Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations

Katarzyna Bryc[1], Wlodek Bryc[2], Jack W. Silverstein[3]

Contact: kasia1@gmail.com

1) Department of Genetics, Harvard Medical School, Boston, MA; 2) Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 3) Department of Mathematics, North Carolina State University, Raleigh, NC

Background

- Principal component analysis (PCA) has been a powerful and efficient method for analyzing large datasets in population genetics since its early applications by Cavalli-Sforza and others.
- PCA of single nucleotide polymorphism genotype data can be used to illuminate population structure.
- A good estimate for the number of populations, K, is needed in Bayesian clustering algorithms such as STRUCTURE (Falush et al. 2003) or ADMIXTURE (Alexander et al. 2009), to infer relationships among individuals.

Table 1: False negative (error) rates of estimates of K, the number of subpopulations, using our threshold for identifying significant eigenvalues, across 50 simulated data from [1] sets under model *Split*. The simulated dataset is small, with only 50 individuals and 100 markers, underscoring the need for larger sample sizes to obtain power.

True K	2	3	4	5
$\mathbb{P}(\hat{K} < K)$	0.0	0.14	0.80	0.98

Summary of results

We present a mathematical model, and the corresponding mathematical analysis, that justifies and quantifies the use of principal component analysis of biallelic genetic marker data for a set of individuals to estimate the number of subpopulations represented in the data.

- The **raw** unprocessed covariance matrix is **more amenable to mathematical analysis**.
- The singular values of such raw data exhibit quantifiable properties that can be used directly to determine the number of populations in the data.
- Works in an almost deterministic fashion, at least when the number of individuals in the study is sufficiently large.

Major result:

We show that for large data sets of individuals from K well-differentiated subpopulations, with overwhelming probability the un-centered sample covariance matrix has K large eigenvalues.

In contrast to previous work, our results describe behavior of the eigenvalues of the sample covariance matrix without centering or normalization, taking into account both the number of individuals and the number of markers.

Conclusions about PCA

- Evidence that PCA is a robust technique for learning about population substructure.
- Contrary to current practice, for inference of substructure, we recommend applying PCA directly on the genotype data without centering or renormalization.
- We obtain K large eigenvalues in the presence of K subpopulations, justified by mathematical theory showing strong separation between the large eigenvalues corresponding to population structure and the remaining bulk of the distribution.
- Largely robust to LD, thinning of markers, and inbreeding.
- Inclusion of cryptic relatives in the dataset can profoundly influence the distribution of the bulk of the eigenvalues, making eigenanalysis challenging.

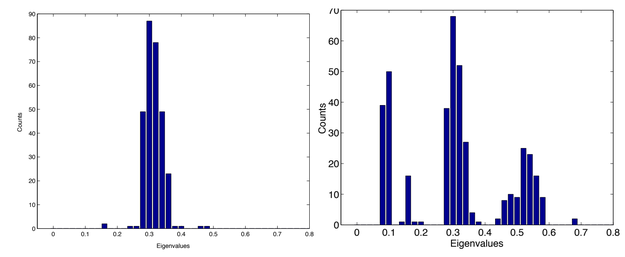

Figure 1: Expected distribution of the bulk of the eigenvalues, which can be confounded by inclusion of related individuals.

LEFT: Histogram of the eigenvalues from PCA of the HapMap CEU, CHB, and YRI unrelated individuals, excluding the three large eigenvalues ($\Lambda \gg 1$), which are omitted to better illustrate the shape of the non-significant eigenvalues.

RIGHT: The multi-modal histogram of the eigenvalues for PCA of three populations of HapMap (CEU, YRI, and CHB) including trios – 297 parents and their related 108 offspring. Large eigenvalues are not shown.

Proof-of-principle simulations that confirm the validity of our model and results

We demonstrate in two proof-of-principle simulations that we are able to obtain evidence of population structure when the number of individuals is large enough. The power to detect substructure relies more on the number of individuals than on the number of markers.

Simulations for a simple model

We generate simulations using overly simple model which allows us to compute all mathematical parameters to assess performance of our method. In this simulation, the site frequency spectra and population structure are known. Further, we fix the subpopulations to be independent. We draw unequal subpopulation sample individuals with proportions $c_1 = 1/6$, $c_2 = 1/3$, $c_3 = 1/2$. The theoretical population proportion $p_r(j)$ at each independent SNP for each subpopulation was selected from the same probability density function $\psi(x) = 0.5/\sqrt{x}$. We simulate individual genotypes for the j-th marker of an individual of the r-th subpopulation by choosing independent binomial values (with 2 trials) with probability of success $p_r(j)$.

Simulations using this simplistic model with $M = 1200$ and $N = 25000$ give eigenvalues:

$$(\mathbf{\Lambda_1}, \mathbf{\Lambda_2}, \mathbf{\Lambda_3}, \Lambda_4, \Lambda_5, \dots) = (\mathbf{48.2}, \mathbf{11.5}, \mathbf{5.8}, 0.27, 0.26, \dots)$$

The simulated eigenvalues match the theoretical predicted significant eigenvalues of $(47.4, 11.5, 5.7)$.

The threshold of 0.5 separates clearly the $K = 3$ largest eigenvalues, set in boldface, from the bulk.

Simulated genetic data under various demographic scenarios

Next we applied our method to simulated substructured datasets generated by coalescent simulations under various demographic scenarios from Gao *et al.* 2011 [1].

As shown in Table 1, In the case of such small sample sizes, we do not find strong power to correctly estimate K, however, we have no overestimates of the number of subpopulations K. in any of these sets of simulations.

Application to human population genotype data

Using HapMap 3 genotype data for the true substructure of the complete set of populations is unknown. We therefore report the performance of our theoretical analysis on the Yoruba, of Ibadan, Nigeria (YRI), European Americans from Utah (CEU), and Han Chinese from Beijing, China (CHB) that should have clear substructure.

We obtain evidence of three subpopulations as the eigenvalues of matrix $\mathbf{X}$ split into two sets: the non-significant, or small, eigenvalues in Figure 1 that lie below the cutoff of 0.5, and three large eigenvalues $\Lambda_1 = 102.0$, $\Lambda_2 = 14.55$, and $\Lambda_3 = 7.37$. The large eigenvalues exceed the cutoff of 0.5 , which matches our prediction for these three populations.

Challenges of cryptic relatives

The shape of the histogram of the distribution of eigenvalues is affected by relationships between the individuals. Including related individuals (offspring) in the HapMap PCA disturbs the unimodal tight distribution, see Figure 1.

Closely related individuals violate our random sampling assumption. These hidden, or "cryptic", relationships among individuals affect the applicability of our method for population structure by changing the distribution of eigenvalues, making it difficult to infer the correct cutoff for substructure. Pruning for LD does not improve the distribution, instead exclusion of related individuals improves the resolution of true substructure.

Robust to other violations of assumptions

Our assumption that that each marker is independent is violated by "linkage disequilibrium" (LD). Simulations indicate that LD does not strongly affect our ability to detect population structure, and thinning the data by removing one SNP from each pair of highly correlated markers (such as via the LD-pruning implemented in *PLINK*) is a simple yet robust technique.

Non-random mating violates stochastic independence of individuals. However, departures from HWE only slightly reduce the power of PCA for detecting population substructure, and it is possible to compensate inbreeding by increasing the number of individuals M.

Details of the model: setup, assumptions, and overview of the proof

Mathematical model setup

In setting up the mathematical model, we follow the notation as in Patterson *et al.* 2006 [2]. We consider M unrelated diploid individuals with N independent biallelic markers, in a large $M \times N$ rectangular array $\mathbf{C}$. The entries $C_{i,j}$ are the number of variant alleles of individual i at for marker j that take values $0, 1$ or 2. The individuals are from K subpopulations, with M_r individuals from subpopulation r.

Population parameters $F_{r,j}$ take values between 0 and 1 and correspond to inbreeding coefficients, or departures from expected allele frequencies. Conservatively, we write F for the largest value of the inbreeding parameter.

Our results describe the asymptotic behavior of the singular values of $\mathbf{C}$ as N increases. In general, our derivations rely on describing population parameters such that each locus or individual is viewed as a random sample from the population of all loci and individuals.

Assumptions

- We assume that if the population sampling information were known, namely, that individual i is from subpopulation r, the genotype probabilities for marker j, $\mathbb{P}(C_{i,j} = 0, 1, 2)$ would be given by the expected allele frequencies in subpopulation r , where $p_r(j)$ is the allele frequency of marker j in subpopulation r.
- For any pair of subpopulations labeled by $r, s \in \{1, \dots, K\}$, we assume that there are numbers $m_{r,s}$ such that

$$m_{r,s} = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} p_r(j) p_s(j) \qquad (1)$$

- We assume that the number of individuals, M, grows proportionally with N

The hidden parameters enter our mathematical analysis through a $K \times K$ (deterministic) symmetric positive matrix $\mathbf{Q}$ with entries

$$[\mathbf{Q}]_{r,s} = \sqrt{c_r c_s} m_{r,s}, \qquad (2)$$

We analyze $\mathbf{C}$ as a random perturbation of a finite-rank matrix The eigenvalues of $\mathbf{CC}'$ which we write in decreasing order $\Lambda_1 \geq \Lambda_2 \geq \cdots \geq \Lambda_M$, or of the alternative scaled matrix,X_N that are larger than the threshold t correspond to population substructure.

$$\mathbf{X}_N = \frac{1}{(\sqrt{M} + \sqrt{N})^2} \mathbf{CC}', \qquad t = \frac{1+F}{2} \qquad (3)$$

If there are K subpopulations present in the data, then as N and M increase without bound (and are subject to certain technical conditions), we prove that with overwhelming probability the smallest $M - K$ eigenvalues are smaller than t, and the largest K eigenvalues are much larger than t.

Increasing the number of individuals, M, increases the magnitude of the significant eigenvalues, and make it possible to resolve which eigenvalues correspond to population structure.

Distribution of the non-significant eigenvalues

Under a simple substructure scenario, the histogram distribution of the small eigenvalues should have a unimodal elliptical shape similar to the Marchenko-Pastur distribution, easily distinguished from large eigenvalues corresponding to substructure (Figure 1).

References

[1] H. Gao, K. Bryc, and C.D. Bustamante. On identifying the optimal number of population clusters via the deviance information criterion. *PloS one*, 6(6):e21014, 2011.

[2] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, 2006.

[3] Trevor J Pemberton, Chaolong Wang, Jun Z Li, and Noah A Rosenberg. Inference of unexpected genetic relatedness among individuals in hapmap phase iii. *The American Journal of Human Genetics*, 87(4):457–464, 2010.

[4] Eric L Stevens, Joseph D Baugher, Matthew D Shirley, Laurence P Frelin, and Jonathan Pevsner. Unexpected relationships and inbreeding in hapmap phase iii populations. *PloS one*, 7(11):e49575, 2012.