding we reduced the Type I error rate by 33% or more.

p	Expected	PC10	cLOGIT
.01	3354	8583	5836
.001	335	1762	1027

Table 2: Expected and Observed number of small p-values

5 Discussion

We will begin this section with a result from Rosenberg et al. [17]: if enough markers are used with a sufficiently large worlwide sample, individuals can be partitioned into genetic clusters that match major geographic subdivisions of the globe, with some individuals from intermediate geographic location having mixed membership in the cluster that correspond to neighboring regions.

As we have already seen, the real data often compare cases and controls either coming from different populations or having different ancestry populations. When multiple continental groups, with clines, are analyzed simultaneously, the PC method faces problems with outliers and regressing beyond the range of data. Unlike eigenstrat, which regresses out the effects attributable to the principal components, the new approach follows the long epidemiological traditional of matched pairing.

Future work

The method proposed by Patterson is appropriate for detecting deviations from the null hypothesis of a homogeneous population. The current distance metric and eigenmap do not scale well as a function of the level of separation between populations. The distance between two nodes i and j changes when we add new data (the distribution of distances for the null case depends only on n, the number of subjects).

Consequently, there are two main issues with standard PC map that can be improved:

1. The sensitivity to outliers and the problem with having spurious "significant" eigenvalues.

2. The map/coordinates change if we add data.

Lets go back to the problem of Multidimensional Scaling and refer to the centered data matrix as a feature matrix

$$F = [f_1, \ldots, f_n]^T,$$

where

$$f_i = [(x_{i1} - \bar{x}_1)/s_1, \ldots, (x_{ip} - \bar{x}_p)/s_p]^T$$

is the feature vector of node i. Then the matrix K defined by $k_{i,j} =< f_i, f_j >= f_i^t f_j$ is an inner product or kernel matrix. The matrix K induces a natural distance between nodes:

$$d_{ij}^2 = k_{i,i} + k_{j,j} - 2k_{i,j},$$

which is exactly the Euclidean distance. The key observation here is that the definition of a kernel and metric is not unique. Any positive semi-definite (p.s.d.) matrix can induce an eigenmap and distance metric.

A possible solution is to define a p.s.d. kernel matrix K, suitable for these problems:

1. Define a p.s.d weight matrix with global weights (in the standard PC map, the weight matrix is $W = \frac{1}{p}FF^T$) and then normalize it. This would result in very little contribution of outliers to principal components structure.

2. Define K so that $k_{i,j}$ is a localized function of the Euclidean distances $(f_i - f_j)^t(f_i - f_j)$. (a Gaussian weight function is commonly used in applications, such as manifold learning, when preservation of local distances is needed).

Spreading of Sample Eigenvalues

Consider $n = 100$ subjects with $p = 400$ markers, from a homogeneous population. The sample covariance matrix S follows a Wishart density, and the population eigenvalues $l_j(I)$ are all equal to 1. There is a large spread of the sample eigenvalues $\hat{l}_j = \hat{l}_j(S)$.

The empirical distribution function of eigenvalues is defined by:

$$G_p(t) = \frac{1}{p}\#\{\hat{l}_j \leq t\}.$$

The limiting density function , if $\frac{n}{p} \to \gamma$, is

$$g(t) = \frac{\sqrt{(b_\pm - t)(t - b_\pm)}}{2\pi\gamma t}, b_\pm = (1 \pm \sqrt{\gamma})^2.$$

The larger p is relative to n, the more dispersed is the limiting density. Returning to the algorithm for finding outliers, we notice that the spread of the first l rescaled eigenvalues will vary with γ. Presently, the estimates for the first l rescaled eigenvalues are determined empirically. Finding the distribution of the lth largest eigenvalue for an identity covariance matrix will solve this problem.

References

[1] Balding DJ, Nichols RA. *A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity.* Genetica. 1995;96:3-12.

[2] Belkin M, Niyogi P. *Laplacian eigenmaps and spectral techniques for embedding and clustering.* Advances in Neural Information Processing Systems 14, Cambridge, MA, 2002. MIT Press.

[3] Campbell CD, Ogburn EL, Lunetta KL, et al:. *Demonstrating stratification in a European American population.* Nat Genet 2005;868-872.

[4] Cavalli-Sforza, L.;Menozzi, P.; Piazza, A. *The history and geography of human genes.* Princeton: Princeton University Press; 1994. 428 p.

[5] Devlin B, Roeder K. *Genomic control for association studies.* Biometrics. 1999;55:997-1004.

[6] Epstein MP, Allen AS, Satten GA. *A simple and improved correction for population stratification in case-control studies.* Am J Hum Genet. 2007 May;80(5):921-30.

[7] Everitt BS *Cluster Analysis* 1993 London: Edward Arnold

[8] Hinds DA, Stokowski RP, Patil N, Konvicka K, Kershenobich D, Cox DR, Ballinger DG. *Matching strategies for genetic association studies in structured populations.* Am J Hum Genet. 2004 Feb;74 (2):317-25

[9] Johnstone I. *On the distribution of the largest eigenvalue in principal components analysis.* Ann Stat. 2001;29:295-327.

[10] Lafon S, Lee AB. *Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning and data set parametrization.* Pattern Analysis and Machine Intelligence, 2006 Sep; 28(9):1393-1403

[11] Lee W-C. *Case-control association studies with matching and genomic controlling.* Genet Epidemiol 2004;27:1-13.

[12] Mardia KV, Kent JT, Bibby JM *Multivariate Analysis* 1979 London; Academic Press.

[13] Patterson N, Price AL, Reich D *Population Structure and Eigenanalysis* PLoS Genetics Vol. 2, 12

[14] Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, et al. *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet. 2006;38:904-909.

[15] Pritchard J, Stephens M, Donnelly P. *Inference of population structure using multilocus genotype data.* Genetics. 2000;155:945-959.

[16] Rosenbaum PR *Observational Studies* New York NY: Springer-Verlag, 1995

[17] Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, et al. (2005) *Clines, clusters and effect of study design on the inference of human population structure.* Plos Genet 1(6): e70

[18] Shi J, Malik J. *Normalized cuts and image segmentation.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1997,731:737

[19] Zhang S, Zhu X, Zhao H. *On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals.* Genet Epidemiol. 2003 Jan;24:44-56.

	PC			Pair matching			Full matching		
	Fst= 0.01	0.03	0.1	0.01	0.03	0.1	0.01	0.03	0.1
2 populations									
Markers									
96	69.350	106.125	211.188	62.050	99.925	202.275	65.000	101.275	205.812
386	54.737	61.175	85.200	44.325	45.225	51.300	46.700	49.212	54.175
1536	52.025	54.300	55.237	45.700	44.612	44.587	46.862	47.062	46.462
6144	52.850	52.000	51.450	44.263	45.050	44.737	46.725	46.763	45.812
12000	52.862	51.587	52.112	43.913	43.812	44.700	47.050	45.450	46.700
24000	52.700	50.062	51.050	43.350	42.312	41.337	45.688	45.788	45.288
Gradient									
Markers									
96	68.812	108.550	221.262	48.675	66.612	109.000	61.288	97.350	201.137
386	53.862	58.250	71.150	43.438	46.362	48.087	47.788	51.538	63.362
1536	52.237	50.987	50.438	44.525	44.188	44.438	46.550	46.663	46.462
6144	52.362	51.263	49.875	45.050	44.600	44.150	46.487	44.925	46.625
12000	52.362	51.837	51.700	44.325	43.925	44.737	47.375	47.075	47.388
24000	52.237	52.275	51.525	44.400	44.663	44.450	47.300	45.812	46.375

Table 3: Size of the 3 tests at level 0.05. The expected number of p-values smaller than 0.05 is 50.

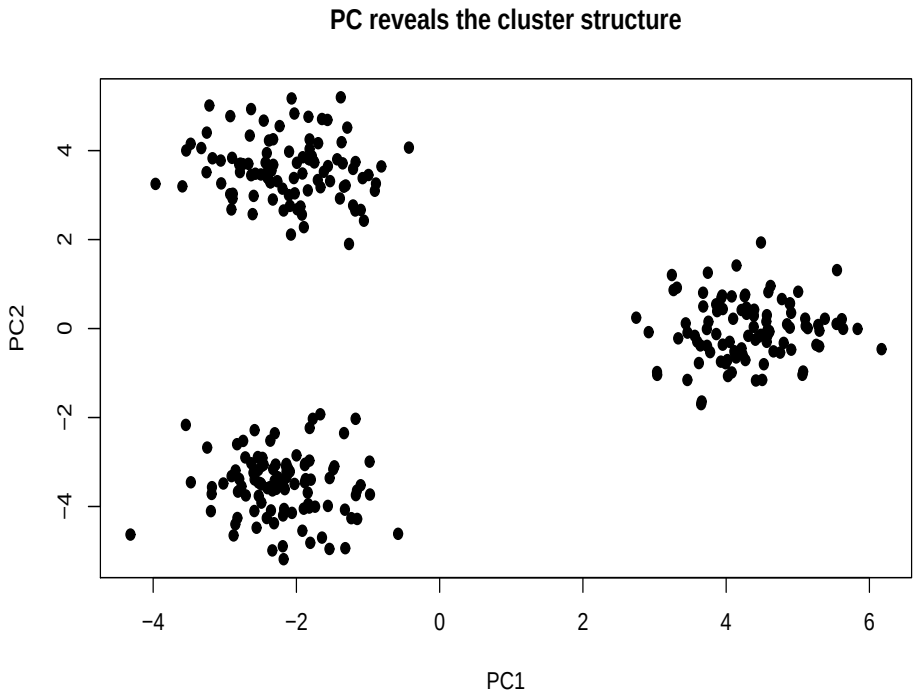
	PC			Pair matching			Full matching		
	Fst= 0.01	0.03	0.1	0.01	0.03	0.1	0.01	0.03	0.1
2 populations									
Markers									
96	782.638	709.975	682.688	693.175	635.175	619.550	754.175	684.750	659.450
386	768.775	701.237	681.587	682.288	631.712	621.038	734.913	673.100	653.250
1536	765.875	701.888	677.413	682.837	634.638	622.038	736.450	673.650	652.312
6144	764.737	694.175	676.263	683.575	630.350	623.087	735.188	670.938	652.600
12000	764.700	696.513	676.100	684.013	632.737	623.638	734.600	672.225	652.438
24000	763.000	696.500	677.413	683.550	632.288	623.638	733.550	670.837	652.612
Gradient									
Markers									
96	939.288	916.763	832.688	885.987	871.625	804.288	922.300	899.775	813.513
386	923.625	891.100	796.487	877.462	856.587	781.800	902.163	869.237	774.938
1536	917.112	875.538	774.475	875.913	849.688	774.800	894.100	855.900	754.038
6144	913.462	873.900	768.087	872.638	848.850	773.362	891.688	853.050	747.487
12000	914.538	873.225	771.188	873.875	848.763	774.025	890.975	850.188	748.888
24000	912.450	873.750	767.962	872.975	849.312	770.138	892.087	852.025	746.913

Table 4: Power of the 3 tests

	PC			Pair matching			Full matching			Pair matching, no outliers			Full matching, no outli		
	Fst= 0.01	0.03	0.1	0.01	0.03	0.1	0.01	0.03	0.1	0.01	0.03	0.1	0.01	0.03	0
Size															
Markers															
96	64.17	96.56	205.66	43.59	56.50	107.16	56.02	82.50	176.00	37.36	34.57	35.14	43.57	43.07	42.
386	56.75	68.13	120.91	37.56	40.08	49.75	45.16	50.66	69.80	36.07	36.36	35.36	42.57	41.21	43.
1536	55.96	62.27	102.65	36.58	36.75	37.08	43.14	42.08	44.87	34.57	38.14	33.36	40.29	43.79	40.
6144	56.31	60.88	84.88	36.72	36.69	35.48	41.27	42.36	40.98	39.79	37.07	37.21	44.07	44.14	42.
12000	58.04	58.02	73.17	36.99	36.05	35.90	41.96	40.27	41.36	38.21	35.29	33.14	41.07	42.36	41.
24000	56.73	58.26	66.58	37.36	36.99	35.27	42.35	41.79	39.62	36.14	39.43	34.64	40.71	45.00	41.
Power															
Markers															
96	804.3	753.2	650.2	589.9	578.7	510.7	769.7	725.8	622.9	706.1	713.2	655.9	771.9	776.1	719
386	784.4	731.0	630.0	583.2	566.1	489.5	720.6	686.5	583.1	698.3	713.4	655.6	770.9	769.4	726
1536	770.6	716.4	614.6	581.3	566.6	482.2	670.8	642.4	548.1	700.5	716.1	660.1	771.4	773.7	727
6144	761.8	711.0	603.8	583.1	566.2	485.0	639.5	620.3	531.3	702.9	716.4	666.0	770.3	774.0	728
12000	751.1	704.4	595.1	581.6	564.5	485.3	637.1	615.0	527.6	703.2	713.1	663.2	769.0	767.1	726
24000	746.1	698.8	593.1	583.9	564.6	484.4	636.9	613.4	529.4	701.1	715.4	667.3	771.0	775.1	726

Table 5: Size and Power the 3 tests before and after removing outliers. The simulated data is a gradient with 9 subpopulations from which 5 populations are outliers (controls with distances to the closest case greater than the maximum distance for the corresponding homogeneous population).

Figure 3:


PC reveals the cluster structure
PC2
PC1
−4
−2
0
2
4
6

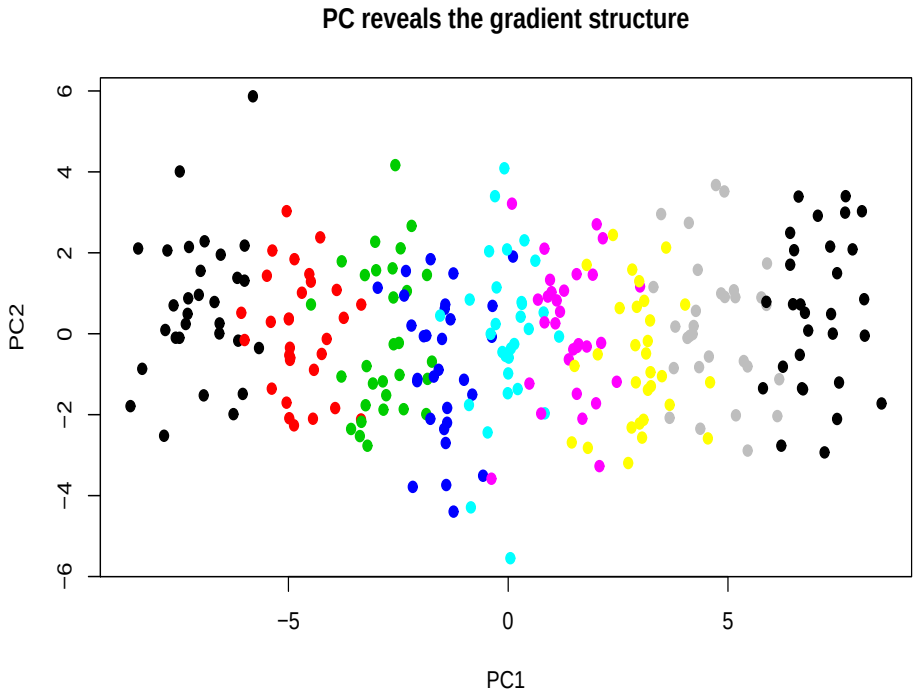

PC reveals the gradient structure
PC2
PC1
−6
−4
−2
0
2
4
6
−5
0
5

Figure 4: First 2 PC for third simulated data

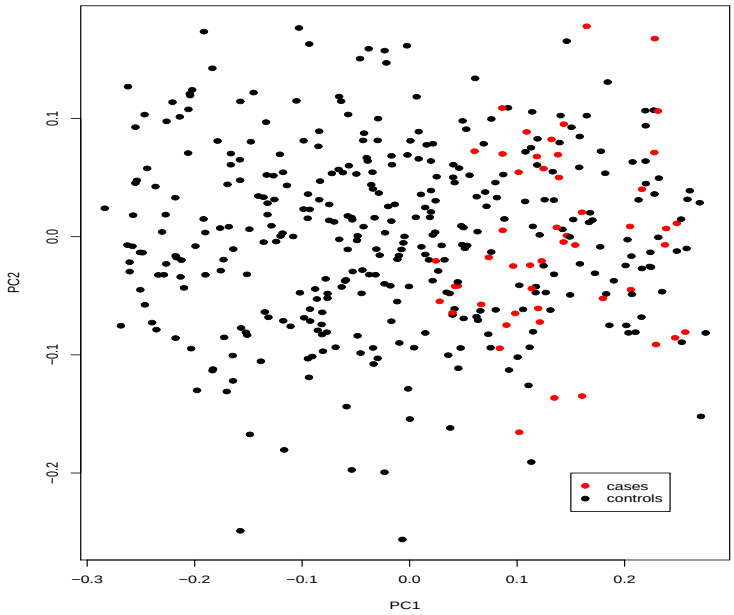


Figure 5: First 2 PC for T1D data

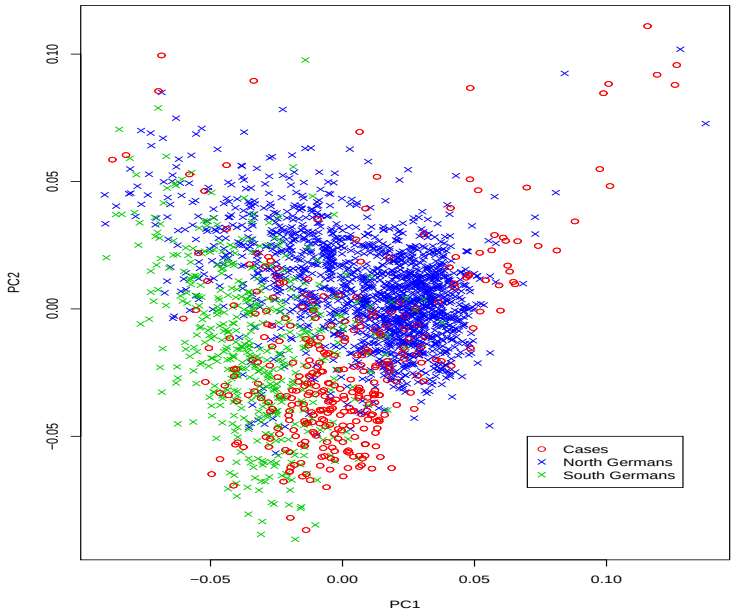
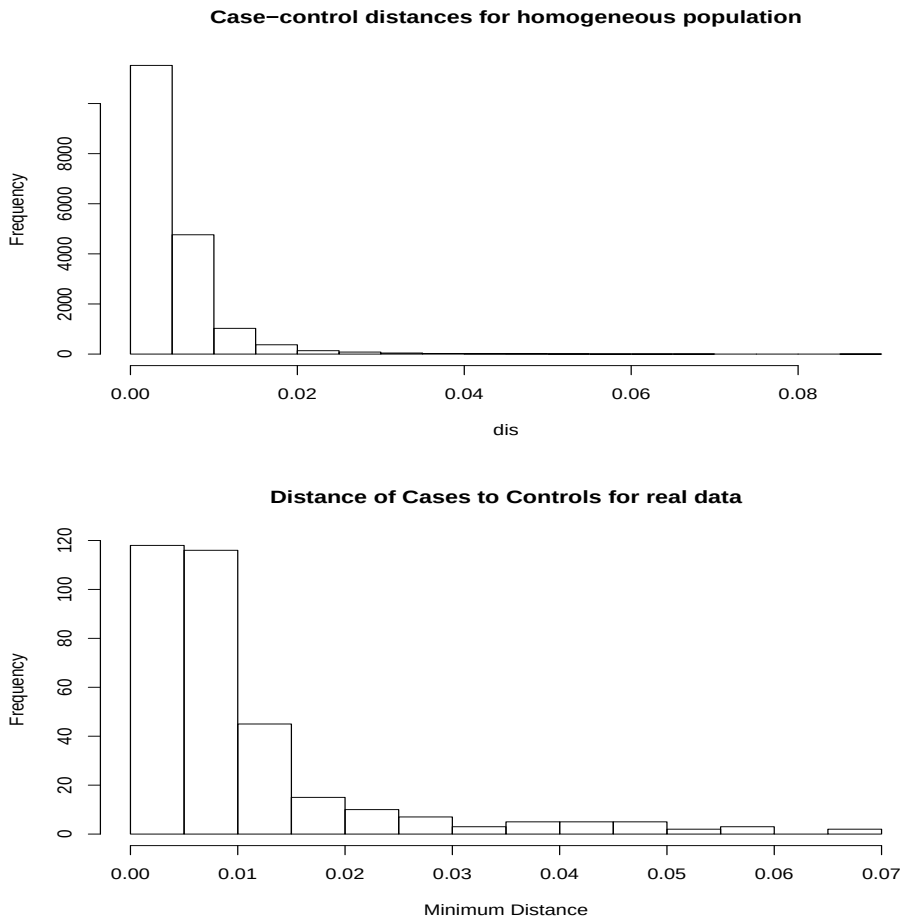


Figure 6:


Case−control distances for homogeneous population
Frequency
0
2000
4000
6000
8000
0.00
0.02
0.04
0.06
0.08
dis
Distance of Cases to Controls for real data
Frequency
0
20
40
60
80
100
120
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Minimum Distance