



LMSE Workshop 2018

Population structure

Reporter: Qinwen Xue

2018.1.9



Content

- Concept
- Sturcture
- PCA
- AMOVA



Concept of population structure

- A population may have substructure – differences in genetic variation among its constituent parts.
- low gene flow, genetic drift, selection, and mutation lead to genetic variation in subpopulation



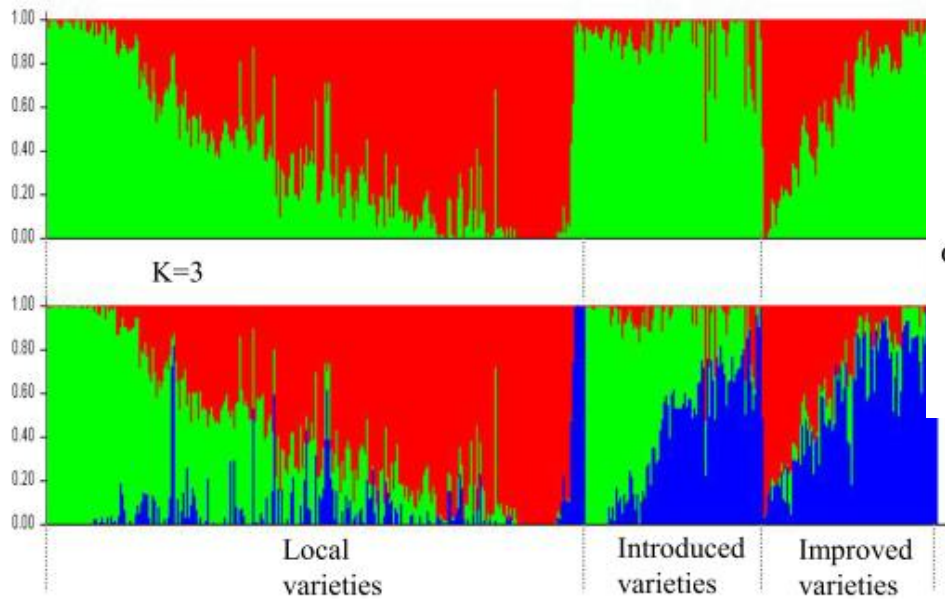
Characteristics

- The number of subpopulations within it.
- The frequencies of different genetic variants (alleles) in each subpopulation.
- The degree of genetic isolation of the subpopulations.

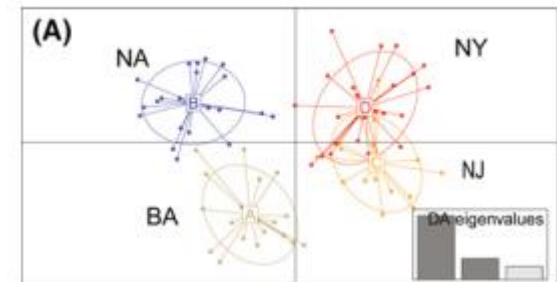
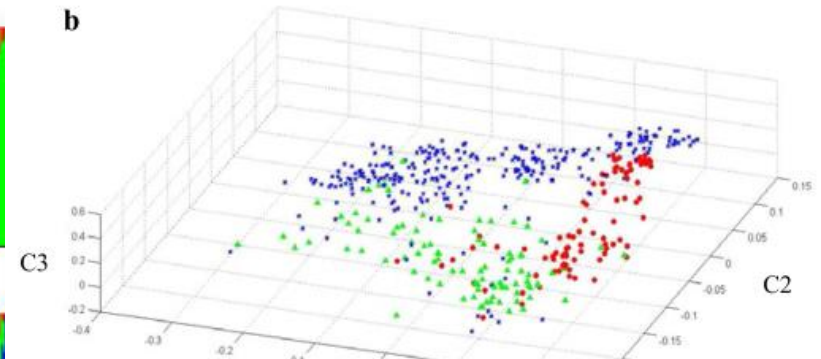


Method in populaton structure

Model-based clustering method



Distance-based (Principal components)





structure

- A model-based clustering method (Pritchard et al. 2000)
- Bayesian approach (MCMC: Markov Chain Monte Carlo)
- Detects the underlying genetic population among a set of individuals genotyped at multiple markers.
- Computes the proportion of the genome of an individual originating from each inferred population.



Concept of structure

Within population

1. Hardy-Weinberg equilibrium

$$(p + q)^2 = p^2 + 2pq + q^2 = 1$$

$$A = p, a = q$$

$$AA, Aa, aa$$

2. linkage equilibrium

The random association of alleles at different loci



Parameter

- Ancestry Models
- Allele frequency models
- Running length



Ancestry Models

no admixture model

Admixture model



Linkage model

Model with informative priors

Allele frequency models

Correlated allele frequencies



Independent model



Ancestor model

1. No admixture model. each individual is assumed to have originated in a single population
2. Admixture model. Individuals may have mixed ancestry, is a common feature of real data
3. Linkage model. This is essentially a generalization of the admixture model to deal with “admixture linkage disequilibrium”(genetic map)
4. Using prior population information. there is often additional information that might be relevant to the clustering



Allele frequency

- 1. Correlated allele frequencies:
Frequencies in the different populations are likely to be similar(migration or share ancestry)
- 2. Independent model:
Allele frequencies in different populations to be reasonably different from each other(improves clustering for closely related populations but over estimate k)
- Same population = correlated model,

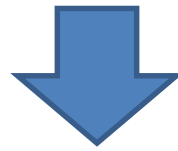


Running length

- length of burn-in period and Number of MCMC Reps



(1) Burnin length: how long to run the simulation before collecting data to minimize the effect of the starting configuration



Typically a burn-in is 10000—100000

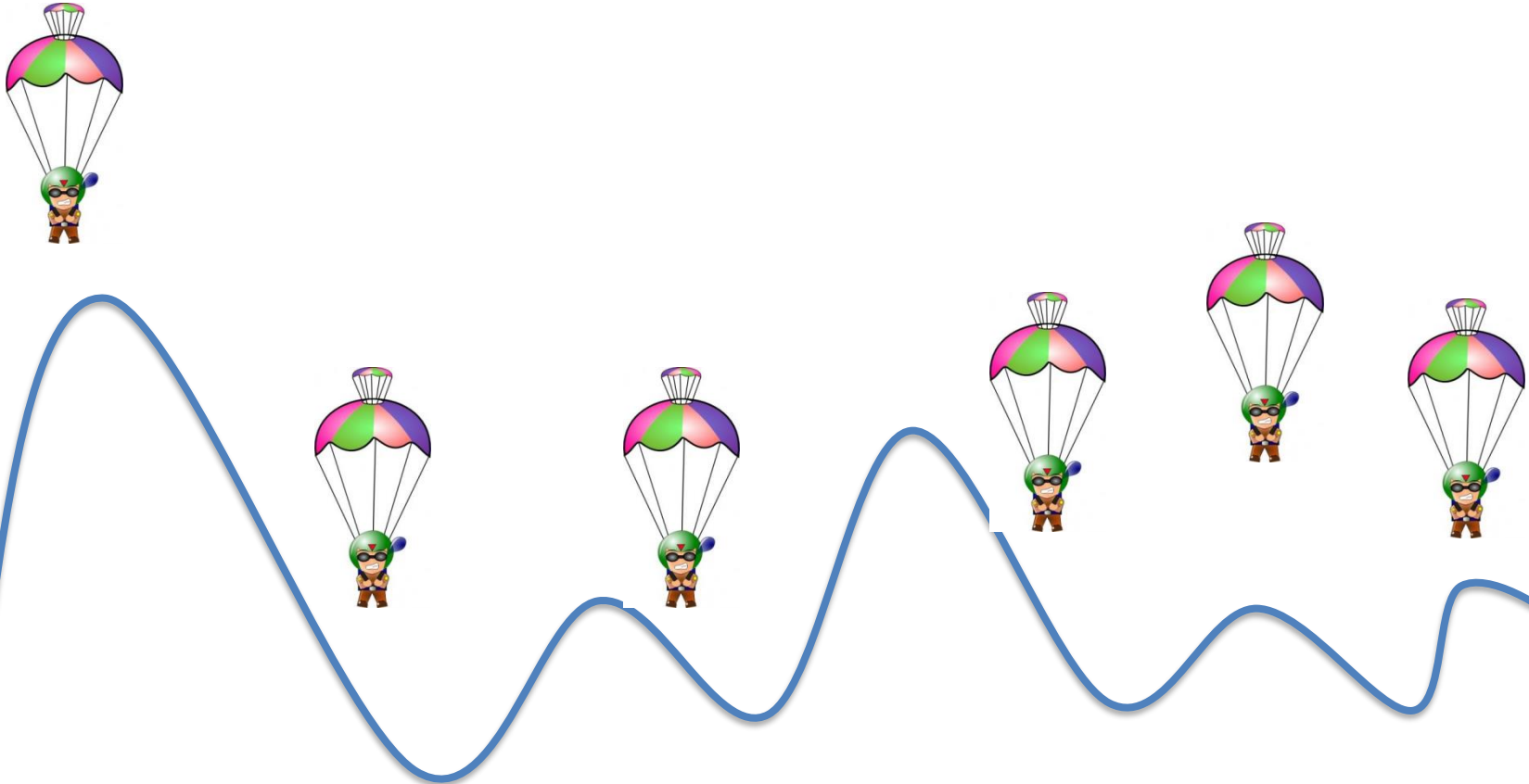


(2) How long to run the simulation after the burnin to get accurate parameter estimates.

Several runs at each K and whether you get consistent answers, 10000-1000000.

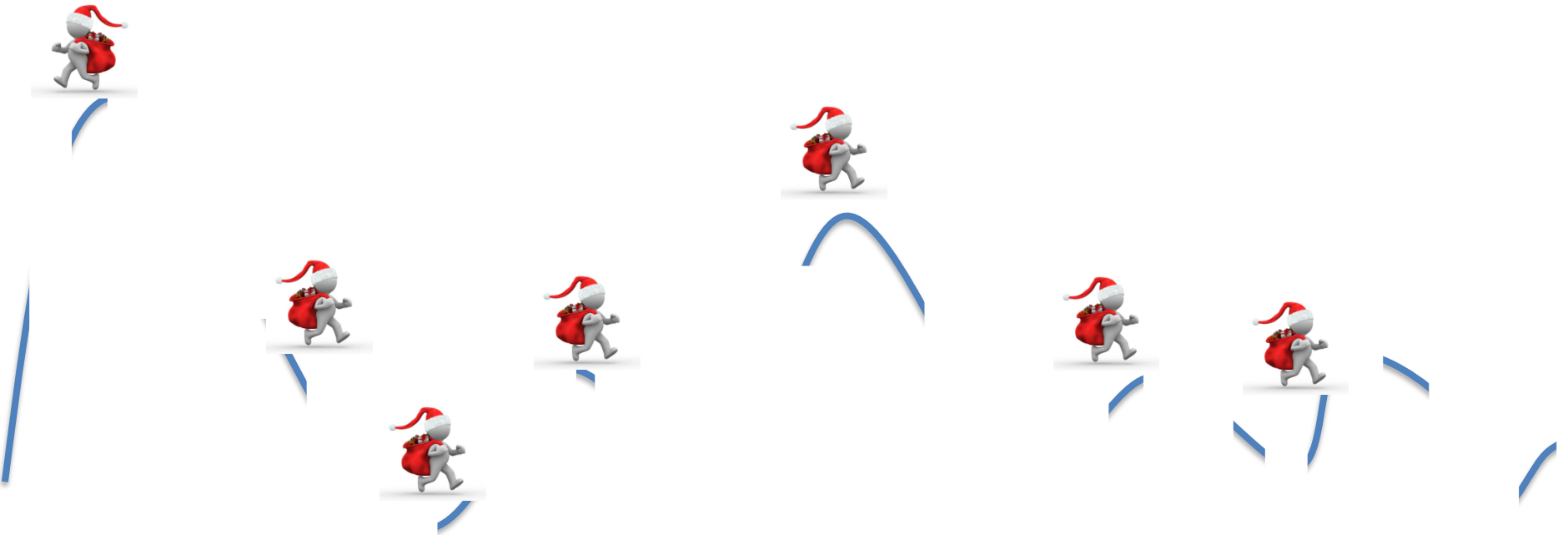


burn-in period



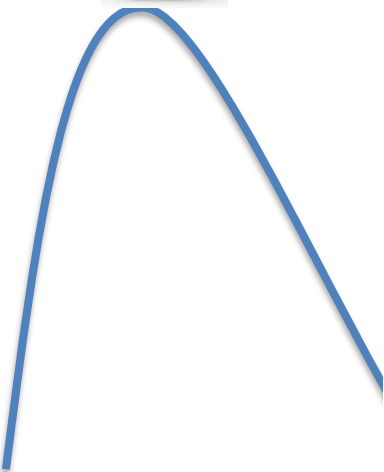


burn-in period





Number of MCMC Reps





Example

File -> new project -> name the project + select directory + choose data file -> number of individual (13) + ploidy of data (2) + number of loci (2444) + missing data value (-9) -> skip -> choose individual ID for each individual -> ProceedParameter set -> new -> length of burnin period (50000) + number of mcmc reps after burnin (500000) -> name parameterProject -> start a job -> set k from m to n (i.e. 2-4) + number of Iterations (3 or 5) -> startCompress the result and upload to Structure Harvester.File -> load structure results -> Browse (choose your result with the best value of K)-> Bar Plot -> show -> sort by Q -> save

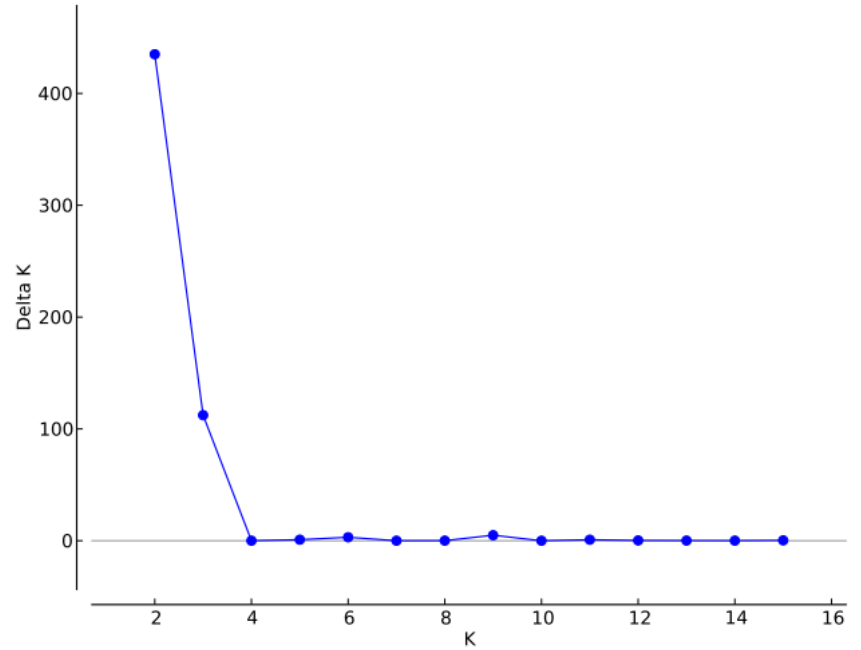


- the result and upload to Structure
- Harvester.File -> load structure results -> Browse (choose your result with the best value of K)-> Bar Plot -> show -> sort by Q -> save

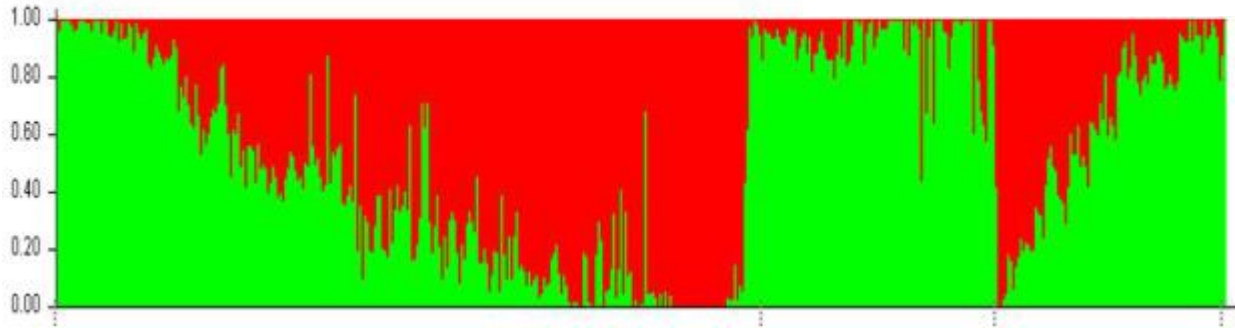


Figure S3 Graphical method (as in Evanno et al., 2005) allowing the detection of the number of groups K using ΔK

$$\Delta K = \text{mean}(|L''(K)|) / \text{sd}(L(K))$$



K=2





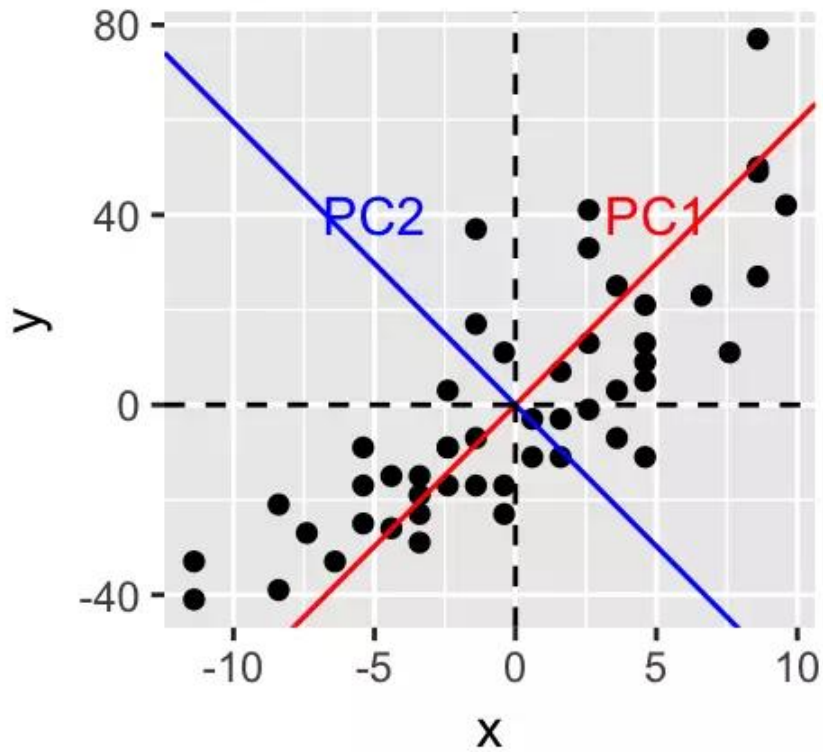
PCA

- PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components(from wikipedia)

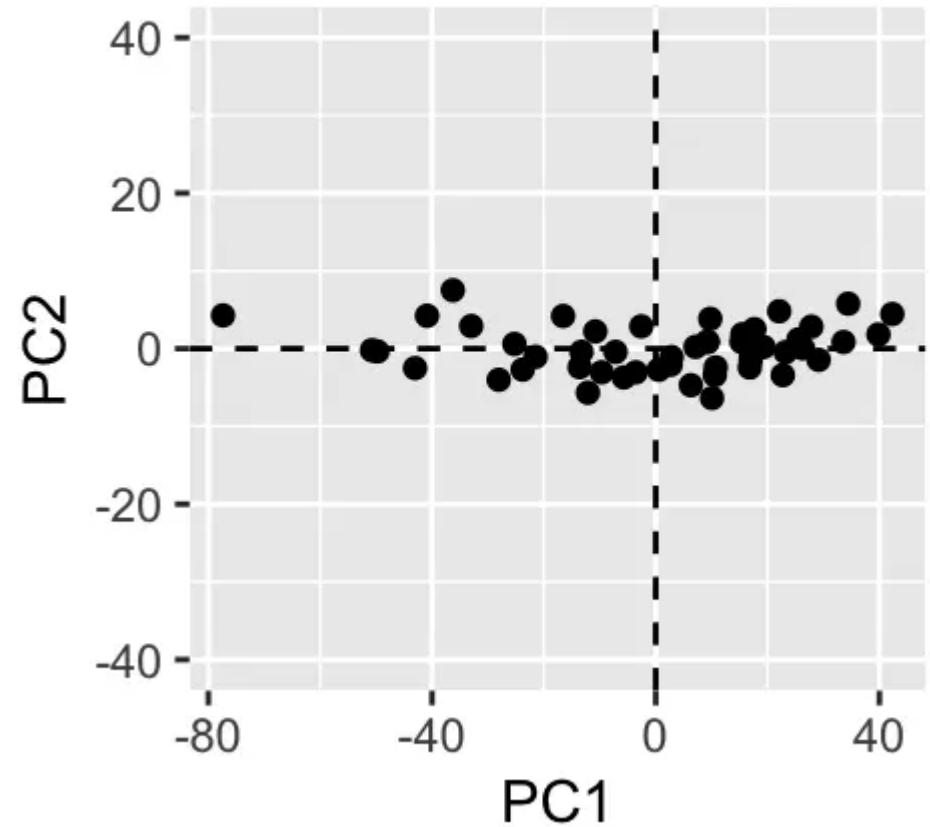


Concept of PCA

Plot 1A

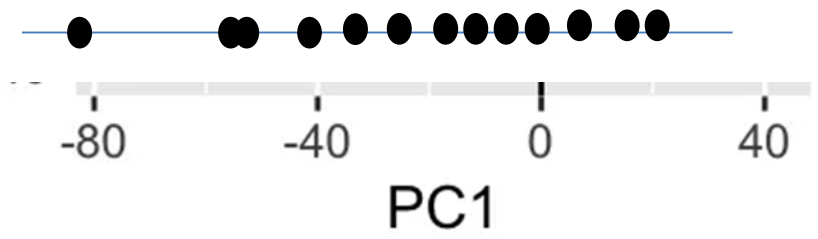
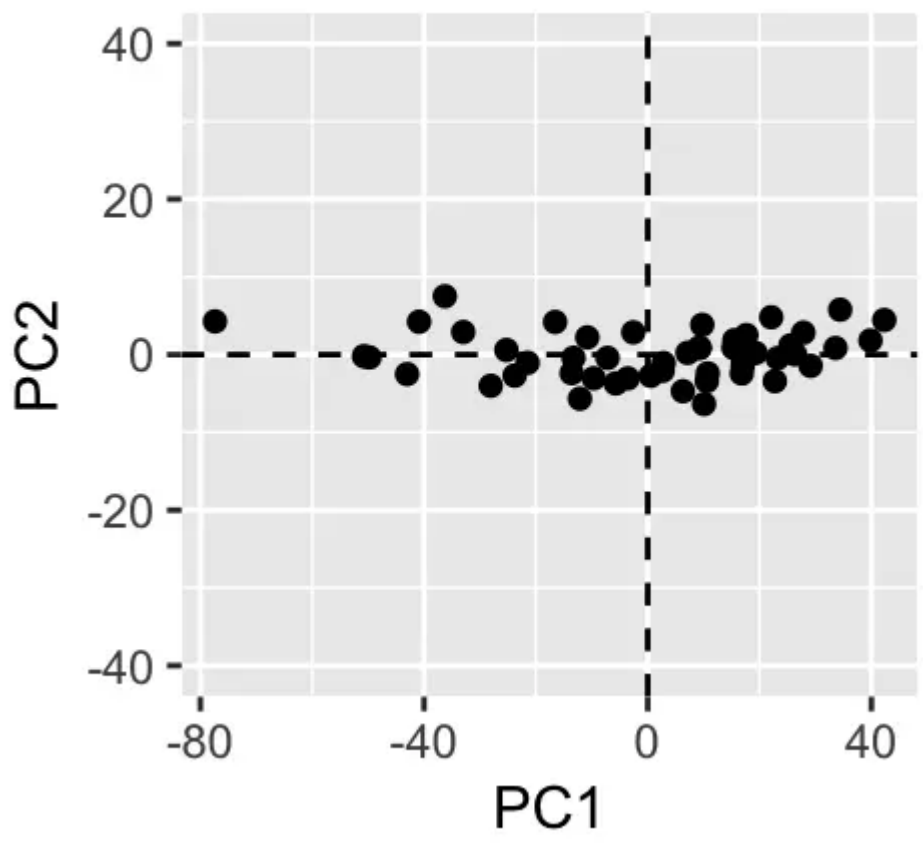


Plot 1B



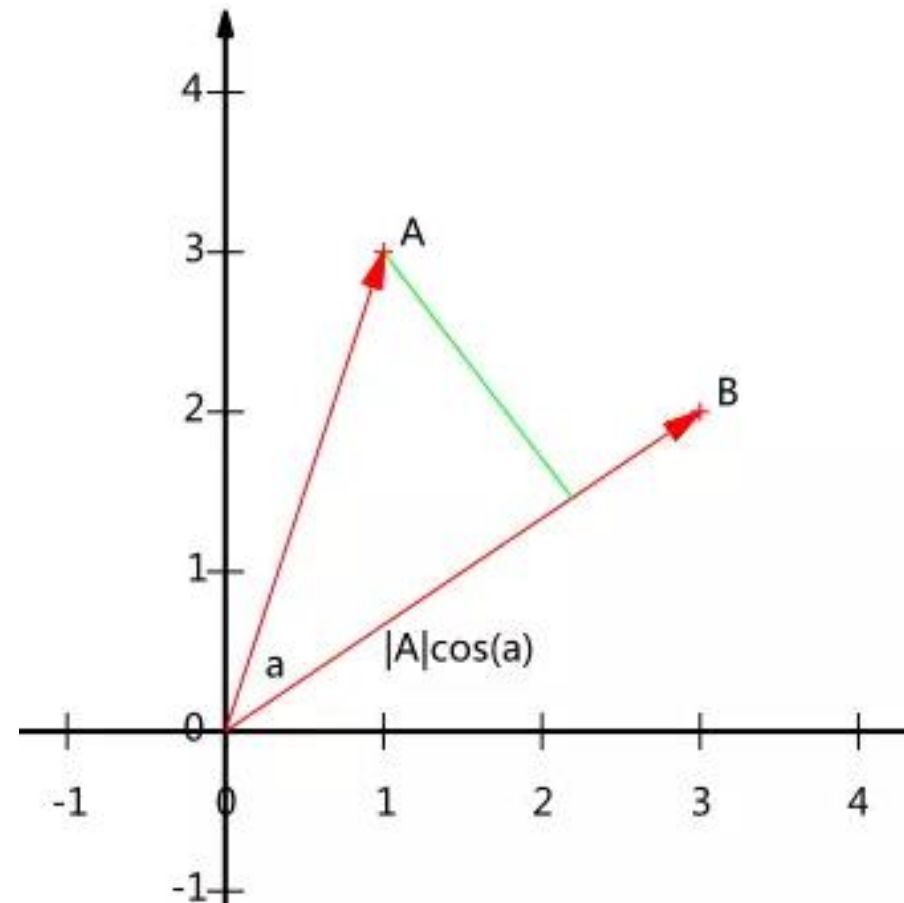


Plot 1B





- $A(a_1, a_2) * B(b_1, b_2) = a_1 * a_2 + b_1 * b_2$
- $A * B = |A| |B| \cos a$
- Let $|B| = 1$, then
- $A * B = |A| \cos a$

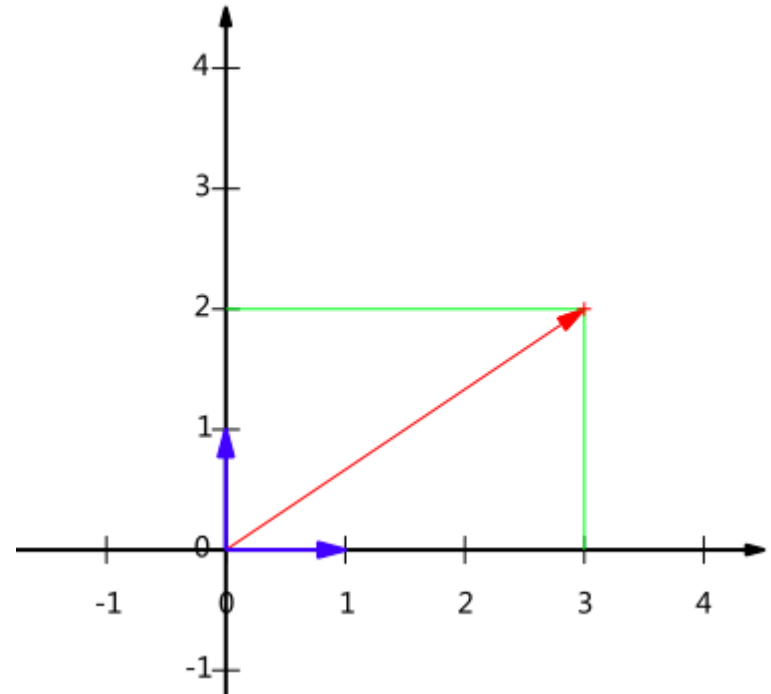




- Vector $(3,2)$ represent a vector which mapped in X vector is 3, in Y vector is 2.

Vector $(3,2)$ is seen as

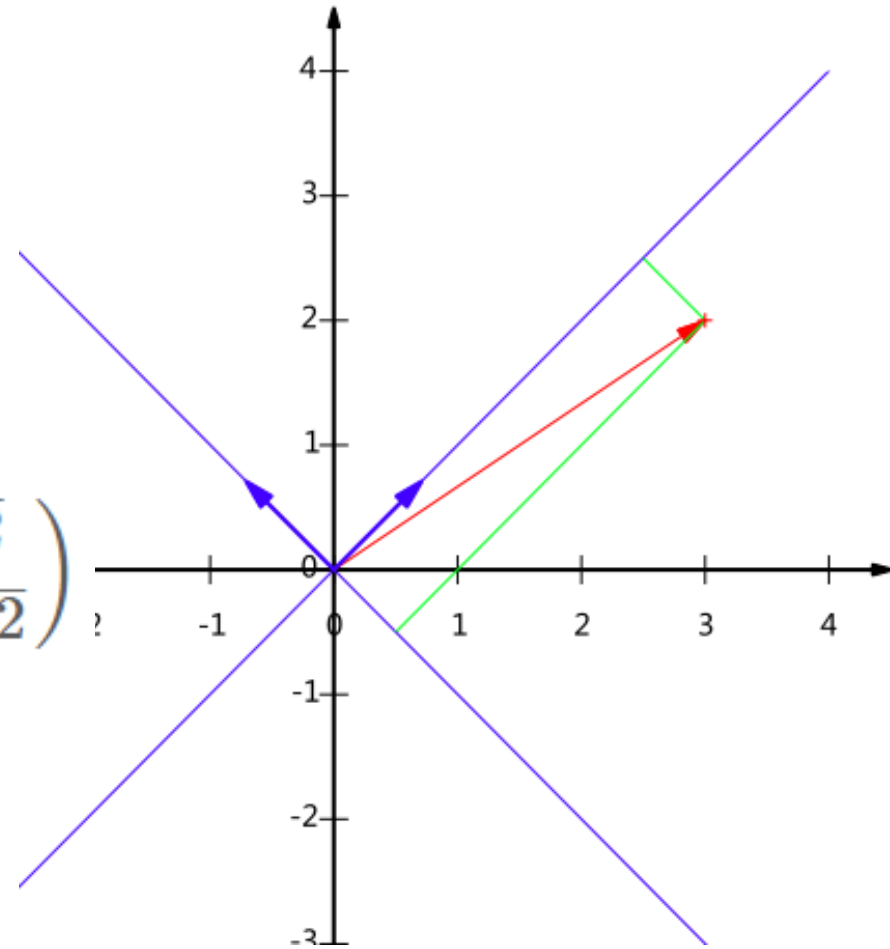
$$x(1, 0)^T + y(0, 1)^T$$





- If we transfer $(3, 2)$ to a new coordinate axis, we can use

$$\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 5/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$





- If we have three vector $(1,1)$, $(2,2)$, $(3,3)$, transfer them to the new coordinate axis

$$\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} = \begin{pmatrix} 2/\sqrt{2} & 4/\sqrt{2} & 6/\sqrt{2} \\ 0 & 0 & 0 \end{pmatrix}$$



- So if we have M n -dimensional vectors , transfer them to a new n -dimensional space with R n -dimensional coordinate axis .

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_R \end{pmatrix} (a_1 \quad a_2 \quad \cdots \quad a_M) = \begin{pmatrix} p_1 a_1 & p_1 a_2 & \cdots & p_1 a_M \\ p_2 a_1 & p_2 a_2 & \cdots & p_2 a_M \\ \vdots & \vdots & \ddots & \vdots \\ p_R a_1 & p_R a_2 & \cdots & p_R a_M \end{pmatrix}$$

$M: (p_1 a_M, p_2 a_M, \dots, p_R a_M)$



Example

- We have $(1,1)$, $(1,3)$, $(2,3)$, $(4,4)$, $(2,4)$

$$\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix}$$

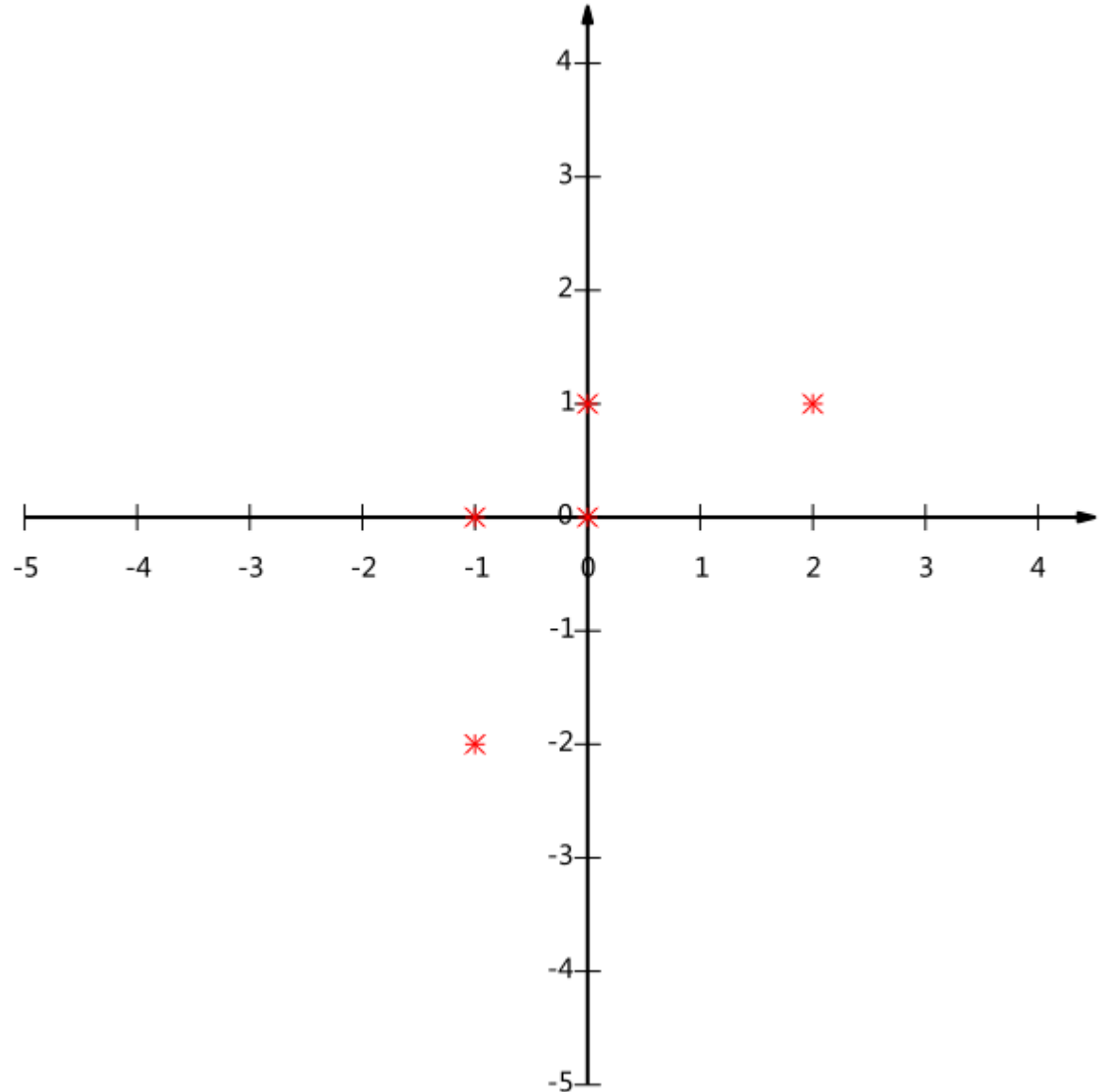


Reduce $\bar{\mathbf{x}}$

$$\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

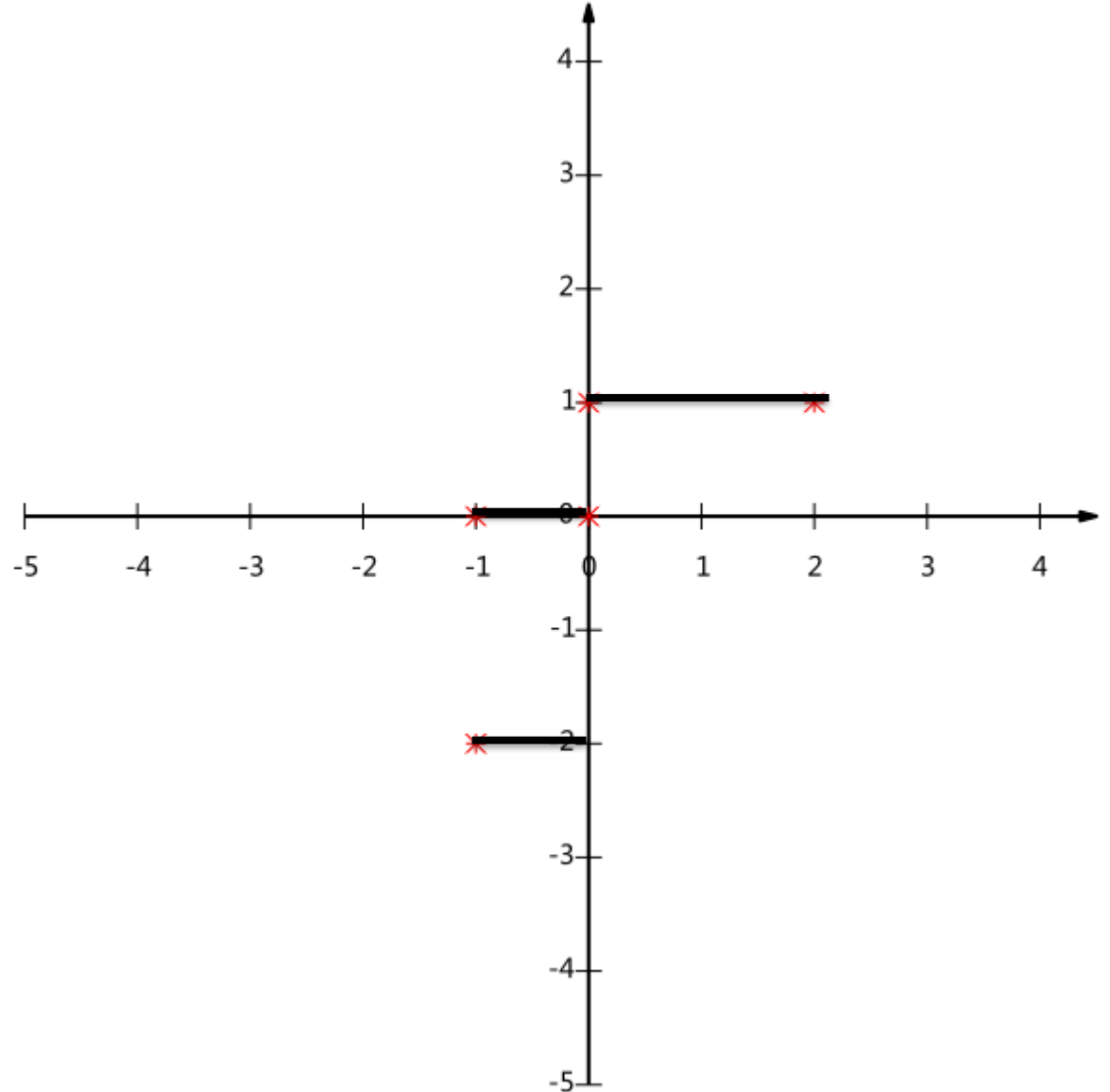


If we want
to use 1-
dimensional
to represent
the data....



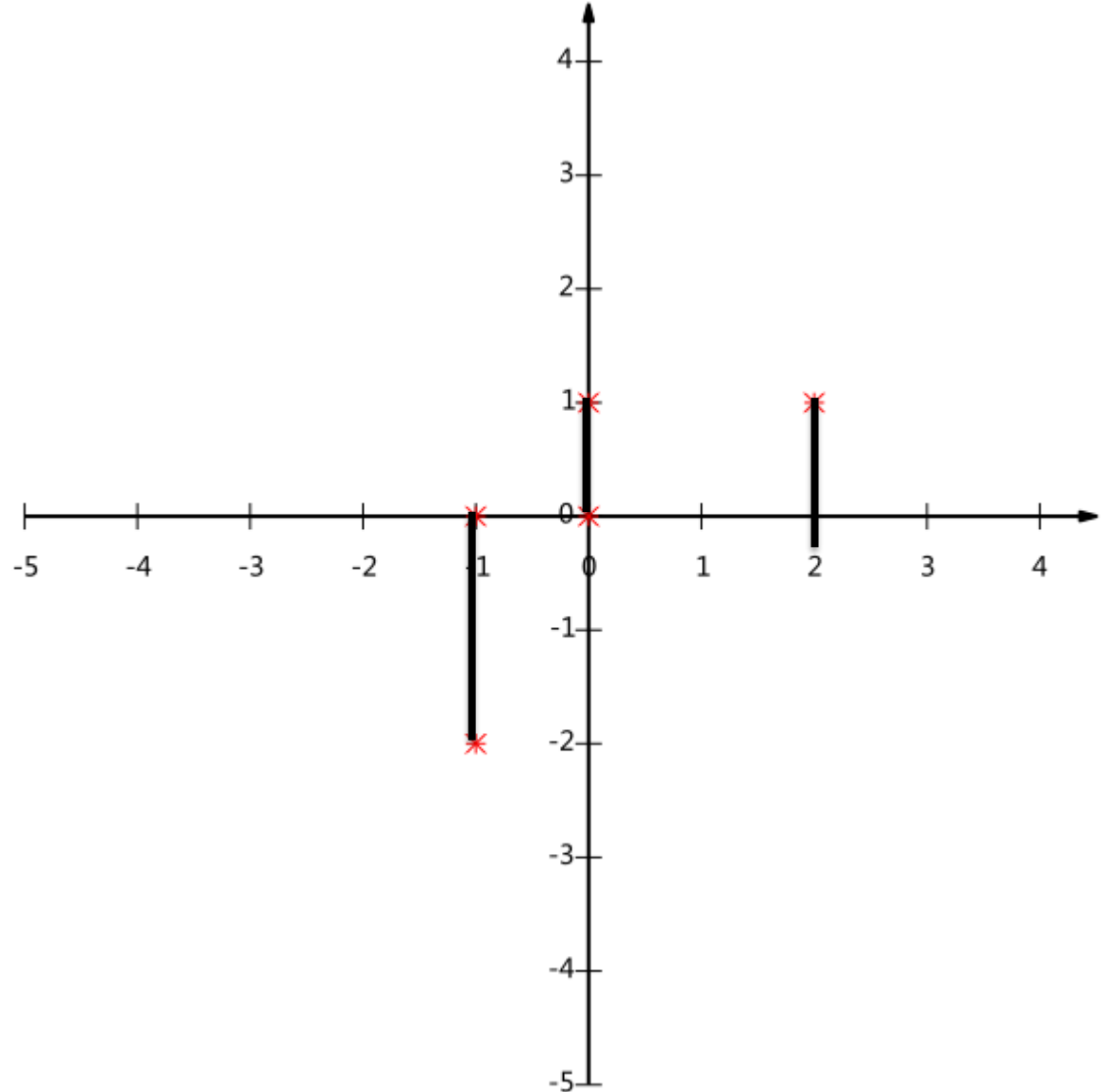


If we want
to use 1-
dimensional
to represent
the data....





If we want
to use 1-
dimensional
to represent
the data....



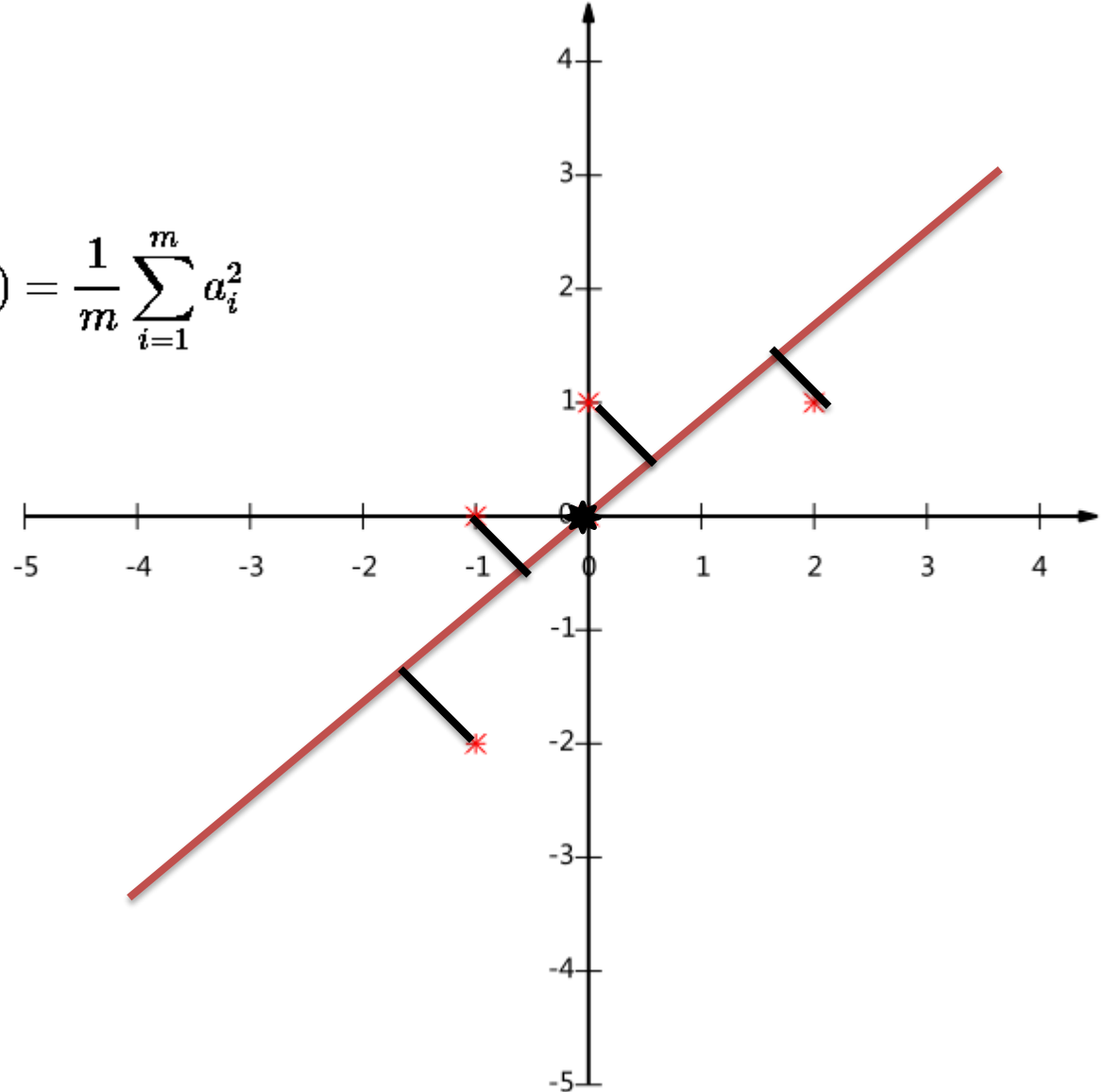


If we want to use 1-dimensional to represent the data....

$$Var(a) = \frac{1}{m} \sum_{i=1}^m a_i^2$$

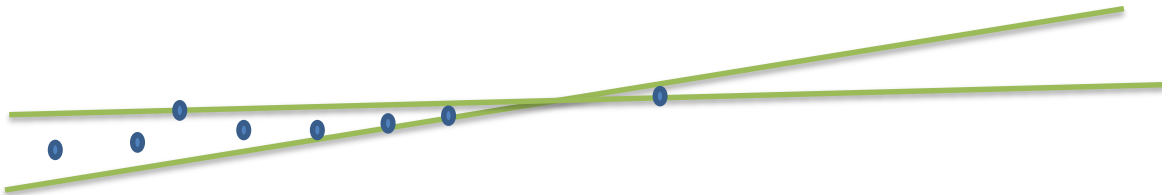
So, the data distribute more dispersive, can represent well.

In mathematic, we can use variance. Variance more large, data more representative.





- If we want to use 2-dimensional form to represent the 3-dimensional data.....
- Firstly, we choose the largest variance vector to be a coordinate axis.
- If we also choose the largest variance to be another coordinate axis...
- It will give us a repetitive information.





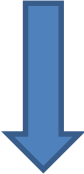
- So the best way is to find the unrelated vector, like the vertical vector.
- In probability theory and statistics, covariance is a measure of how much two random variables change together.
- If two variance both are independent , covariance of them is zero.

$$\text{Cov}(a, b) = \frac{1}{m} \sum_{i=1}^m a_i b_i$$



- How to find the covariance equal to zero but variance is the largest?



$$X = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{pmatrix} \times X^T = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \cdots & \cdots \\ a_m & b_m \end{bmatrix}$$


$$\frac{1}{m} X X^T = \begin{pmatrix} \frac{1}{m} \sum_{i=1}^m a_i^2 & \frac{1}{m} \sum_{i=1}^m a_i b_i \\ \frac{1}{m} \sum_{i=1}^m a_i b_i & \frac{1}{m} \sum_{i=1}^m b_i^2 \end{pmatrix}$$



- If we have M N -dimensional data, we can arrange them to be a $M \times N$ matrix X . And then set matrix $C = \frac{1}{m} XX^T$.
- The diagonals in the C yield variance and the i th row j th line value as same as j th row i th line value to yield covariance of vector i and vector j .
- And then we get the eigenvalues and eigenvectors from C . we choose K eigenvectors whic eigenvalues from largest , if we need K -dimensional data, to arrange a matrix P .
- Finally , $Y = PX$



Example

- `dudi.pca`



AMOVA

- Analysis of Molecular Variance (AMOVA) is a method of estimating population differentiation directly from molecular data and testing hypotheses about such differentiation.
- A variety of molecular data – molecular marker data (for example, RFLP or AFLP), direct sequence data, or phylogenetic trees based on such molecular data – may be analyzed using this method (Excoffier, et al. 1992).



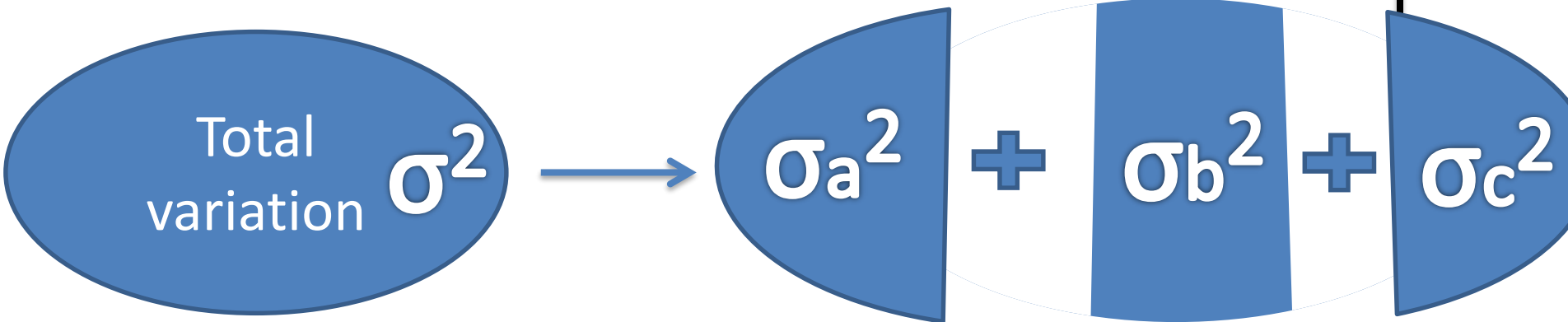
Concept of AMOVA

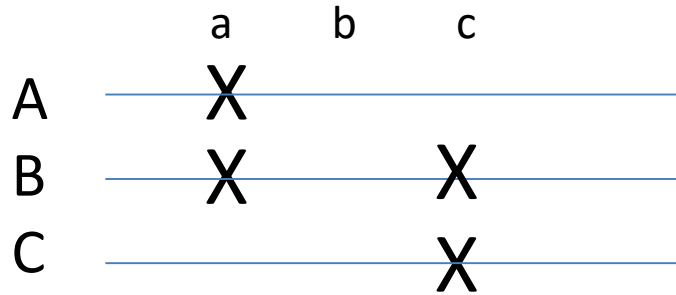
- $X_{jig} = X + a_k + b_{jk} + a_{ijk}$

Variation Among
group

Variation within a
group among demes

Variation
within deme



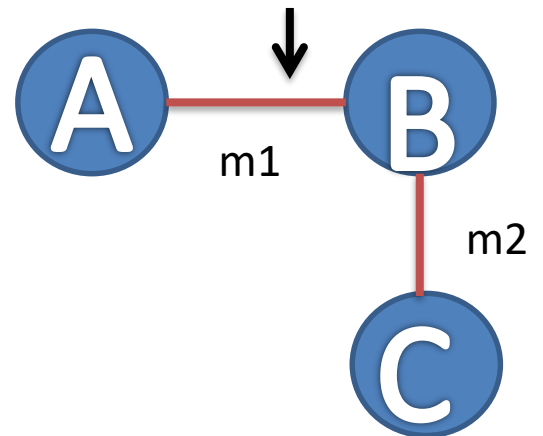


	a	b	c
A,	(1,	0,	0)
B,	(1,	0,	1)
C,	(0,	0,	1)

Boolean vector



A mutational event



	m1	m2
A,	(1,	0)
B,	(0,	0)
C,	(0,	1)



- Squared Euclidean distances are calculated for all pairwise arrangements of Boolean vectors, which are then arranged into a matrix, and partitioned into submatrices corresponding to subdivisions within the population (Excoffier, et al. 1992)

$$\mathbf{D}^2 = \begin{bmatrix} \begin{bmatrix} [\mathbf{D}_{11}^2] & [\mathbf{D}_{12}^2] \end{bmatrix} & \dots & [\mathbf{D}_{1I}^2] \\ \begin{bmatrix} [\mathbf{D}_{21}^2] & [\mathbf{D}_{22}^2] \end{bmatrix} & \dots & [\mathbf{D}_{2I}^2] \\ \dots & \dots & \dots \\ [\mathbf{D}_{I1}^2] & \dots & [\mathbf{D}_{II}^2] \end{bmatrix},$$



- The sums of the diagonals in the matrix and submatrices yield sums of squares (SS) for the various hierarchical levels of the population.
- The sum of squares of submatrices is the sum of squared (SSD) deviations to describe the differentiation of population.



- How do we get the effect of variance in different levels? (covariance components)

3.2.1.1 Haplotypic data, one group of populations

Source of variation	Degrees of freedom	Sum of squares (SSD)	Expected mean squares
Among Populations	$P - 1$	$SSD(AP)$	$n\sigma_a^2 + \sigma_b^2$
Within Populations	$N - P$	$SSD(WP)$	σ_b^2
Total	$N - 1$	$SSD(T)$	σ_T^2



- The variance components can be used to calculate a series of statistics called phi-statistics (ϕ), which summarize the degree of differentiation between population divisions and are analogous to F-statistics. ϕ -statistics are derived as follows (Excoffier, et al. 1992; Excoffier 2001):

Level of population hierarchy

Φ -statistic

Among demes within group

$$\Phi_{SG} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2}$$

Among groups within population

$$\Phi_{GT} = \frac{\sigma_a^2}{\sigma^2}$$

Among demes within population

$$\Phi_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma^2}$$



Thanks!!