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

Katarzyna Bryc 1, Włodzimierz Bryc 1, Jack W. Silverstein 1

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

Contact: kasia1@gmail.com

Abstract

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.

Summary of results

We have a probabilistic 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.

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

In contrast to previous work, our results describe behavior of the eigenvalues of the sample covariance matrix with 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 N 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, shrinking 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.

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, needed in Bayesian clustering algorithms such as STRUCTURE (Falush et al. 2003) or ADMIXTURE (Alexander et al. 2008), to infer relationships among individuals.

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 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 and can be used in simulations to be independent. We draw unequal subpopulation sample individuals with proportions (0.5, 0.2, 0.3) = (0.5, 0.2, 0.3). The theoretical population proportion p(r,s) at each independent SNP for each subpopulation was selected from the same probability density function f = 0.5 N (0, 0.25).

Simulations using this simple model with K = 1200 and N = 20000 give eigenvalues:

\[
(\lambda_1, \lambda_2, \lambda_3, \ldots) = (-48.2, 11.5, 5.8, 2.7, 0.6, \ldots)
\]

The simulated eigenvalues match the theoretical predicted significant eigenvalues of \((1, 0.1, 0.01, \ldots)\).

The threshold of 0.5 separates clearly the K = 1 largest eigenvalues, set in boldface, from the bulk.

Application to human population genotype data

Using HapMap 3 genotype data for the true substructure of the complete set of populations is unknown. We report the performance of our theoretical analysis on the Yoruba, of Benin, with Asian 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 \(X^T X\) split into two sets: the 1, 2, 3 largest, or small, eigenvalues in Figure 1 that below the cutoff of 0.5, and three largest eigenvalues \(\lambda_1, \lambda_2, \lambda_3\), of which 2, 3, 7.37. The large eigenvalues exceed the cutoff of 0.5, which matches our prediction for these three populations.

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 independent individuals with \(K\) independent and statistically identical. In a large \(M = N\) rectangular array \(C\). The entries \(C_{i,j}\) are the number of variant alleles of individual \(i\) for marker \(j\) that take values 0 or 1. The individuals are from \(K\) subpopulations, with \(N\) individuals from subpopulation \(r\).

Population parameters \(P_{r,s}\) take values between 0 and 1 and correspond to inbreeding coefficients, or departures from expected allele frequencies. Conservatively, we write \(P\) for the largest value of the inbreeding parameter.

Our results describe the asymptotic behavior of the singular values of \(C\) 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 \(j\), \(P_{r,s}(i,j)\), were 0 or 1, would be given by the expected allele frequencies in subpopulation \(r\), \(r(j)\) is the allelic frequency of marker \(j\) in subpopulation \(r\).
- For any pair of populations labeled by \(r, s\in\{1,\ldots,K\}\), we assume that there are numbers \(m_{r,s}\), such that

\[
m_{r,s} = \max_{i,j} N^{-1} N_{r,i} N_{s,j} p(j|P_{r,s}(i,j)).
\]

Our results describe the asymptotic behavior of the singular values of \(C\) 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.

References