Published on *Plant Breeding E-Learning in Africa* (https://pbea.agron.iastate.edu)

Measures of Similarity

By William Beavis, Mark Newell (ISU)

Objectives

- Utilize coefficients of inbreeding and parentage to construct the numerator relationship matrix
- Utilize molecular marker information to construct a realized kinship matrix

Introduction

In an ideal reference breeding population, there is no structure consisting of sub-populations or aggregates of relatives organized into families and tribes. Plant Breeding populations, on the other hand, are organized into sub-populations. Perhaps the best known example is represented by the heterotic germplasm pools in maize, e.g., Stiff Stalks, Non-Stiff Stalks, Lancasters and Iodents. In cytoplasmic male sterile hybrid systems such as sorghum, the restoration pattern can be the primary divider of germplasm with additional subdivisions based on morphological characteristics and geographic origins, e.g., Kaoliang, Durra, and Feterita. Alternatively, coefficients of relationship and inbreeding among members of a breeding population can be used to represent the structure of the breeding population. Also, with the emergence of high throughput molecular marker technologies, it is possible to represent relationships among members of a breeding population using identity in state to produce a realized kinship matrix.

Population Structure Based on Pedigree Information

Animal breeders were the first to utilize relationships among individuals for purposes of providing Best Linear Unbiased Predictions in linear mixed models. The “A” matrix in the linear mixed model equation, also known as the Numerator Relationships Matrix (NRM) was originally used by Henderson to capture information from relatives to predict breeding values of animals. In essence, the A-matrix provides information on the proportion of alleles that are identical by descent between all pairs of individuals in a breeding population.

Specifically, the numerator relationships are equal to twice the coefficient of coancestry between any pair of individuals. In other words, $A_{x,y} = 2\Theta_{x,y}$. Thus, if we know the pedigrees of all members of a breeding population we can construct an A-matrix using a recursive tabular method.

Recursive Tabular Method

Recursive Tabular Method For Constructing The A-Matrix

1. Order members of a pedigree chronologically, i.e., list parents before offspring. Assume that founder lines are not inbred and are not related to each other.
2. Transpose the list and use this to represent columns for the A matrix.
3. Beginning with the cell represented by A1,1 compute Θ1,1.
4. Move to cell $A_{1,2}$ and compute $\Theta_{1,2}$. This will be the same value that can be used for cell $A_{2,1}$
5. Move to cell $A_{2,2}$ and compute $\Theta_{2,2}$.
6. Move to cell $A_{1,3}$ and compute $\Theta_{1,3}$. This will be the same value for $A_{3,1}$
7. Move to cell $A_{2,3}$ and compute $\Theta_{2,3}$. This will be the same value for $A_{3,2}$
8. Move to cell $A_{3,3}$ and compute $\Theta_{3,3}$
9. Repeat until all elements of the A matrix are completed.

Population Structure Based on Markers

The Realized Kinship Matrix

Consider two cultivars scored for 1400 SNPs. We can ask whether this pair of cultivars have the same or different alleles at each locus. Intuitively, if they had the same allele at all 1400 loci, we would say that there are no detectable allelic differences between the two genotypes, i.e., that they are identical in state or that their similarity index = 1.0. Alternatively, if none of the alleles are the same at all 1400 loci, then we would say that the genotypes have no alleles in common, i.e., that their similarity index is zero. In practice, the two genotypes will exhibit a measure of similarity somewhere between these extremes.

Quantitative Measure for Similarity

Let's take this intuition and develop a quantitative measure for similarity. If the two cultivars (x and y) have the same pair of alleles at a locus, score the locus = 2, if one of the alleles is the same, score the locus = 1, otherwise the score = 0. If we sum these up across all loci the maximum score would be 2800. If we divide the summed score by 2800 we would obtain a proportion measure (designated $s_{x,y}$) to quantify the similarity between the pair of lines. This concept can be represented algebraically as:

$$S_{x,y} = \frac{1}{2n} \sum_{i=1}^{n} X_i Y_i$$

Such a similarity measure could be converted into an "intuitive genetic distance" measure by subtracting $S_{x,y}$ from 1.

Measures of Distance

Our intuitive genetic distance would make sense if

1. there are only two alleles per locus,
2. our interpretation of the result does not include inferences about identity by descent, and
3. if there is no LD among the SNP loci.

However, most populations are more complex requiring more nuanced measures of genetic distance. Population geneticists tend to use three distance measures depending upon the inference about population structure they are trying to understand. These are:

- **Nei's Distance** assumes all loci have the same neutral rate of mutation, mutations are in equilibrium with genetic drift and the effective population size is stable. The interpretation is a measure of the average number of changes per locus and that differences are due to mutation and genetic drift.
- **Cavalli-Sforza's Distance** assumes differences are due to genetic drift between populations with no mutation and interprets the genetic distance as a Euclidean Distance metric.
- **Reynolds Distance** is applied to small populations, thus it assumes differences are due to genetic drift and is based on knowledge about coancestry, i.e., identity by descent for alleles that are the same.

Application of Distance and Similarity Measures

There are a large number of additional distance and similarity measures that can be applied to molecular marker scores including Euclidean, Mahalanobis, Manhattan, Chebyshev, and Goldstein. Also Bayesian Statistical approaches can be used to identify structure in the population (Pritchard et al, 2000) without resorting to calculation of distance metrics. The choice of an appropriate method depends upon the type of molecular marker data and the research question. A thorough presentation of distance measures is beyond the scope of this course, but there are graduate courses on multivariate statistics in which issues associated with each of the distance metrics can be explored.

For now, let's assume that we decided to use our $S_{x,y}$ to represent differences between all pairs (x_i, y_j) of breeding lines. Next, suppose we extend the example from two lines to 1800 lines scored for 1400 SNPs. In this case, there are $[n\times(n-1)]/2] = 1{,}619{,}100$ estimates of pairwise distances among the lines.

Clearly any attempt to find patterns in a data matrix consisting of all pairwise measures of similarity or distance will take considerable effort. Yet, these patterns in the data are essential to quantifying the structure in a breeding population, because the structure will affect inferences about genetic effects. It is the need to find patterns in such large data sets that motivated application of multivariate statistical methods such as principle components and cluster analyses in plant breeding populations.

Principal Component Analysis (1)

The primary purpose for applying principal component analysis (PCA) to genetic distance matrices is to summarize, i.e., reduce dimensionality, so that the underlying population structure can be visualized.

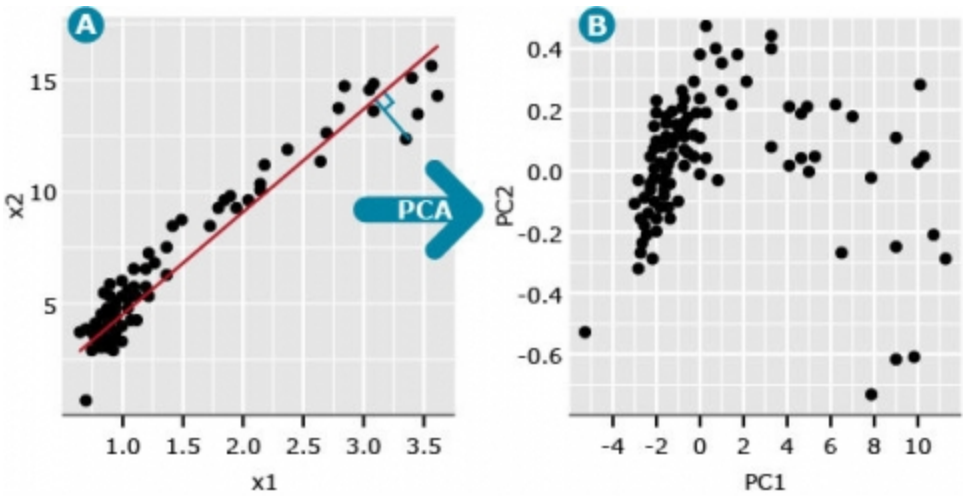

Fig. 1 Effect of principal component analysis.

Conceptual Interpretation

Imagine we have two variables, denoted x1 and x2, where x1 represents the distance scores between cultivar 1 and all other cultivars and x2 represent the distance scores between cultivar 2 and all other cultivars. If we plot the x1,x2 pairs of data we might generate a plot such as seen in Fig. 1A. We could add distance data for a third cultivar and represent the data with a 3 dimensional plot. We could obtain data for as many cultivars as we might have interest in, but the ability to plot these in multi-dimensional space is not possible.

Principal Component Analysis

The primary purpose for applying principal component analysis (PCA) to genetic distance matrices is to summarize, i.e., reduce dimensionality, so that the underlying population structure can be visualized.

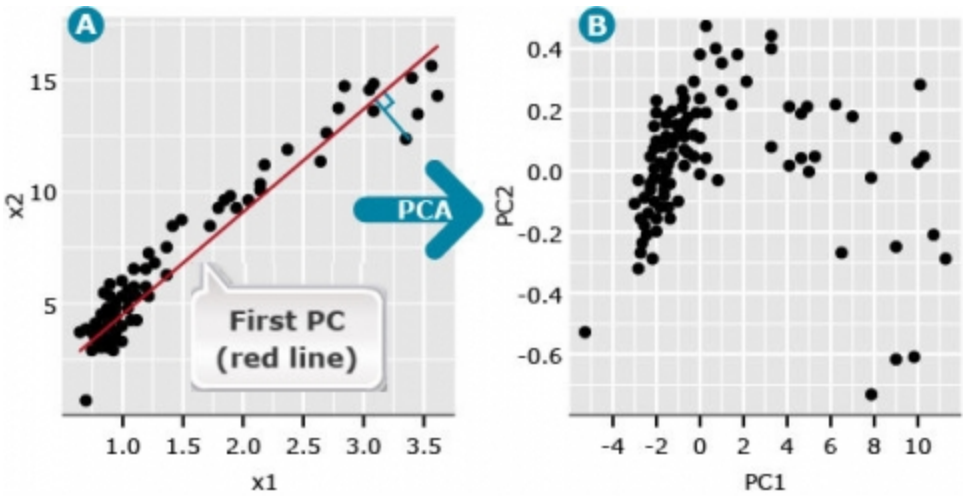

Fig. 2 Effect of principal component analysis.

Conceptual Interperpretation

We refer to the first principal component (PC), also known as the first eigenvector, as a line (red) that minimizes the perpendicular distances (blue line) between the red line and the data points (Fig. 2 A).

Principal Component Analysis - Interpretation

Conceptual Interpretation

The second PC follows the same definition except that it represents a line through the data that minimizes distance between a second line, that is orthogonal (at a right angle) to PC1. The second PC minimizes the distance between the data and the second line. Since the second PC is orthogonal to the first the distance among the data points represented by each PC is maximized. Thus we can plot data points represented by the first two principle components (Fig. 3.B). By plotting the PC's instead of the raw data we often find hidden structure in the data (compare Fig. 3.A vs. 3.B).

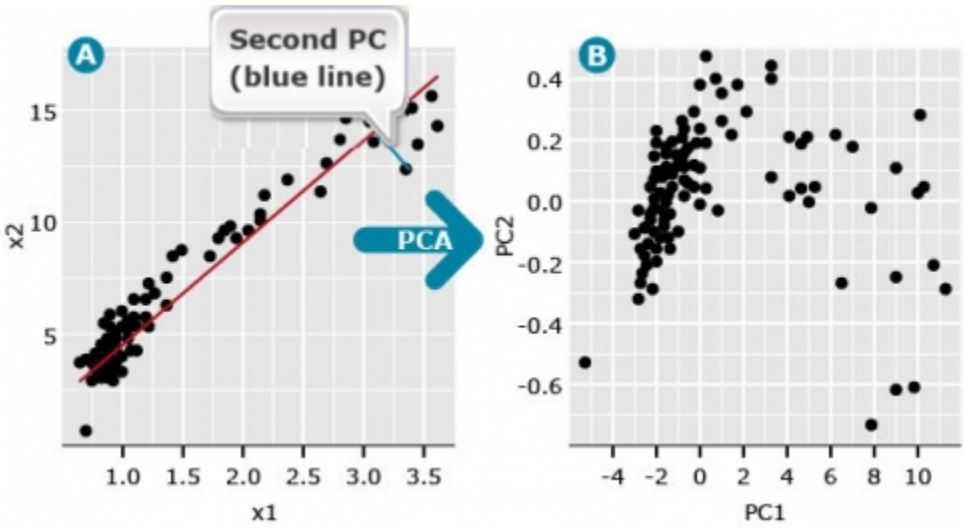

Fig. 3 Effect of principal component analysis.

Subsequent PCs represent lines that are orthogonal to all previous PCs and minimize distance between each PC and data points that maximize the variability among the orthogonal PCs. This means that each PC is uncorrelated to all other PCs.

A useful measure in PCA is the eigenvalue associated with each eigenvector (PC). The first eigenvalue is the proportion of maximum variability among the multidimensional data that is explained by the first PC. For the data depicted in Fig. 3.B, the first eigenvalue is 0.997 and the second eigenvalue is equal to 0.003. Since the first PC is the vector (or line) that is plotted in the direction of maximum variability among data points, the first eigenvalue is always the largest and each consecutive eigenvalue accounts for less variability than the prior PCs.

PCA Example

Let's consider an example from a set of 1816 barley lines scored for 1416 SNPs (Hamblin et al. 2010). In this analysis, there were

$$\binom{1816}{2})or(\frac{n \times (n-1)}{2} = 1,648,020$$

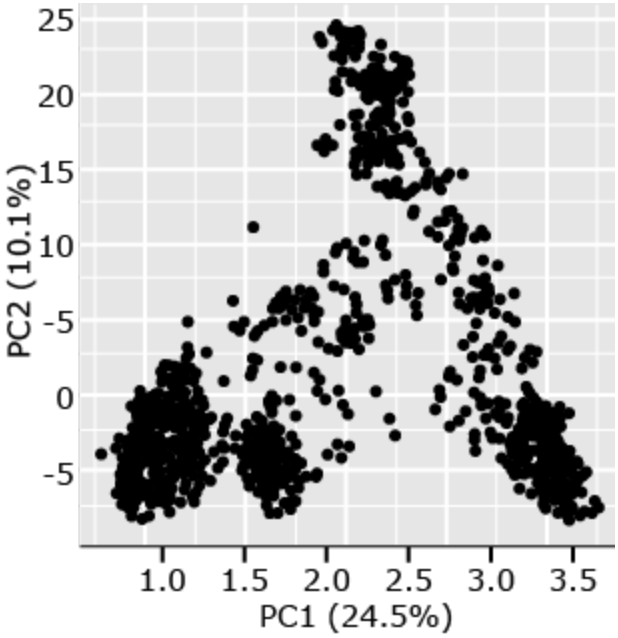

Fig. 4 Four distinct clusters produced by PCA.

estimates of pairwise distances based on 1416 SNP scores for each of the barley lines. Eigenvalues for PC1 and PC2 accounted for 24.5% and 10.1% of the variability among pairwise genotypic distances. By plotting PC1 versus PC2 (Fig. 2), we observe four distinct clusters. Subsequent analyses of the lines represented by each point in the clusters revealed that the members of each cluster are from 2-row, 6-row, spring, or winter barley types. From a breeding perspective, we can see that most breeding for barley occurs within types rather than between types. The population structure is a result of breeding processes of selection, drift and non-random mating.

Cluster Analysis

Similar to PCA, the purpose of applying cluster analysis to matrices of pairwise distance measures among a set of genotypes is to segregate the observations into distinct clusters. There are many types of cluster analyses, a primary distinction is between supervised and non-supervised clustering. K-means is one of the supervised methods that have been widely adopted by plant population geneticists. The clustering method is supervised in the sense that K represents a pre-determined number of clusters. Designating the number of clusters is usually based on prior knowledge about groups of lines that are being clustered. For example, it might make sense to designate the four clusters of barley lines based on known breeding history in which different barley agronomic types are not inter-mated. K-means represents an iterative procedure with the following steps:

i. An initial set number of K means (seed points) are determined (also called initialization); these are the initial means for each of K clusters.
ii. Each genotype is then assigned to the nearest cluster based on its pairwise distances to all other genotypes within and among clusters.
iii. Means for each cluster are then re-calculated and genotypes are re-assigned to the nearest cluster.
iv. Steps ii and iii are then repeated until no more changes occur.

Cluster Analysis Example

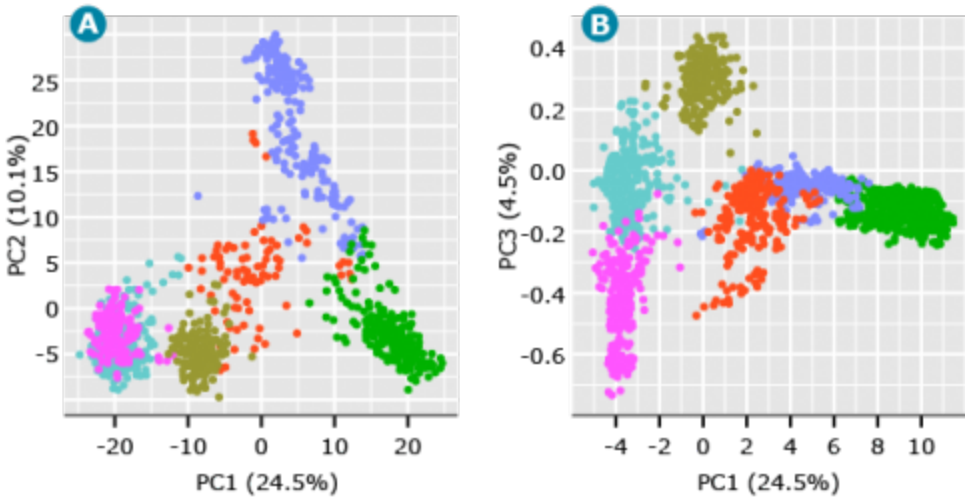

Fig. 5 PCA-produced k-means.

For the barley data, since the inter-mating rule is not absolute, i.e., some agronomic types are occasionally inter-mated, it could be informative to designate K = 6 (Fig. 5). Note that a plot of PC1 vs PC3 (Fig. 5.B) demonstrates the value of plotting PCs beyond the first two. While the third PC accounts for only 4.5% of the variability among genotypes, the third PC helps to distinguish what appear to members of the same cluster in Fig. 5.A.

Hierarchical Clustering

An unsupervised approach to clustering genotypic distance data is hierarchical clustering. This approach sequentially lumps or splits observations to make clusters. Applying the hierarchical approach to the barley data set we can visualize the results using a dendrogram (Fig. 6). In the dendrogram observations are arrayed along the x-axis and the y-axis refers to the average genetic distance between breakpoints. For example, the horizontal line at 4e+05 indicates that there are two major groups with a distance between them of 4e+05. The user determines the height (distance along the y-axis) at which a horizontal line is drawn and the number of clusters is chosen, this is drawn below in red for 6 clusters. The user may determine this by using the PC plots, cluster dendrogram, and any prior information that is known about the germplasm.

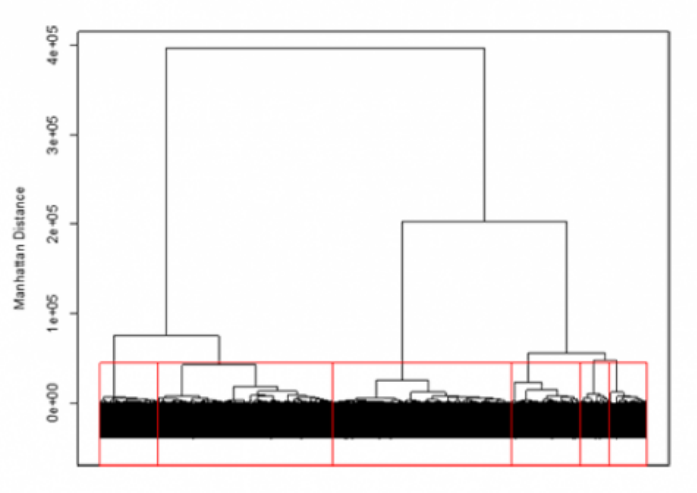

Fig. 6 Dendogram observations data

Hierarchical clustering can be implemented in many different ways. For genotypic data, the most common method is Ward's, which attempts to minimize the variance within clusters and maximize the variance between clusters. Similar to K-means clustering, we can look at the PC plots to explore the results for hierarchical clustering to see how the lines were assigned to clusters.

Acknowledgements

This module was developed as part of the Bill & Melinda Gates Foundation Contract No. 24576 for Plant Breeding E-Learning in Africa.

Quantitative Genetics Measures of Similarity Author: William Beavis, and Mark Newell (ISU)

Multimedia Developers: Gretchen Anderson, Todd Hartnell, and Andy Rohrback (ISU)

How to cite this module: Beavis, W. and M. Newell. 2016. Measures of Similarity. *In* Quantitative Genetics, interactive e-learning courseware. Plant Breeding E-Learning in Africa. Retrieved from **https://pbea.agron.iastate.edu**.