ON THE EQUIVALENCE OF METHODS FOR POPULATION STRATIFICATION AND THEIR APPLICATION IN GENETIC ASSOCIATION STUDIES

Nubia Esteban DUARTE[1]
Suely Ruiz GIOLO[2]
Mariza de ANDRADE[3]
Júlia Pavan SOLER[4]

- *ABSTRACT: Population-based association studies with unrelated individuals have been used in the mapping of genes involved in the regulation of complex diseases. However, when subjects are from different ethnic ancestries, these studies may yield spurious associations due to population stratification, with an excess of false positive or negative results. Principal components analysis based either on genotype values from known genetic markers (columns of the matrix) or on individuals (rows of the matrix) are the most common approaches used for correction of the confounding effect due the population stratification in genetic association studies. In this paper, results from the singular value decomposition theory of matrices are used to show the analytical equivalence between these approaches, focusing mainly in their relevant role in population stratification analysis. It is also shown the importance of using the biplot as a visualization tool not only to explain the joint information of samples and genetic markers but also to detect informative markers. Although both procedures can be used to correct for population stratification, principal components analysis based on samples is more computationally feasible due to the large number of genetic markers ($n << p$ problem). As an application, it is used genotype data from four HapMap populations.*

- *KEYWORDS: Principal components; singular value decomposition; duality in genetic matrix.*

[1] Universidade de São Paulo – USP, Faculdade de Medicina, Laboratório de Genética e Cardiologia Molecular, CEP: 05403-900, São Paulo, SP, Brasil. E-mail: *nued03@gmail.com*

[2] Universidade Federal de Paraná – UFPR, Departamento de Estatística, CEP: 81531-990, Curitiba, PR, Brasil. E-mail: *giolo@ufpr.br*

[3] Mayo Clinic, Division of Biomedical Statistics and Informatics, Rochester, Minnesota, USA. E-mail: *mandrade@mayo.edu*

[4] Universidade de São Paulo – USP, Instituto de Matemática e Estatística, CEP: 05508-090, São Paulo, SP, Brasil. E-mail: *pavan@ime.usp.br*

1 Introduction

The mapping of human genes is an important step for personalized medicine and other aspects of health care (Altshuler, 2008). In this context, one important result of the Human Genome Project was the discovery of millions of DNA sequence variants in the human genome. Since the majority of these variants is single nucleotide polymorphisms (SNPs), such markers have received special attention in recent years (Bush, 2012). High-dimensional SNPs platforms have been used to represent the human genome make up to differentiate the phenotypes between individuals and some SNPs have been directly associated to the genetic mechanism of complex diseases in humans.

One issue in those studies of complex diseases is that population stratification may yield spurious phenotype-genotype association due to the differences in SNP allelic frequencies between populations. This leads to a confounding effect between the SNP and the ancestry of the individuals. Thus several methods to detect and correct for population stratification were proposed to control the excess of false positive or negative rates (Tiwari et al., 2008).

In this article, the focus is on two alternative methodologies, one by Zhang et al. (2003) and the other by Price et al. (2006), to estimate genetic background variables as global ancestry coefficients. In the situation of having a matrix $\mathbf{G}$ of dimension $N \times M$, where N and M are the numbers of individuals and of genetic markers, Zhang et al. (2003) proposed the spectral decomposition of the $M \times M$ covariance matrix among markers (standard principal components based on the variables/columns space), while Price et al. (2006) performed such decomposition on the $N \times N$ covariance matrix among individuals (eigenanalysis based on the individuals/rows space).

Several authors had also introduced methods to correct for population stratification using either the M-columns or the N-rows of genetic matrix. For instance, Bauchet et al. (2007) suggested the use of principal coordinates analysis to summarize genetic marker data by operating on the rows. Chae and Warde (2006) had also argued that principal coordinates analysis was more powerful than principal component analysis to ensure identification of groups of individuals only if some conditions were satisfied, while Zhu et al. (2008) applied principal component analysis to adjust for the effect of population stratification, and suggested the equivalence of the information obtained by these two methods due to the duality of these techniques. Zheng et al. (2012) provided a simple proof showing that eigenanalysis based on individuals and the standard principal component analysis based on genetic variables were the same up to a constant. Furthermore, Ringnér (2006) showed how principal component analysis could be used to explore high-dimensional gene expression data set.

Many applications in multivariate analysis usually start on the rows and columns of a rectangular data matrix. For example, Gower (1966) described the duality of Q and R techniques, where the Q technique focused on a $N \times N$ matrix, whose values were measures of association between the individuals, and the R technique focused on $M \times M$ matrix, whose values defined the relationship among the variables (markers). Furthermore, it also provided analytical discussion of the Q and R techniques on the context of principal components, factor analysis, and multidimensional scaling (or principal coordinate analysis). Gabriel (1971) introduced the use of biplots as a graphical display of rows and columns of a rectangular data matrix. Subsequently, the concept of biplots was extended to different representations of individuals, variables, and their interaction in two dimensions plots (Greenacre, 1984, Sajesh and Srinivasan, 2008). Ringnér (2006) applied this graphic to visualize the interdependency between individuals and the identification of clusters, with the assumption that there was an association between individuals and markers.

Despite the rich literature using principal component analysis for correction of population structure in genetic association studies, there is still room to formalize and extend the equivalence between eigenanalysis based on the variables and individuals space, mainly on the big data context under the $n << p$ problem. Then, the specific aims of this paper are, firstly to establish an analytical equivalence between these two approaches using the theorem of singular-value decomposition in large genomic data matrices, and secondly to explore the use of biplots to represent and visualize the joint information of individuals and molecular markers.

The paper is organized as follows. The Methods section presents the analysis in the spaces $\Re^M$ and $\Re^N$ with the methods proposed by Zhang et al (2003) and Price et al. (2006) by emphasizing their analytical equivalence, and the characterization of biplots. In the Results section is present our analysis performed on a sample of unrelated individuals from the HapMap Project. The relevance of the analytical equivalence is emphasized in the Discussion and Conclusions section.

2 Methods

2.1 Methodology for population stratification

Let $\mathbf{G} = \{g_{ij}\}$ be a matrix of dimension $N \times M$, where g_{ij} represents the genotypic value (0, 1 or 2 for aa, Aa and AA, respectively) of the i-th individual in the j-th marker, $i = 1, \ldots, N$ and $j = 1, \ldots, M$. Let define the normalization,

$$\mathrm{x}_{ij} = \frac{1}{\sqrt{N}}\left(\frac{g_{ij} - \bar{g}_j}{s_j}\right), \tag{1}$$

where $\bar{g}_j$ is the arithmetic mean and s_j is the standard deviation of variable j (j-th marker). By defining $\mathbf{X}_{N\times M} = \{\mathrm{x}_{ij}\}$ the matrix with the elements of $\mathbf{G}_{N\times M}$ after the normalization described in equation (1), one can obtain two matrices. The first, defined by $\mathbf{\Sigma}_{M\times M}$, is the covariance among the M markers, and the second, $\mathbf{\Psi}_{N\times N}$, is the covariance among the N individuals.

Since the interpretation between rows and columns of a data matrix allows for the representation of the distances between individuals as well as the covariances between variables (markers), one can easily represent geometrically the N rows and

M columns of a matrix by points, where the set of points can be: a) the set of N individuals in the M-dimensional space $\Re^M$, or b) the set of M variables in the N-dimensional space $\Re^N$.

In the case where there are missing genetic markers, it is recommended to remove the missing x_{ij} prior to the genotype normalization, and after the normalization to replace the missing x_{ij} by the value zero (Patterson et al., 2006).

2.2 Analysis in the $\Re^M$ space

In the $\Re^M$ space (R technique) the main purpose is to evaluate the relationship between individuals, where the distance between individuals i and i' is defined as

$$d^2(i,i') = \sum_{j=1}^{M}(x_{ij} - x_{i'j})^2 = \frac{1}{N}\sum_{j=1}^{M}\left(\frac{g_{ij} - g_{i'j}}{s_j}\right)^2, \tag{2}$$

$j = 1, ..., M$ and $i, i' = 1, ..., N$.

The spectral decomposition of the matrix $\mathbf{X}'\mathbf{X}$, denoted by $\mathbf{\Sigma}_{M\times M}$, can be written as $\mathbf{\Sigma}_{M\times M} = \mathbf{V}\mathbf{\Lambda}_{\Sigma}\mathbf{V}'$, where $\mathbf{V}_{M\times M} = (\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_M)$ is the set of the M orthonormal eigenvectors and $\mathbf{\Lambda}_{\Sigma}$ is the diagonal matrix of eigenvalues. The principal components obtained from the spectral decomposition of the matrix $\mathbf{\Sigma}_{M\times M}$ are the columns of the matrix $\mathbf{T}$ given by

$$\mathbf{T} = \mathbf{X}\mathbf{V}, \tag{3}$$

where the coordinates of the N individuals in the factorial axis $\mathbf{V}_k$ (k-th eigenvector) are the N components of the vector $\mathbf{T}_k = \mathbf{X}\mathbf{V}_k$.

In order to find latent variables to correct for population stratification in genetic association studies, Zhang et al. (2003) used the standard principal components of the spectral decomposition of the covariance matrix $\mathbf{\Sigma}_{M\times M}$ between markers given by equation (3). For instance, to obtain the k-th latent variable t_{ik} (for each individual i), the eigenvector $\mathbf{V}_k$ associated with the k-th eigenvalue of $\mathbf{\Lambda}_{\Sigma}$ is used in the following equation,

$$t_{ik} = \mathbf{X}_i\mathbf{V}_k = \sum_{j=1}^{M} x_{ij}v_{jk},$$

where $\mathbf{X}_i$ is the row vector with the genotypic information of the M markers of the i-th individual (row i of the matrix $\mathbf{X}$). If the distance between individuals i and i' is small, it is expected that the values of t_{ik} and $t_{i'k}$ are similar. Under Zhang et al.'s approach, the latent variable t_{ik} can be included as a covariate in the semiparametric regression model to correct for population stratification in genetic association studies.

2.3 Analysis in the $\Re^N$ space

In the $\Re^N$ space (Q technique) the goal is to evaluate the relationship between the variables (markers), where the Euclidean distance between two variables j and j' is given by,

$$\begin{aligned} d^2(j,j') &= \sum_{i=1}^{N}(x_{ij} - x_{ij'})^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{g_{ij}-\bar{g_j}}{s_j} - \frac{g_{ij'}-\bar{g_{j'}}}{s_{j'}}\right)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{g_{ij}-\bar{g_j}}{s_j}\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{g_{ij'}-\bar{g_{j'}}}{s_{j'}}\right)^2 \\ &\quad - 2\frac{1}{N}\sum_{i=1}^{N}\left(\frac{g_{ij}-\bar{g_j}}{s_j}\frac{g_{ij'}-\bar{g_{j'}}}{s_{j'}}\right). \end{aligned}$$

Thus, the distance between two variables depends on their correlation, $d^2(j,j') = 2(1 - r_{jj'})$, where $r_{jj'}$ is the correlation coefficient between variables j and j'. Therefore, the proximity between two variables can be explained by their correlations. Strong positive correlation ($r_{jj'} \approx 1$) means that the variables are close, and strong negative correlation ($r_{jj'} \approx -1$) indicates that they are distant. Intermediate distances ($r_{jj'} \approx 0$) correspond to independent variables.

To represent the variables, let us consider the decomposition of the covariance matrix among the N individuals, $\mathbf{\Psi}_{N\times N}$, where $\Psi_{ii'}$ is the covariance between rows i and i' of matrix $\mathbf{X}$, or equivalently, the spectral decomposition of the matrix $\mathbf{XX}'$ given as $\mathbf{\Psi}_{N\times N} = \mathbf{U\Lambda_\Psi U}'$, where $\mathbf{U}_{N\times N} = (\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_N)$ denotes the N eigenvectors and $\mathbf{\Lambda}_\Psi$ contains on its diagonal the eigenvalues of $\mathbf{\Psi}$. The elements of $\mathbf{U}_k$, u_{ik}, $i = 1, \ldots, N$, are the ancestry of the individuals over the k-th axis of variation (Price et al., 2006).

The principal coordinates obtained from the spectral decomposition of matrix $\mathbf{\Psi}_{N\times N}$ are the columns of the matrix $\mathbf{F}$ given by,

$$\mathbf{F} = \mathbf{X}'\mathbf{U}, \tag{4}$$

where the coordinates of M variables in the factorial axis $\mathbf{U}_k$ (k-th eigenvector) are the M components of the vector $\mathbf{F}_k = \mathbf{X}'\mathbf{U}_k$, and the coordinate f_{jk} of the variable j on axis k is given as $f_{jk} = \sum_{i=1}^{N} x_{ij}u_{ik}$.

In genetic association studies, Price et al. (2006) proposed to correct for population stratification using the ten most significant eigenvectors $\mathbf{U}$ (that contain the ancestry of the individuals) as covariates in the regression model. This approach simplifies the analysis because the eigenvectors are orthogonal and uncorrelated. If the correlation between the variables j and j' is small, the values of u_{ik} and $u_{ik'}$ are expected to be similar. This approach is implemented in the EIGENSOFT package (Price et al., 2006; Patterson et al., 2006).

2.4 Analytical equivalence between the methodologies

The principal components analysis on a symmetric matrix $M \times M$ and the principal coordinates analysis on a matrix $N \times N$ are considered dual to each other when both lead to a set of points with the same inter-points distances as shown by Gower (1966). Then, the goal of this section is to investigate this duality using the singular-value decomposition of the rectangular array $\mathbf{X}_{N\times M}$ of rank N (Johnson and Wichern, 2007).

The relationship between the vectors of the matrices $\mathbf{U}_{N\times N}$ and matrix $\mathbf{V}_{M\times M}$ can be established by the relationship given by,

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}', \tag{5}$$

where the columns of $\mathbf{U}_{N\times N}$ and $\mathbf{V}_{M\times M}$ are the eigenvectors associated with the nonzero eigenvalues of $\mathbf{X}\mathbf{X}'$ and $\mathbf{X}'\mathbf{X}$, respectively, and $\mathbf{\Lambda}_{N\times N}$ is the diagonal matrix of the eigenvalues, also known as singular values of matrix $\mathbf{X}$. By multiplying both sides of the equation (5) by $\mathbf{V}$, we obtain $\mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\mathbf{V}$. Since $\mathbf{V}'\mathbf{V}$ is equal to the identity matrix, then $\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{\Lambda}^{-1/2}$, where $(\mathbf{U}_1, \ldots, \mathbf{U}_k, \ldots, \mathbf{U}_N)$ is equal to $(\mathbf{X}\mathbf{V}_1\lambda_1^{-1/2}, \ldots, \mathbf{X}\mathbf{V}_k\lambda_k^{-1/2}, \ldots, \mathbf{X}\mathbf{V}_N\lambda_N^{-1/2})$.

It is important to know that the k-th column $\mathbf{U}_k$ of matrix $\mathbf{U}_{N\times N}$ contains the ancestry information u_{ik} of each individual i, where u_{ik} is a linear combination of the genotypic information of the M markers of the individual i (row i of $\mathbf{X}$) and the coordinates of each marker over the vector $\mathbf{V}_k$ standardized by the root square of the eigenvalue λ_k. Then, the scores of ancestries can be interpreted as the preponderance of the markers on the variability of the individuals, where

$$\mathbf{U}_k = \mathbf{X}\mathbf{V}_k\lambda^{-1/2} = \lambda^{-1/2}\mathbf{T}_k, \tag{6}$$

whereby

$$u_{ik} = \mathbf{X}_i\mathbf{V}_k\lambda^{-1/2} = \lambda^{-1/2}\sum_{j=1}^{M} x_{ij}v_{jk} = \lambda^{-1/2}t_{ik}.$$

Thus, the ancestries of individuals in the k-th axis can be obtained by the scores of the k-th principal component and vice-versa, establishing the analytical equivalence between the approaches proposed by Zhang et al. (2003) and Price et al. (2006) to obtain latent variables to control for population stratification in genetic association studies. Similarly, by multiplying both sides of equation (5) by $\mathbf{U}'$, we have $\mathbf{U}'\mathbf{X} = \mathbf{U}'\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'$ and as $\mathbf{U}'\mathbf{U}$ is equal to the identity matrix, then $\mathbf{U}'\mathbf{X} = \mathbf{\Lambda}^{1/2}\mathbf{V}'$ and therefore, $(\mathbf{V}_1, \ldots, \mathbf{V}_k, \ldots, \mathbf{V}_N)$ is equal to $(\mathbf{X}'\mathbf{U}_1\lambda_1^{-1/2}, \ldots, \mathbf{X}'\mathbf{U}_k\lambda_k^{-1/2}, \ldots, \mathbf{X}'\mathbf{U}_N\lambda_N^{-1/2})$.

Note that the k-th column $\mathbf{V}_k$ of matrix $\mathbf{V}_{M\times N}$ contains the coordinates v_{jk} for each marker j. In this case, v_{jk} is the linear combination of column j of matrix $\mathbf{X}$, which has the genotypic information of the marker j of all N individuals, and the vector $\mathbf{U}_k$ represents the ancestry of each individual standardized by the root square of the eigenvalue λ_k. Hence, the coordinates v_{jk} can be interpreted as the contribution of the N individuals to the variability among the markers.

By replacing $\mathbf{U}_k = \mathbf{X}\mathbf{V}_k\lambda^{-1/2}$ of equation (6) in equation (4), we have,

$$\mathbf{F}_k = \mathbf{X}'\mathbf{U}_k = \mathbf{X}'\mathbf{X}\mathbf{V}_k\lambda^{-1/2} = \mathbf{X}'\mathbf{T}_k\lambda^{-1/2},$$

so that the coordinate of the variable j on the axis k is given by,

$$f_{jk} = \sum_{i=1}^{N} x_{ij}\frac{t_{ik}}{\sqrt{\lambda_k}} = r_{jk},$$

where r_{jk} is the correlation coefficient between the variable j and the main component k. It is worth noticing that the above results were obtained directly from the factor analysis approach.

As shown above, the vectors of matrix $\mathbf{V}$ are functions of the vectors of matrix $\mathbf{U}$ and vice versa. Thus, this result is very useful to circumvent computational problems arising from the decomposition of the matrix $\mathbf{\Sigma}_{M\times M}$ whose dimension can be extremely large, especially in association studies with large number of molecular markers (such as SNPs). Furthermore, from a matrix $\mathbf{X}$, in which each row or column vector represents a point in the $\Re^M$ or $\Re^N$ spaces, it was showed the correspondence of axes (ie, t_{ik} is the coordinate of the individual i in the k-th axis and f_{jk} is the coordinate of the variable j in the k-th-axis). This correspondence is the key concept for the study of biplots described in the next section.

2.5 Biplots

Biplots are commonly used to display a graphical joint representation of the rows and columns of the matrix $\mathbf{X}$. Therefore, it can be very useful to perform a simultaneous analysis of the relationships among individuals and markers (variables). It is based on the same principle of dimensionality reduction techniques through the singular-value decomposition of the matrix $\mathbf{X}$ (equation 5). The fundamental difference is that the goal is to reproduce the original data through a joint representation of rows and columns (Gabriel, 1971). In the graphical representation, biplots are usually plotted in two dimension (using two singular vectors), where the individuals are represented by points and the variables (markers) are displayed either as a vectors, linear axes or trajectories. Hence, the matrix $\mathbf{X}_{N\times M}$ is approximated by

$$\begin{aligned}\mathbf{X}_{N\times M} \approx \mathbf{U}_{(2)}\mathbf{\Lambda}_{(2)}^{1/2}\mathbf{V}'_{(2)} &= (\mathbf{U}_{(2)}\mathbf{\Lambda}_{(2)}^{1/2-c/2})(\mathbf{\Lambda}^{c/2}\mathbf{V}'_{(2)}) \\ &= \mathbf{G}\mathbf{H}', \end{aligned} \tag{7}$$

where $\mathbf{U}_{(2)}$ and $\mathbf{V}_{(2)}$ indicates only two eigenvectors of the matrices $\mathbf{U}$ and $\mathbf{V}$, respectively, $\mathbf{G} = \mathbf{U}_{(2)}\ \mathbf{\Lambda}_{(2)}^{1/2-c/2}$ and $\mathbf{H}' = \mathbf{\Lambda}_{(2)}^{c/2}\mathbf{V}'_{(2)}$ with $c \in [0, 1]$. Here $\mathbf{G}$ is the representation of the N rows of the matrix $\mathbf{X}$ in a bi-dimensional space and $\mathbf{H}$ is the representation of the M columns of the matrix $\mathbf{X}$ in this same space. Values most commonly used for c are 0, 1/2 and 1. When $c = 1$, the matrices $\mathbf{G}$ and $\mathbf{H}'$ from equation (7) are written as

$$\mathbf{G} = \mathbf{U}_{(2)} = \{u_{ij}\} \qquad \text{and} \qquad \mathbf{H}' = \mathbf{\Lambda}_{(2)}^{1/2}\mathbf{V}'_{(2)}$$

and by equation (7), $\mathbf{H}' = \mathbf{U}'_{(2)}\mathbf{X}$, where, from equation (4), $\mathbf{H}' = \mathbf{U}'_{(2)}\mathbf{X} = \mathbf{F}'_{\mathbf{(2)}} = \{f_{jk}\}$. Then, the matrix $\mathbf{G}$ contains the ancestries of N individuals associated with the axes $k = 1$ and $k = 2$, and the matrix $\mathbf{H}$ contains the scores of M markers associated with these same axes. The term principal coordinates is used to refer to the singular vectors scaled by the singular values (for instance, matrix $\mathbf{H}$ with $c = 1$, or matrix $\mathbf{G}$ with $c = 0$), while standard coordinates are the unscaled singular vectors (Greenacre, 1984).

Finally, the biplot can be displayed in a bi-dimensional plane of the joint representation of N points of individuals ($\mathbf{U}_1\lambda^{1/2-c/2} \times \mathbf{U}_2\lambda^{1/2-c/2}$) and M points corresponding to the molecular markers ($\mathbf{V}_1\lambda^{c/2} \times \mathbf{V}_2\lambda^{c/2}$). For $c = 0$, rows are in principal coordinates and columns are in standard coordinates, called the form biplot, which favors the display of the individuals; for $c = 1$, rows are in standard coordinates and columns are in principal coordinates, called the covariance biplot, which favors the display of the variables. When $c = 1/2$, the biplot favors the display of interaction effect.

It is important to notice that the biplot is a principal component analysis, in which information of the columns (markers) are added on the same graph in which the rows (individuals) are represented. Since, in the biplots, individuals and variables are represented on the same plot, it makes sense to evaluate associations between individuals and variables (markers) as the preponderance of one variable (marker) to explain an individual, or the contribution of individuals to the values of a variable (marker). The variability explained by the biplot axes is similar to the variability explained by the principal component analysis, where the axes are obtained in the direction of greatest variability.

3 Results

As an application, it was used the genotyped data of M = 491 SNPs (molecular markers) for N = 397 unrelated individuals from four populations of the HapMap Project, Phase III, where 113 are Utah residents with ancestors from Northern and Western Europe (CEU); 85 Chinese from Denver, Colorado (CHD); 86 Japanese from Tokio, Japan, and 113 Yoruba from Nigeria (YRI) (HapMap, 2003). The set of 491 SNPs consists of ancestry informative markers (AIMs), selected as the markers with the highest loading scores associated with the ten first principal components of the spectral decomposition of the covariance matrix between all the common SNPs available for the eleven populations of the HapMap Project, Phase III (http://hapmap.ncbi.nlm.nih.gov/).

For this data set, it was performed the decomposition in singular values and vectors of the matrix $\mathbf{X}$ obtained after the normalization of the original data described in equation (1). Figure 1 displays the biplot using $c = 1/2$ with the 397 individuals and the 491 SNPs where individuals are represented by points (gray) and SNPs are represented by vectors (black).

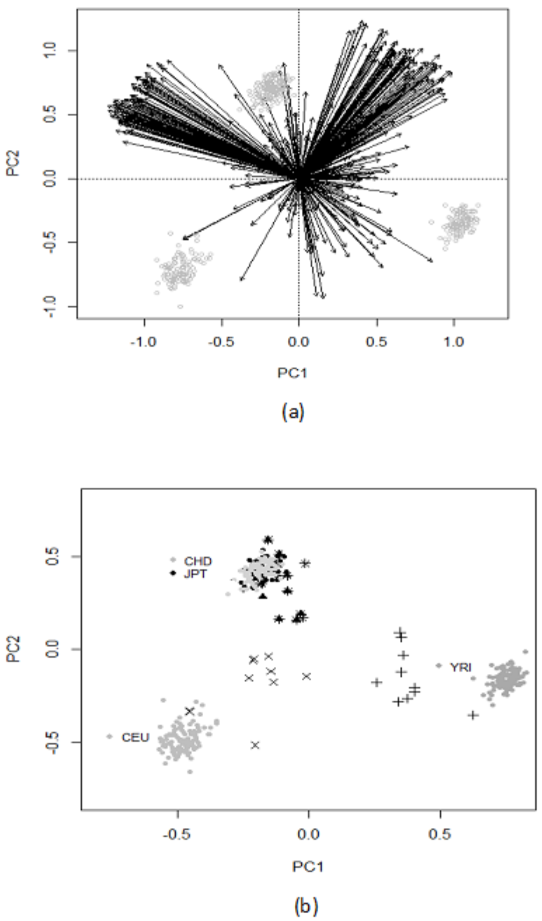

Figure 1 - Biplots display: (a) Biplot showing the 397 individuals (gray points) and the 491 SNPs (black vectors); (b) Biplot showing the 397 individuals divided into the four populations and the ten SNPs closest to the centroid of each population (indicated as cross signals). The gray points represents the European population (CEU), black and light gray points represents the Chinese and Japanese populations (CHD, JPT) and dark gray points represents the Yoruba population (YRI).

In Figure 1(a) can be observed three clusters, one of Asian individuals (CHD and JPT), another of African individuals (YRI) and the last of European individuals (CEU). Figure 1(b) shows the ten SNPs closest of the centroid of each cluster, indicating that these ten SNPs are those which mostly contribute to the discrimination of the population.

Table 1 shows the distribution of the ten SNPs closest to the centroid of each cluster, over the 22 chromosomes, as well as that one for the hundred SNPs closest to the centroids. Specifically, the first column identify the chromosome, the following four columns presents the frequency of the 10 SNPs nearest to the centroid of each population (the 10 SNPs distributed in the 22 chromosomes) and, analogously, the last four columns present the frequency of the 100 SNPs closest to the centroid.

Table 1 - **Frequency distribution over the 22 autosomal chromosomes of the 10 and 100 SNPs closest to the centroid of each population.**

	10 SNPs				100 SNPs			
Chromosomes	CEU	CHD	JPT	YRI	CEU	CHD	JPT	YRI
1	2				10	8	8	9
2	2	1	1	1	10	8	9	12
3		1	1		9	9	9	7
4		2	2	1	9	11	12	6
5		1			4	6	5	6
6					6	5	5	4
7			1		0	3	3	3
8	1	2	2	3	6	9	9	11
9	2	1	1	1	5	4	4	4
10					8	7	7	8
11	1			1	6	2	3	3
12					4	4	3	3
13				1	3	2	2	3
14					3	4	4	0
15		1	1	1	3	4	3	2
16					4	1	1	3
17					1	2	2	0
18		1	1		2	3	3	4
19					2	1	1	4
20	1				3	1	1	5
21					0	1	1	1
22	1			1	2	5	5	2

Table 2 identifies such 10 SNPs closest to the centroids by their names. When the 10 SNPs closest to the centroid are considered, the highest concentration of them is observed in chromosome 8, while for 100 SNPs, highest concentrations are observed in chromosomes 1, 2, 3, 4 and 8.

Table 2 - **Annotation of the ten SNPs closest to the centroid of each population in the HapMap Project.**

	CEU	CHD	JPT	YRI
1	rs10511625	rs12954302	rs7168890	rs10458360
2	rs4961736	rs7168890	rs12954302	rs41170
3	rs916204	rs6770520	rs6770520	rs12680859
4	rs4668082	rs349200	rs349200	rs355183
5	rs2403732	rs6722663	rs7864878	rs11787250
6	rs7522358	rs12235034	rs4373455	rs12618983
7	rs11237061	rs7864878	rs12235034	rs9599954
8	rs7266805	rs7695667	rs6722663	rs10993768
9	rs1543061	rs4256368	rs7695667	rs4871822
10	rs7365057	rs2354895	rs2354895	rs12577127

4 Discussion and conclusion

In this paper, was discussed the duality of Q and R techniques in the context of the spectral decomposition of rectangular matrices using genomic data. The proposal of this article was to show the issues related to spurious association caused by population-stratification in genetic association studies, as well as the computational difficulties and methodologies currently available in the analysis of genomic data of high dimensionality in low-dimensionality (known as the $n << p$ problem).

In this context, it was highlighted two methodologies proposed by Zhang et al. (2003) and Price et al. (2006) respectively to correct for population stratification. In the analytical equivalence between them, it was possible to conclude that correction for population-stratification in association studies can be done by either of these two alternatives. The Price methodology has however the advantage to be more computationally feasible when the dimension of the variables space (markers) is much larger than the dimension of the individuals space, i.e. $n << p$.

In spite of such results being well known in the singular value decomposition framework, the opportunity to further strengthen the analytical equivalences of both methodologies allows to emphasize the different solutions for the ancestry estimation. For instance, considering that the solution adopted by Price et al. (2006) corresponds to the classical solution of factor analysis via principal components, i.e., the scores of ancestry are eigenvectors coordinates standardized by their corresponding eigenvalues, one can obtain more accurate coefficients of ancestry imposing solutions of factor analysis via maximum likelihood methods. Alternatively, since Price et al. (2006) ancestry coefficients are also standard coordinates, i.e., unscaled singular vectors, there is still room to extend the ancestry estimation to robust standardization of the matrix $\mathbf{G}$. Following this rational, the EIGENSOFT package (Patterson et al., 2006) estimates of the global ancestry scores using measures of information content, such as allele frequency differences or FST statistics, to assign loadings for the molecular markers.

The duality between the row and column spaces of rectangular matrices was also discussed in order to introduce biplots in the analysis of genomic data. This graphic allows the representation of the ancestry scores of individuals and molecular markers (as SNPs) in the same axes, i.e., their joint visualization. To illustrate the usefulness of the biplot, genotype data from four populations of the HapMap Project (CEU, CHD, JPT and YRI) were analyzed. The results allowed us to discriminate SNPs associated with individuals from specific populations. Since it is of interest to identify markers that are more informative, establishing such associations may show relevant patterns on the distribution of SNPs in the worldwide populations.

Acknowledgments

This paper was partially supported by the National Council to Technological and Scientific Development (CNPq) and Foundation for Research Support of the State of São Paulo (FAPESP) (grant *no* 2012/000316-5).

DUARTE, N. E.; GIOLO, S. R.; de ANDRADE, M.; SOLER, J. M. P. Sobre a equivalência de métodos para populações estratificadas e sua aplicação em estudos de associação genética. *Rev. Bras. Biom.,* São Paulo, v.33, n.4, p.494-507, 2015.

▪ RESUMO: Estudos de associação em indivíduos não relacionados têm sido comumente usados no mapeamento de genes envolvidos na regulação de doenças complexas. No entanto, quando os indivíduos são de diferentes ancestralidades, estes estudos podem produzir associações espúrias devido à estratificação da população, resultando em excesso de resultados falsos positivos ou negativos. A análise de componentes principais da matriz de genótipos realizada com base nos marcadores genéticos (colunas da matriz) ou nos indivíduos (linhas da matriz) são as abordagens mais comuns utilizadas para corrigir a estratificação da população em estudos de associação genética. Neste artigo, resultados da teoria de decomposição em valores singulares de matrizes são usados para demonstrar a equivalência analítica entre estas duas abordagens e sua relevância para a correção de estratificação genética de populações. Além disso, é mostrada a importância de usar o biplot como uma ferramenta de visualização, que explica não somente a informação conjunta de amostras e marcadores genéticos mas também para detectar marcadores informativos. Embora ambos os procedimentos possam ser utilizados para corrigir para a estratificação populacional, o cálculo com base nos indivíduos é computacionalmente mais recomendável para a análise de conjuntos de marcadores genéticos de alta dimensão (problema $n << p$). Como aplicação são usados os dados genotípicos de quatro populações mundiais do Projeto HapMap.

▪ PALAVRAS-CHAVE: Componentes principais; decomposição em valores singulares; dualidade da matriz genética.

References

ALTSHULER, D.; DALY, M. J.; LANDER, E. S. Genetic mapping in human disease. *Computacional Statistics and Data Analysis*, v.322, p.881-888, 2008.

BUSH, W. S.; MOORE, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*, v.8, p.e1002822, 2012.

BAUCHET, M.; MCEVOY, B.; PEARSON, L. N.; QUILLEN, E. E.; SARKISIAN, T.; HOVHANNESYAN, K.; DEKA, R.; BRADLEY, D. G.; SHRIVER, M. D. Measuring European Population Stratification with Microarray Genotype Data. *The American Journal of Human Genetics*, v.80, p.948-956, 2007.

CHAE, S. S.; WARDE, W. D. Effect of using principal coordinates and principal components on retrieval of clusters. *Computacional Statistics and Data Analysis*, v.50, p.1407-1417, 2006.

GABRIEL, K. R. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, v.58, p.453-467, 1971.

GOWER, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, v.53, p.325-338, 1966.

GREENACRE, M. J. *Theory and application of Correspondence Analysis*. London: Academic Press, 1984.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall, 2007.794p.

PRICE, A. L.; PATTERSON, N. J.; PLENGE, R. M.; WEINBLATT, M. E.; SHADICK, N. A.; REICH, D. Principal component analysis corrects for stratication in genome-wide association studies. *Nature Genetics*, v. 38, p.904-909, 2006.

PATTERSON, N.; PRICE, A. L., REICH, D. Population structure and eigenanalysis. *PLoS Genetics*, v.2, e190, 2006.

RINGNÉR, M. What is principal component analysis?. *Nature Biotechnology*, v.26, p.303-304, 2006.

SAJESH, T. A.; SRINIVASAN, M. R. Detection of multidimensional outliers using biplot analysis. *MJS*, v.7, p.10-24, 2008.

The International HapMap Project. *Nature*, v.426, p.789-796, 2003.

TIWARI, H. K.; BARNHOLTZ-SLOAN, J.; WINEINGER, N.; PADILLA, M. A.; VAUGHAN, L. K.; ALLISON, D. B. Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. *Human Heredity*, v.66, p.67-86, 2008.

ZHANG, S.; ZHU, X.; ZHAO, H. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic Epidemiology*, v.24, p.44-56, 2003.

ZHENG, G.; YANG, Y.; ZHU, X.; ELSTON, R. C. *Analysis of Genetic Association Studies*. New York: Springer, 2012.

ZHU, X.; LI, S.; COOPER, R.; ELSTON, R. C. A unified association analysis approach for family and unrelated samples correcting for stratification. *The American Journal of Human Genetics*, v.82, p.352-365, 2008.

Received in 13.05.2015.

Approved after revised in 01.09.2015.