Population genetics 1: exploratory analyses

Fernando Racimo

Copenhagen, August 2018

Today

- Exploratory vs. hypothesis-driven analyses
- PCA
- Latent mixed-membership models (“Structure”)

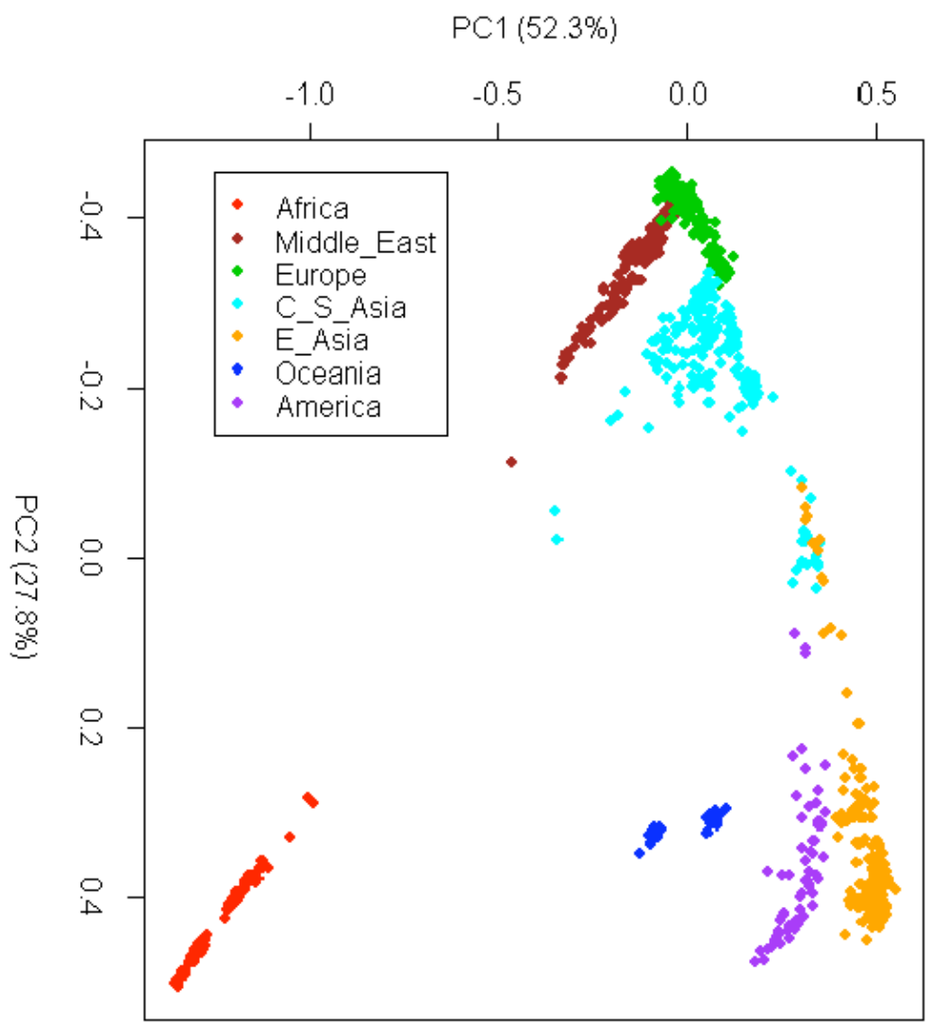
Today

- **Exploratory vs. hypothesis-driven analyses**
- PCA
- Latent mixed-membership models (“Structure”)

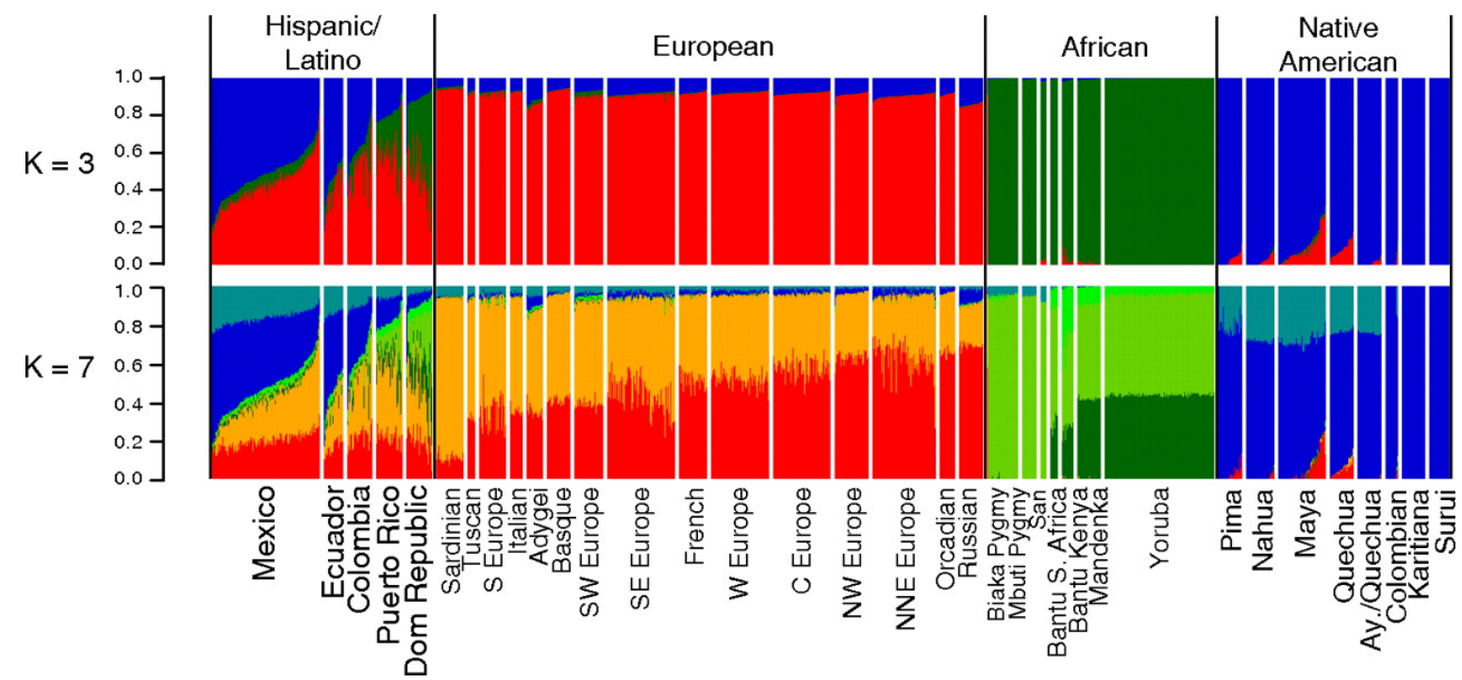
Exploratory analyses

- When we've just gotten some population genomic data (ancient or modern) and don't know where to start with it.
- What are the general patterns of variation? How much structure is there in my data?
- Which groups can be clustered together? Which groups are best modeled as a mixture of other groups?
- Are certain samples particularly interesting?

Exploratory analyses



PCA

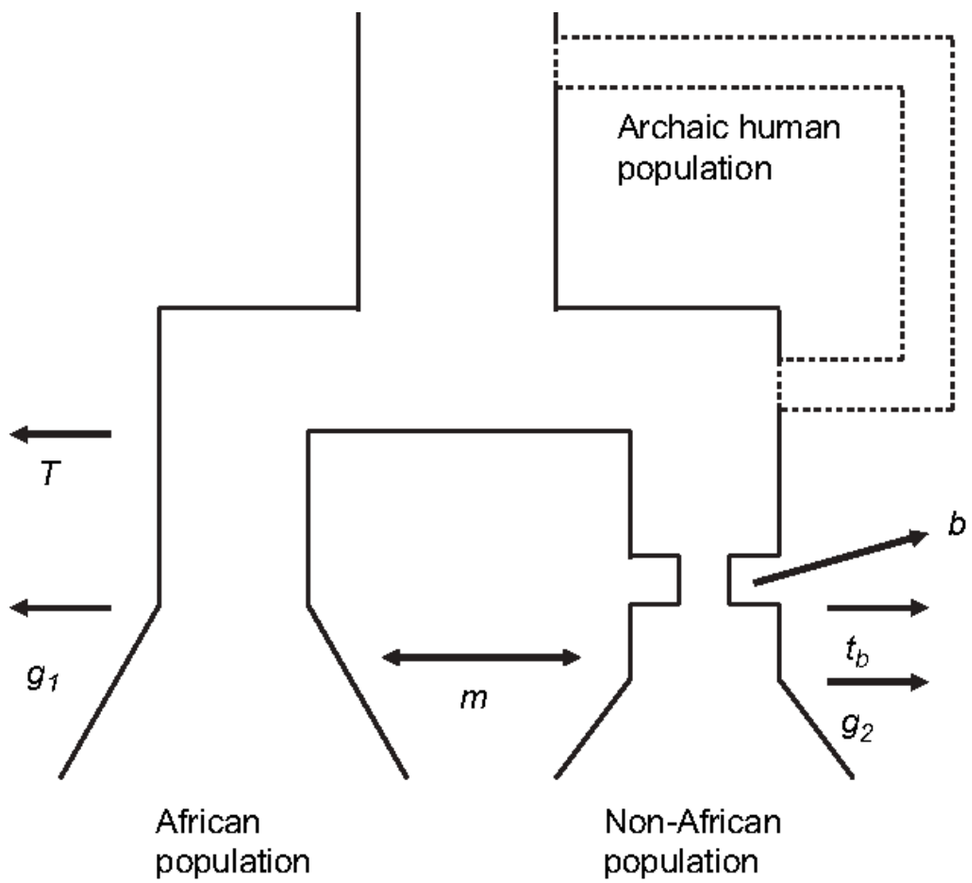


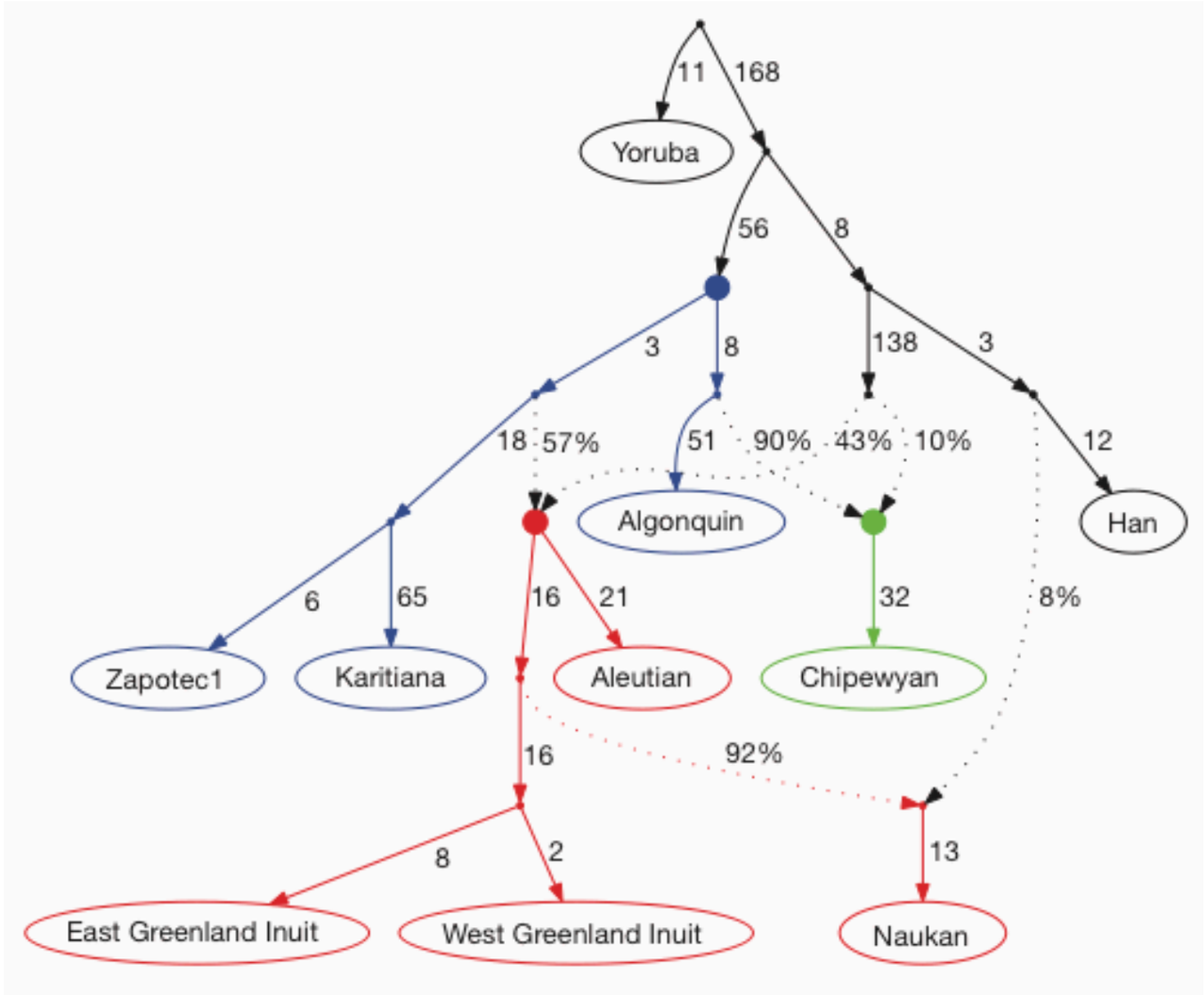
Latent mixed-membership models
(“Structure”)

Hypothesis-driven analyses & parameter estimation

- When we want to start building models of population history and testing particular hypotheses about the past.
- Is a particular population the result of an admixture event? What are the admixture proportions? When did the event happen?
- When did two populations diverge? When did a population contract or expand?
- What is the best history (or set of histories) that can best describe my data?

Hypothesis-driven analyses & parameter estimation





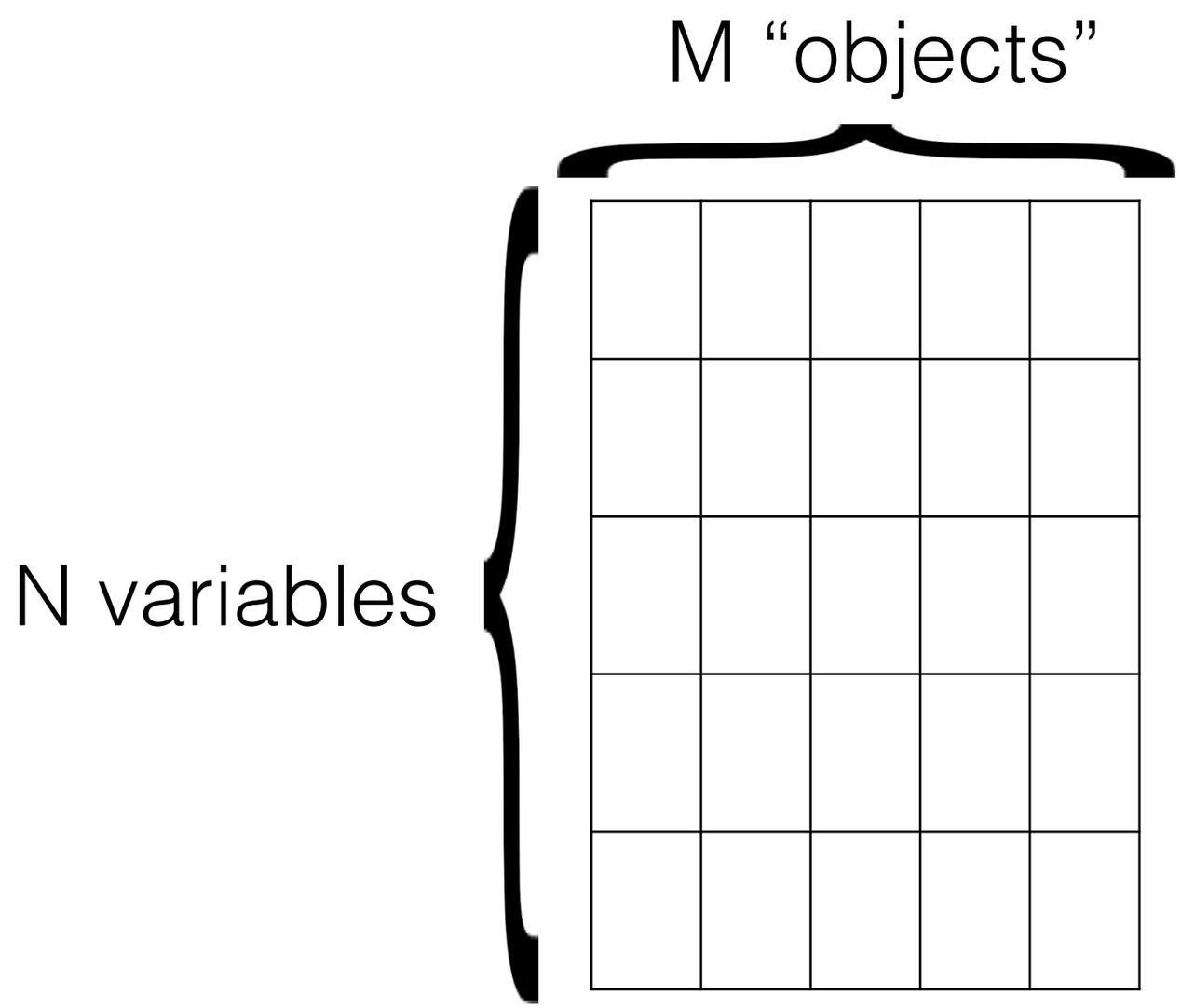
Today

- Exploratory vs. hypothesis-driven analyses
- **PCA**
- Latent mixed-membership models (“Structure”)

PCA

- Useful for **exploratory data analysis**
- Widely used in many fields, including **population genetics, community ecology, macroevolutionary analyses**, etc.
- Useful when we have a set of “objects” (individuals, species, etc), and a (large) set of variables associated with each object
- The variables are **numerous** and may be **correlated in unknown ways**

Multivariate data



Genotype data

M diploid genomes

N loci

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 2 \\ 2 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 2 & 2 & 1 & 1 & 0 \end{bmatrix}$$

Motivation

- Order objects in a way that **similar objects are near each other** and dissimilar objects are farther from each other

- Reduce data to a few axes of variation (dimensionality-reduction) to facilitate **recognition of patterns**

- Gradients reflect **underlying factors or processes**

PCA

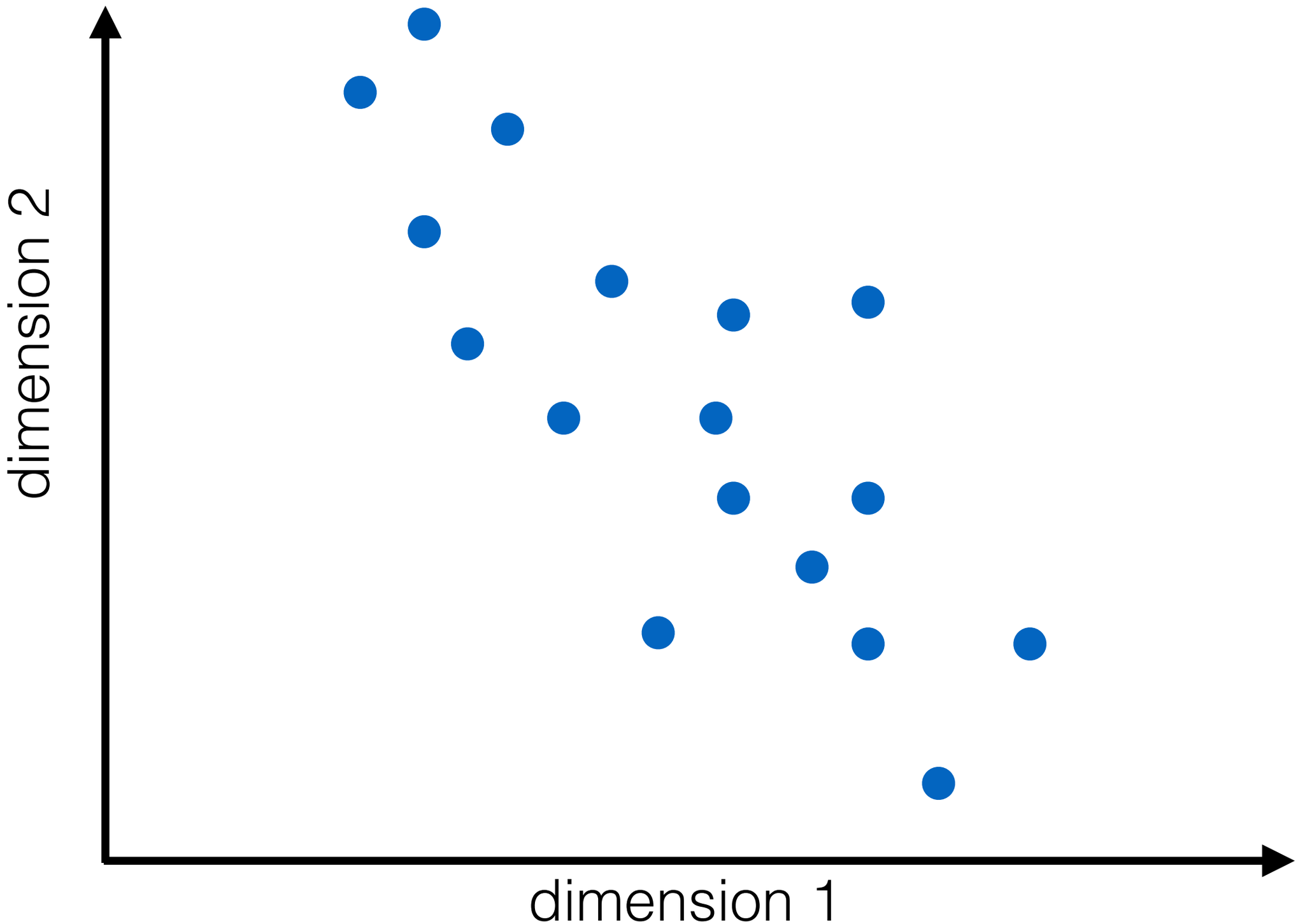
- Principal Component Analysis: an orthogonal transformation of a set of observations of correlated variables into a set of values of linearly uncorrelated variables

- A technique for **dimensionality reduction**

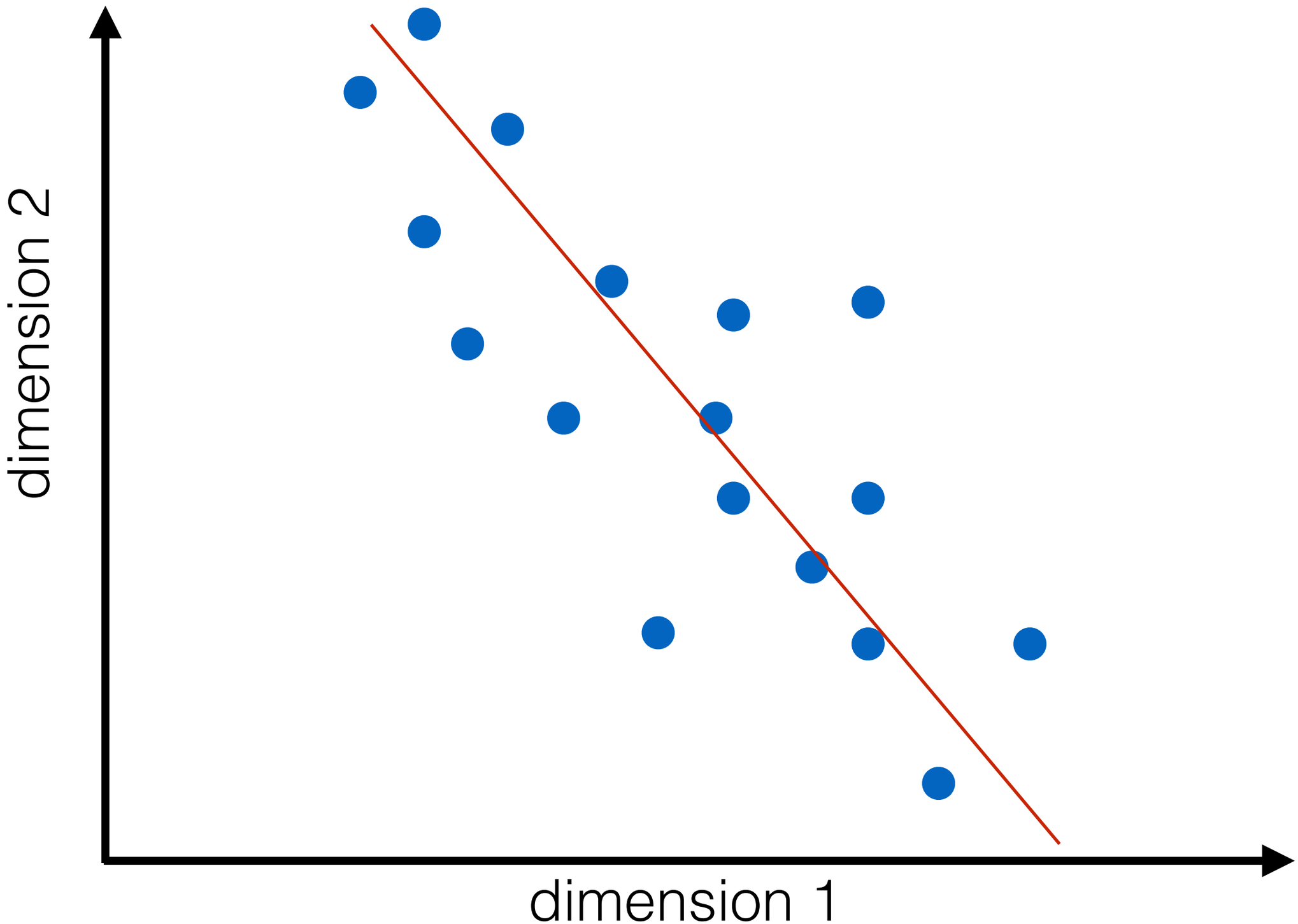
- A technique for extracting the **principal axes of variation** in a dataset

- These axes are orthogonal to each other (and are therefore uncorrelated)

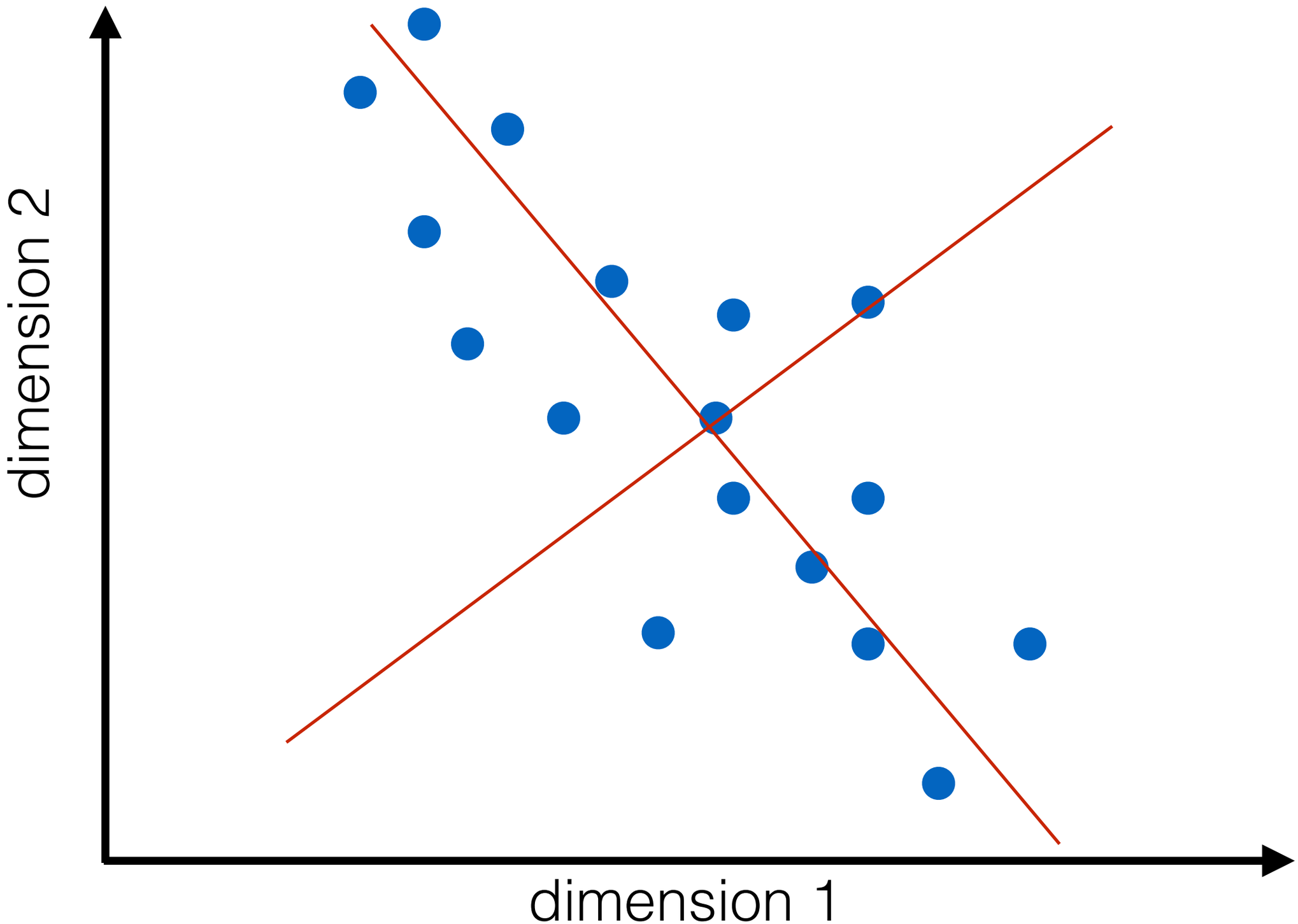
Finding the best **orthogonal** axes of variation



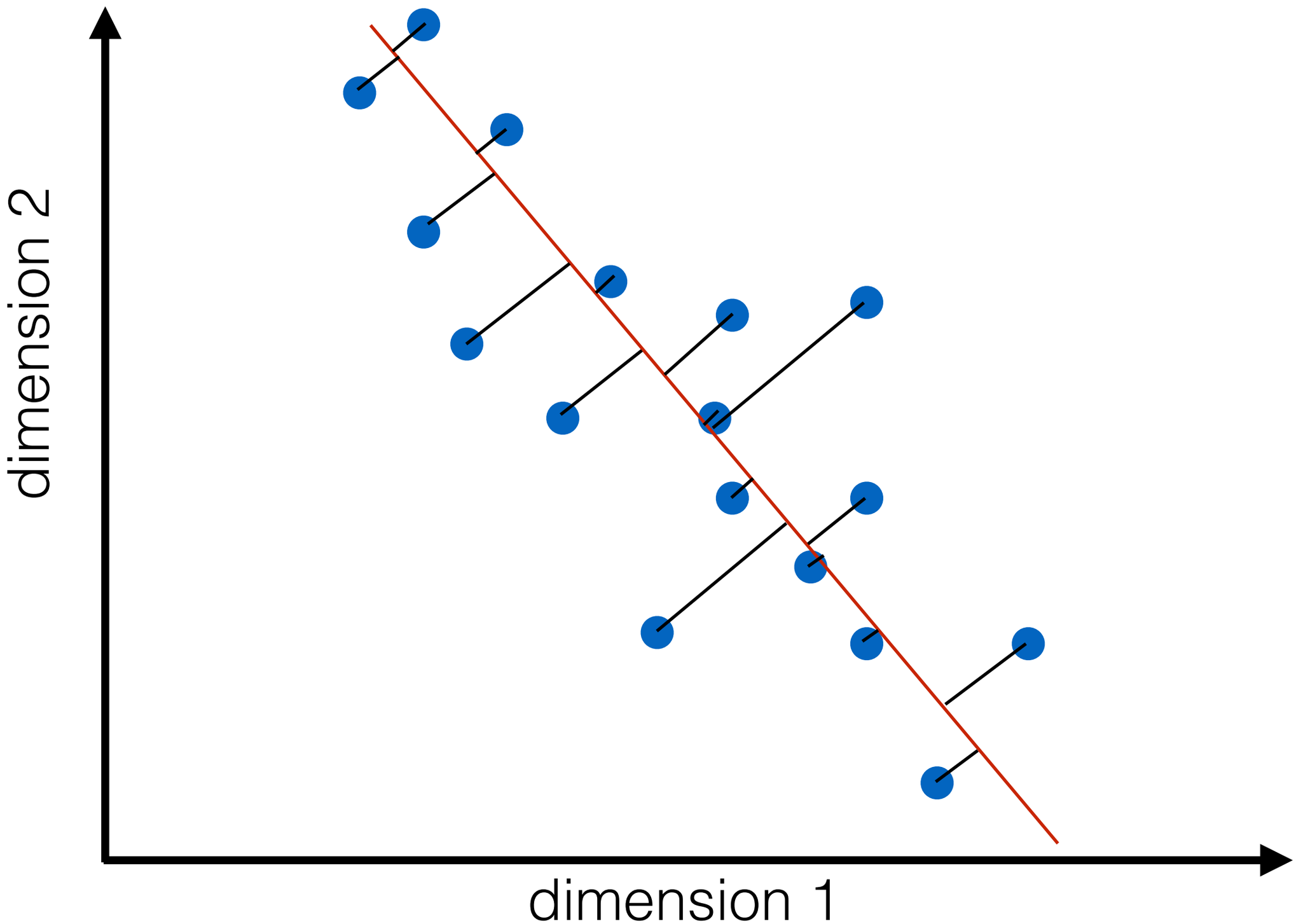
Finding the best **orthogonal** axes of variation



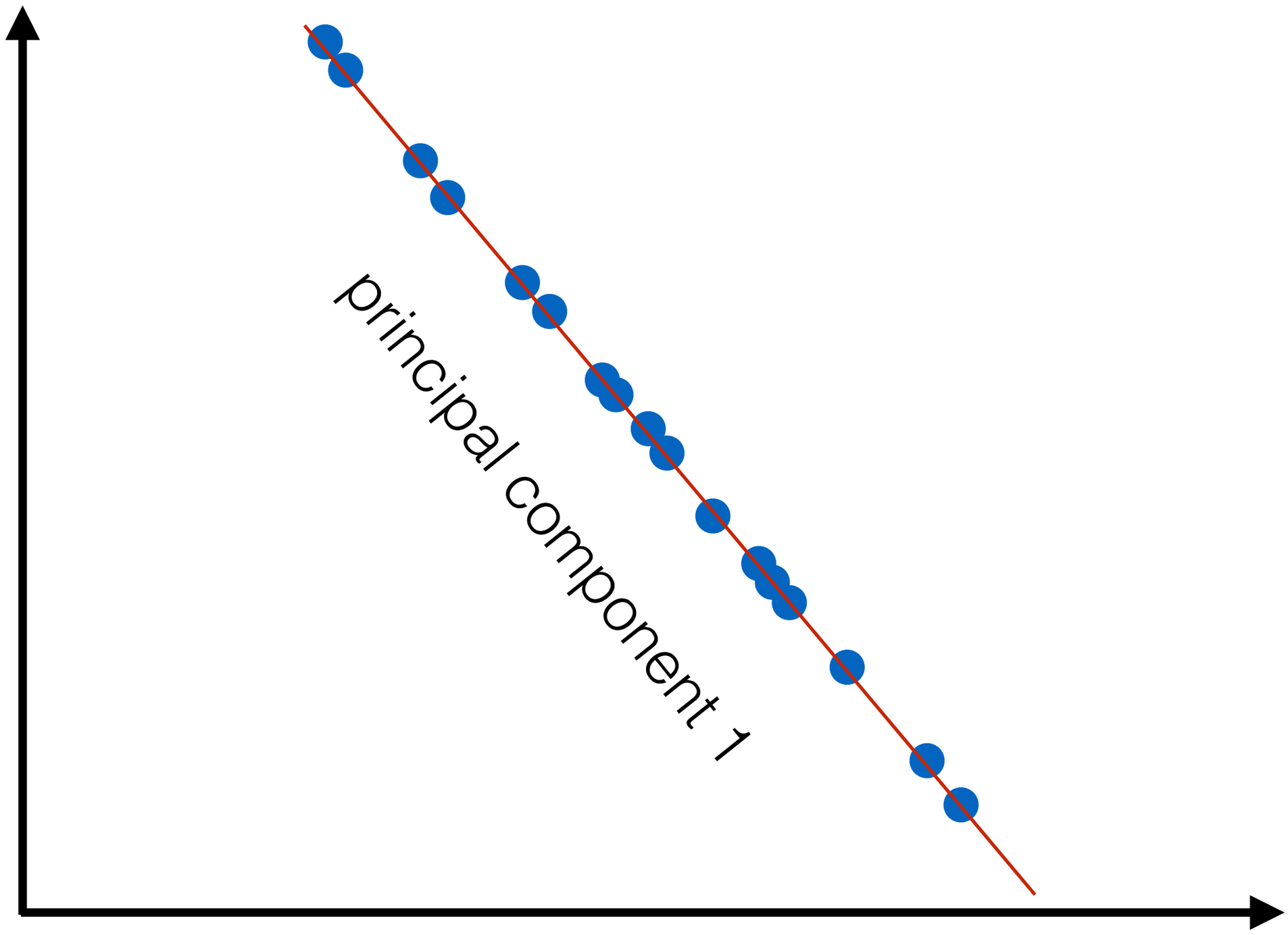
Finding the best **orthogonal** axes of variation



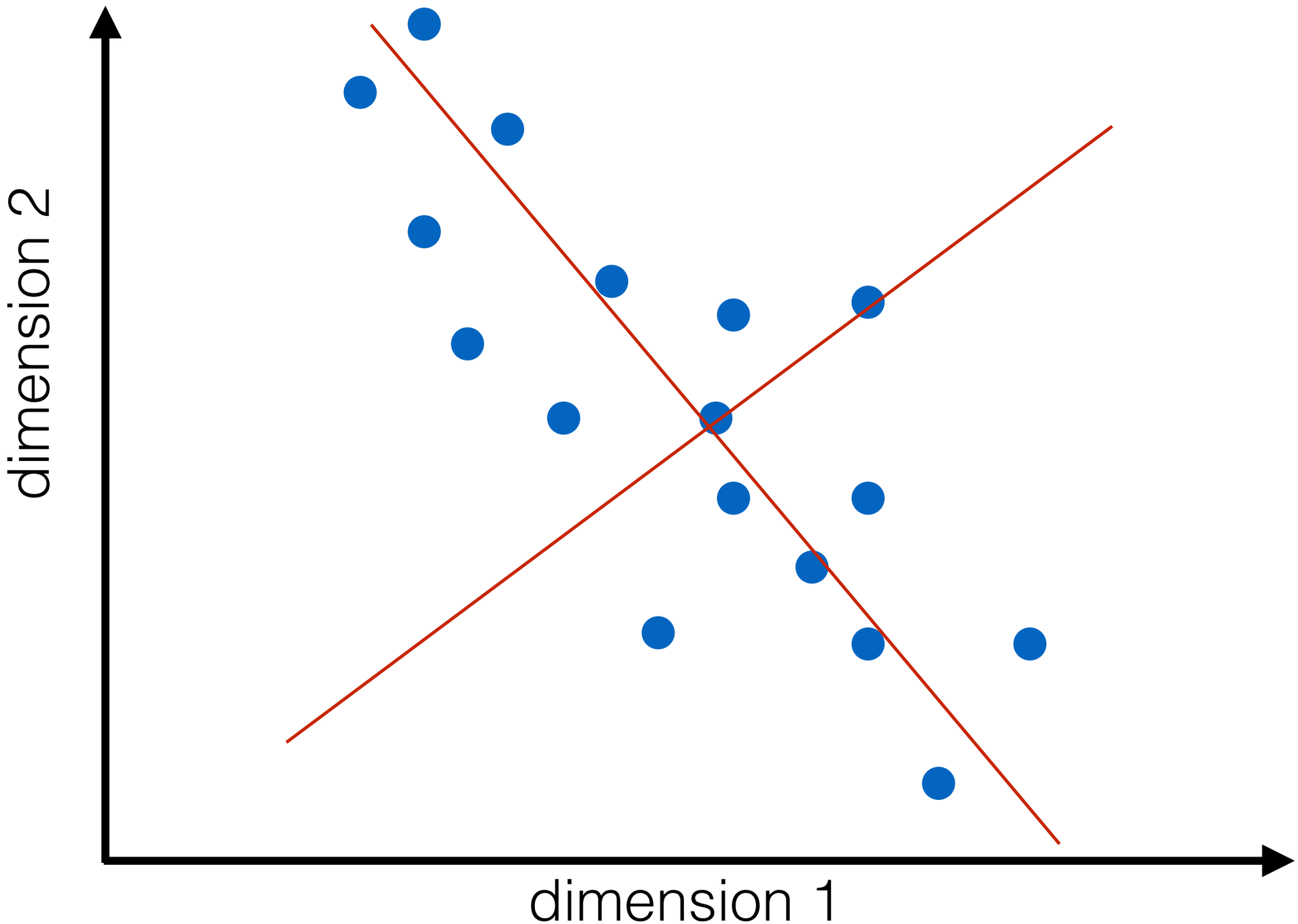
Projecting data onto **orthogonal** axes



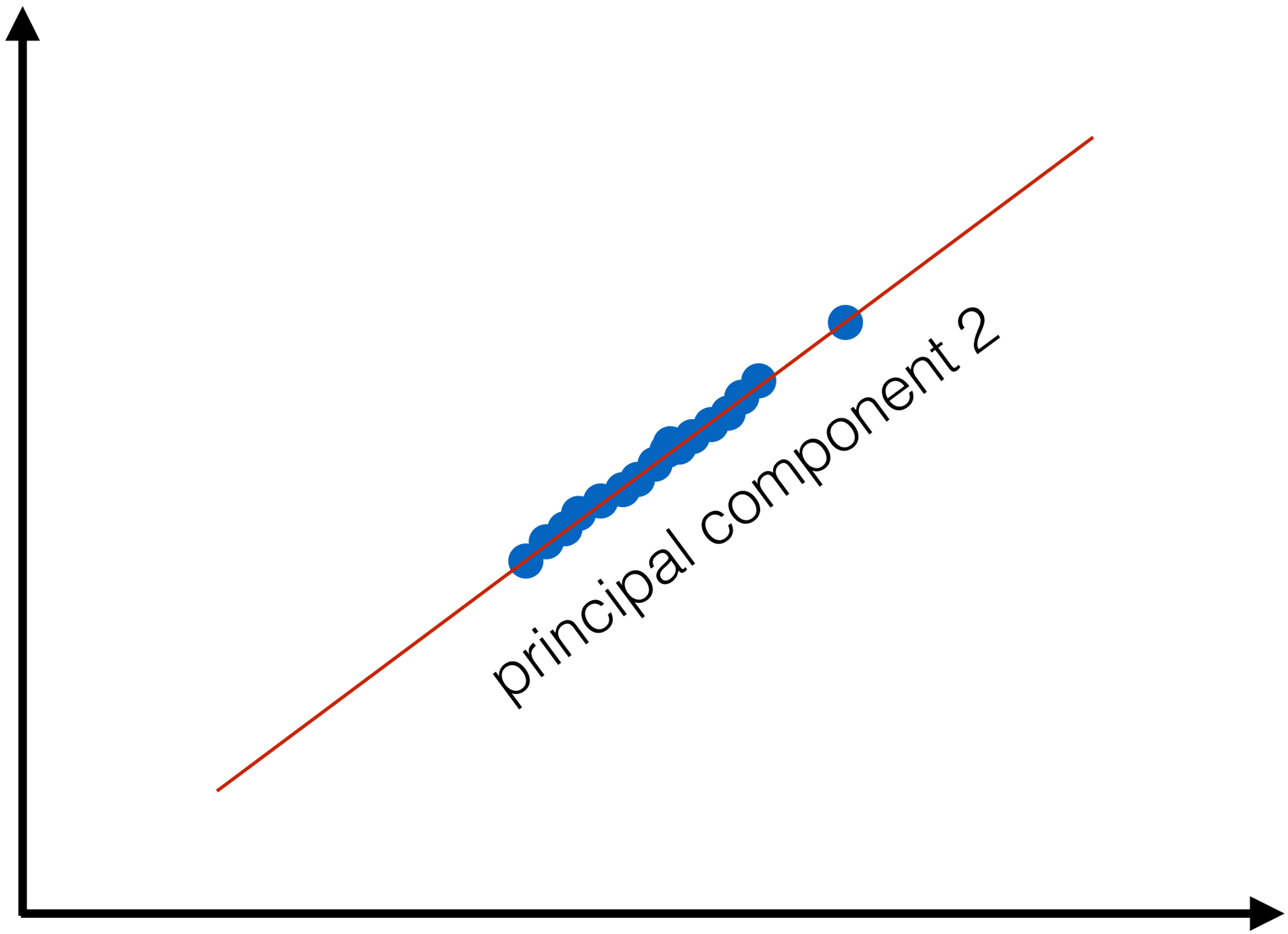
Projecting data onto **orthogonal** axes



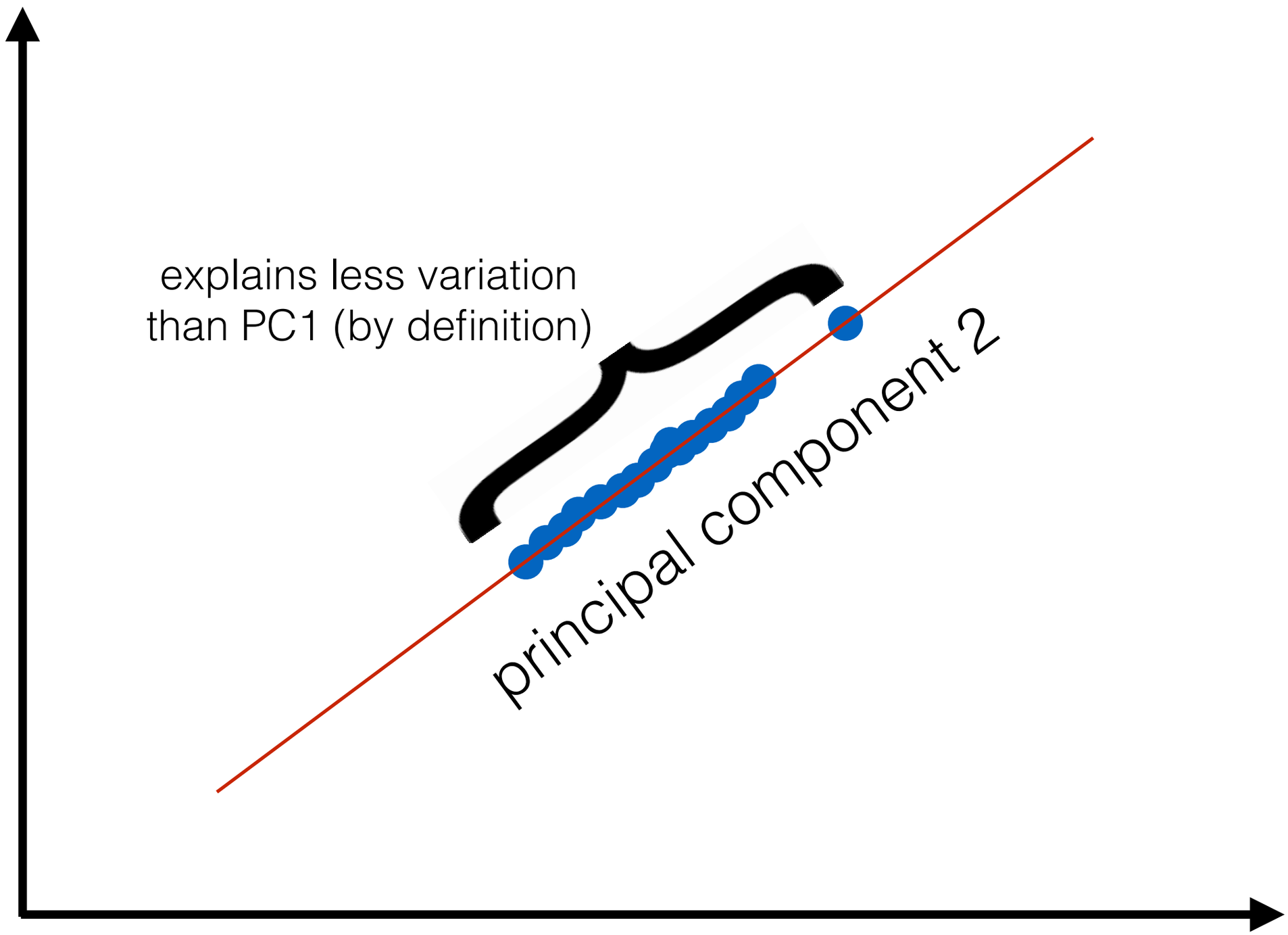
Projecting data onto **orthogonal** axes



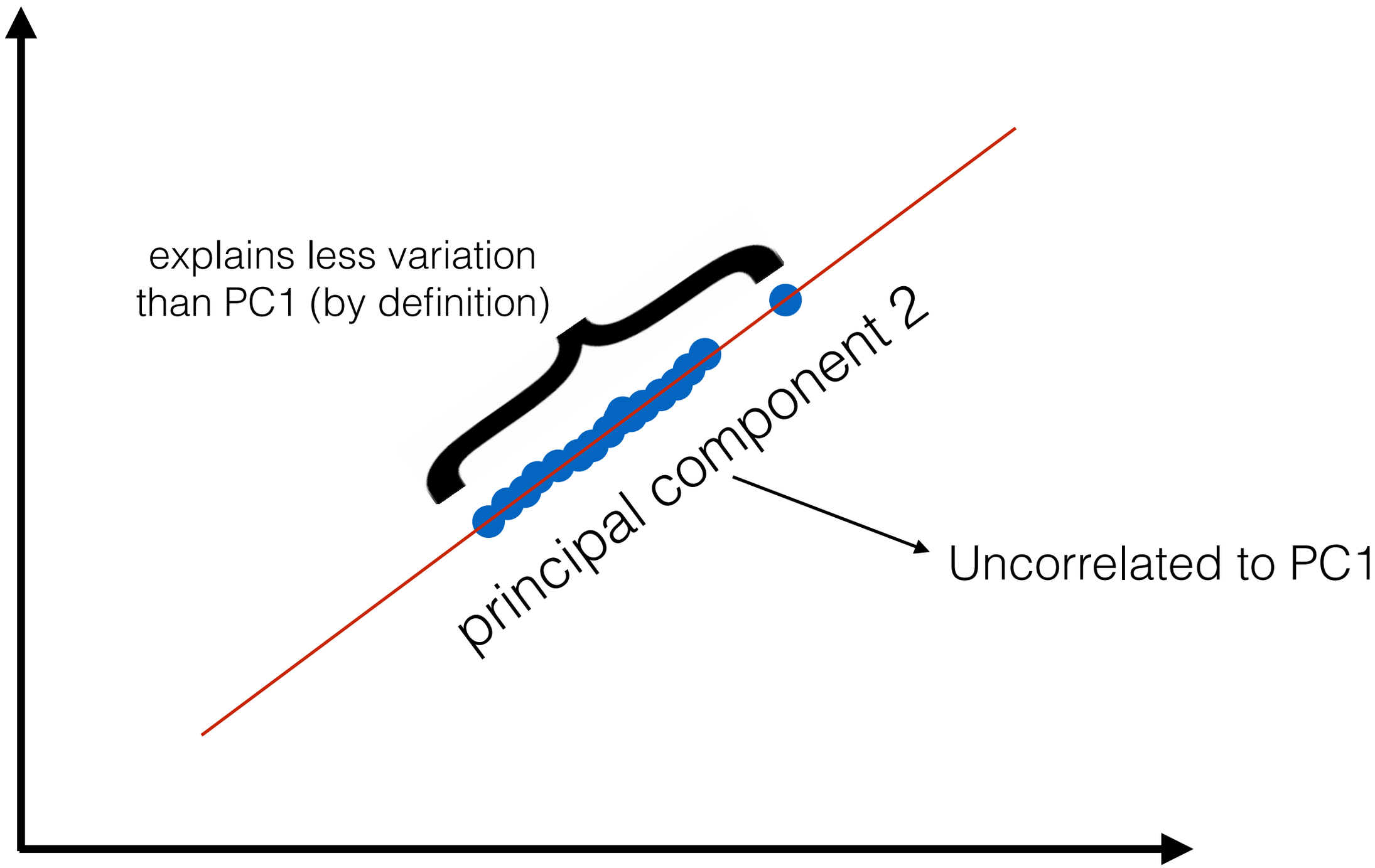
Projecting data onto **orthogonal** axes



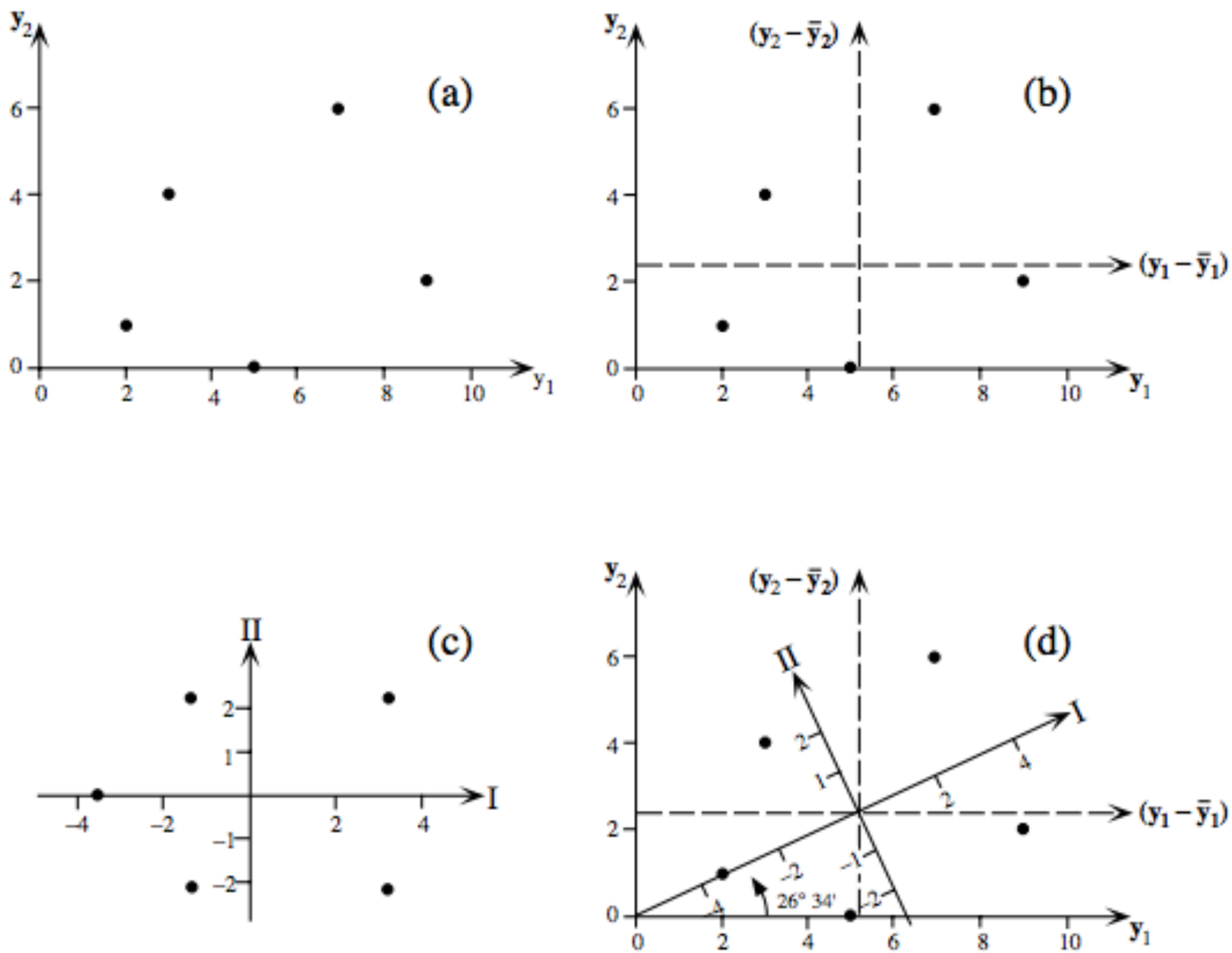
Projecting data onto **orthogonal** axes



Projecting data onto **orthogonal** axes



PCA as a **centered rotation** in N-dimensional space



Step 1: Mean-center data matrix

M objects

N variables

1	1	1	0	0
0	1	2	1	2
2	1	1	0	1
0	0	1	2	2
2	1	1	0	0
0	0	1	1	1
2	2	1	1	0
…	…	…	…	…

Mean-center each variable →

0.4	0.4	0.4	-0.6	-0.6
-1.2	-0.2	0.8	-0.2	0.8
1.0	0.0	0.0	-1.0	0.0
-1.0	-1.0	0.0	1.0	1.0
1.2	0.2	0.2	-0.8	-0.8
-0.6	-0.6	0.4	0.4	0.4
0.8	0.8	-0.2	-0.2	-1.2
…	…	…	…	…

Step 2: Compute covariance matrix (example with N=2)

$$\mathbf{S} = \frac{1}{n-1}\mathbf{Y}_c{}'\,\mathbf{Y}_c = \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix} \longrightarrow \text{Covariance matrix}$$

$$\begin{bmatrix} -3.2 & -2.2 & -0.2 & 1.8 & 3.8 \\ -1.6 & 1.4 & -2.6 & 3.4 & -0.6 \end{bmatrix} \qquad \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix}$$

Step 2: Compute covariance matrix (example with N=2)

$$\mathbf{S} = \frac{1}{n-1}\mathbf{Y}_c{}'\,\mathbf{Y}_c = \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix}$$

Covariance matrix

Cov[1,2]

Cov[1,1] = Var[1]

Cov[2,2] = Var[2]

$$\begin{bmatrix} -3.2 & -2.2 & -0.2 & 1.8 & 3.8 \\ -1.6 & 1.4 & -2.6 & 3.4 & -0.6 \end{bmatrix}$$

$$\begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix}$$

Step 3: Find eigenvectors and eigenvalues of Cov. Mat.

To find eigenvalues, solve this equation:

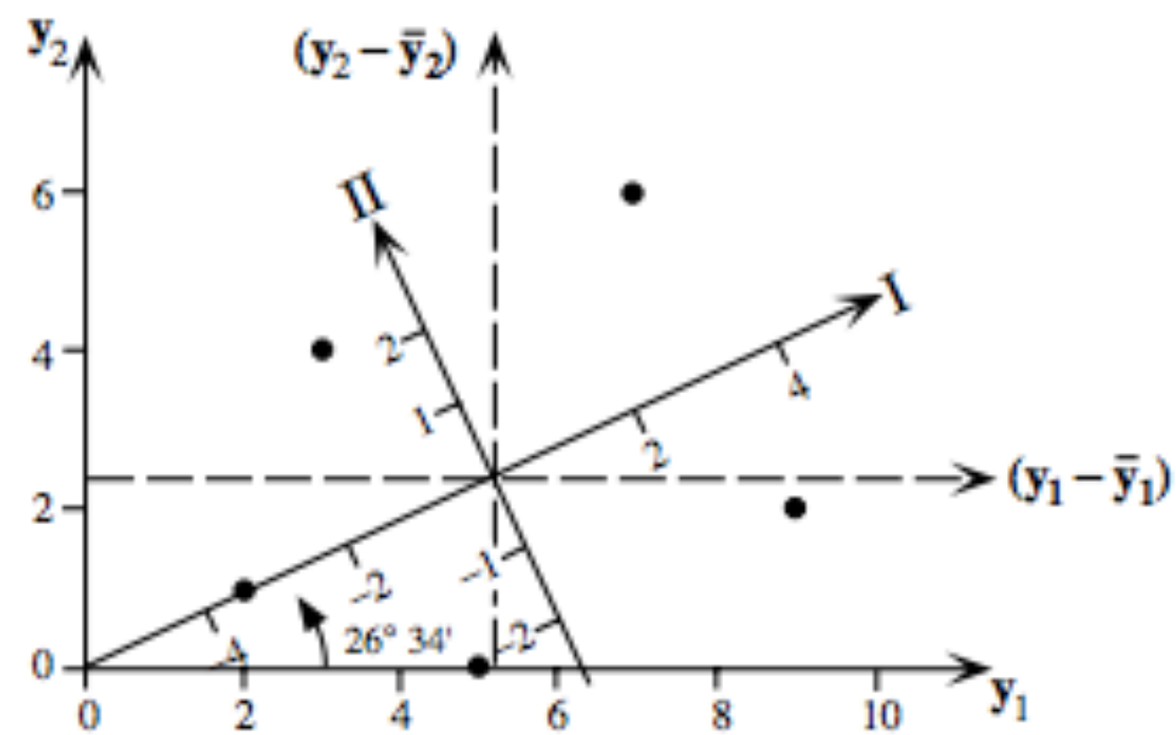
$$|\mathbf{S} - \lambda_k \mathbf{I}| = \left| \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix} - \begin{bmatrix} \lambda_k & 0 \\ 0 & \lambda_k \end{bmatrix} \right| = 0$$

To find eigenvectors, solve this equation:

$$(\mathbf{S} - \lambda_k \mathbf{I})\mathbf{u}_k = \mathbf{0}$$

Remember: eigenvectors are a new **perpendicular** (orthogonal) set of **coordinate axes**

$$\mathbf{U} = \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix}$$



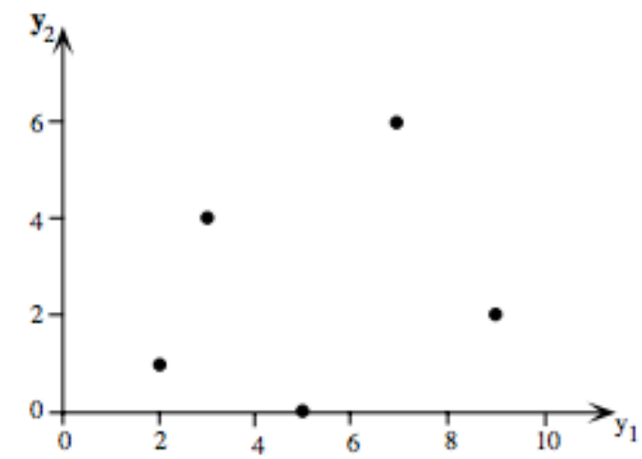
$$\mathbf{u}'_1\mathbf{u}_2 = (0.8944 \times (-0.4472)) + (0.4472 \times 0.8944) = 0$$

Step 4: Project data points into new axes

$$\mathbf{F} = \mathbf{Y}_c\,\mathbf{U}$$

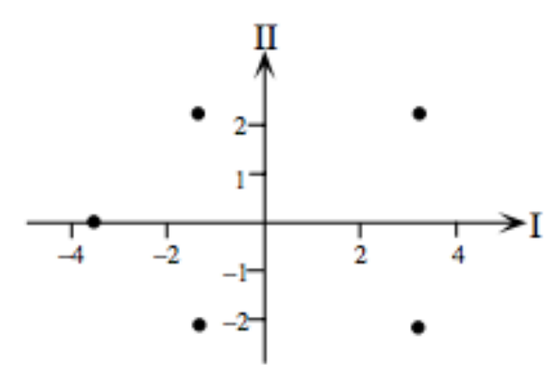
$$\mathbf{F} = \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix} \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} = \begin{bmatrix} -3.578 & 0 \\ -1.342 & 2.236 \\ -1.342 & -2.236 \\ 3.130 & 2.236 \\ 3.130 & -2.236 \end{bmatrix}$$

original data | new axes | data in PCA space



$\mathbf{Y}_c$

$\mathbf{U}$



$\mathbf{F}$

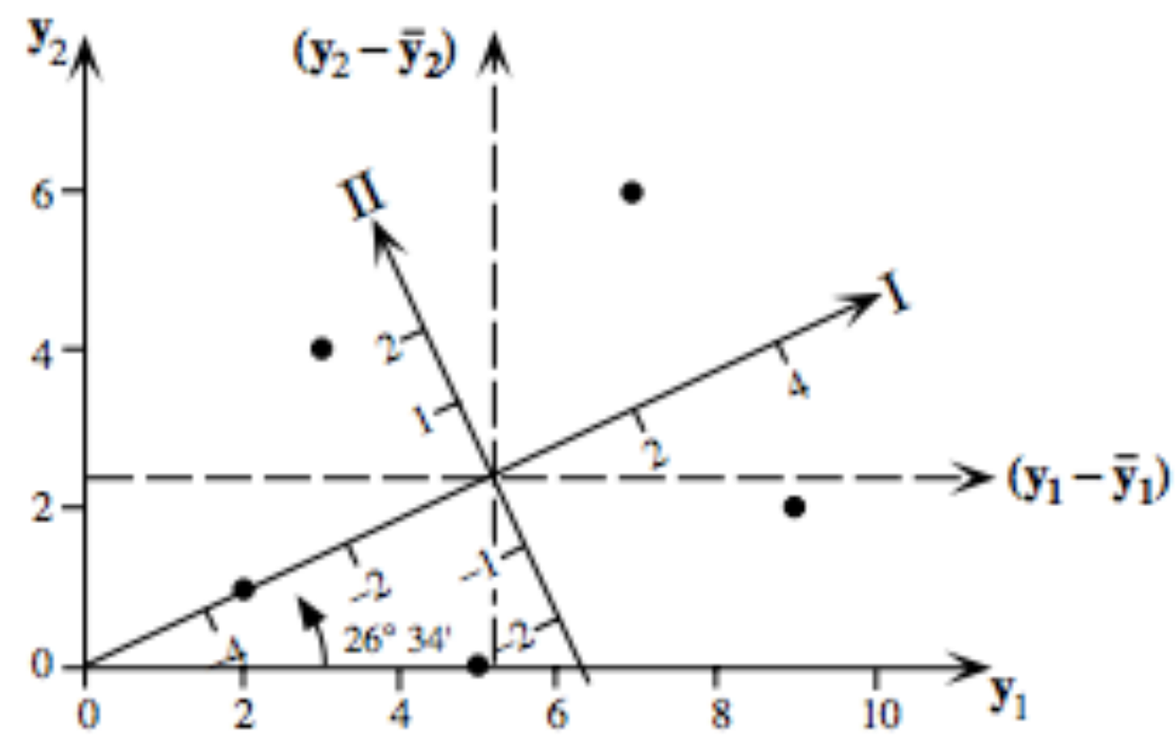
Why are we doing this rotation?

- It is easy to see which axes are the ones that explain the most variation in 2-dimensional space

- It is **much harder to do this (visually) when N is large** (multi-dimensional data)

Why are we doing this rotation?

2 variables $\left\{ \begin{matrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 2 \end{matrix} \right.$



Why are we doing this rotation?

7 variables

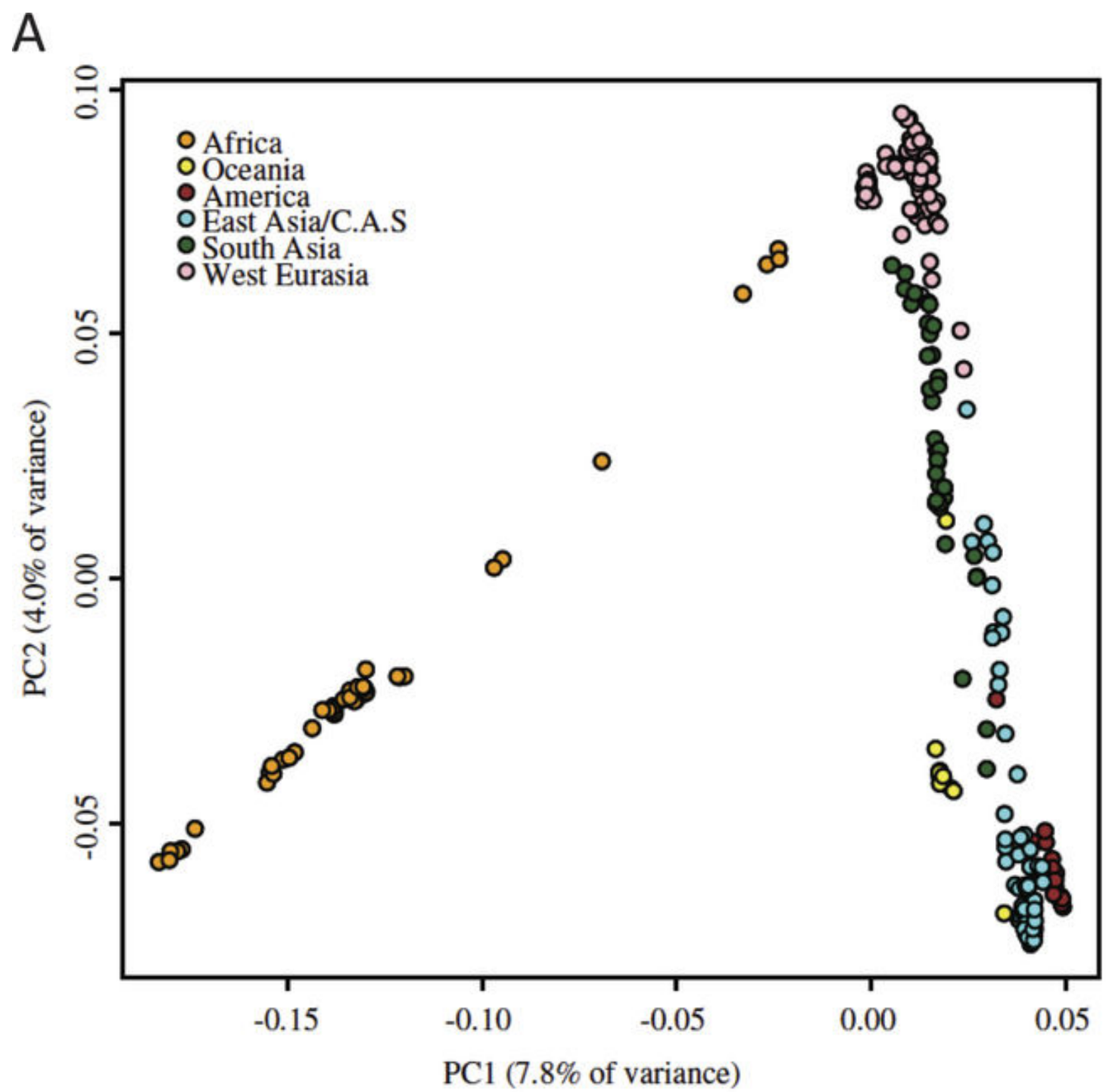
$$\left\{\begin{matrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 2 \\ 2 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 2 & 2 & 1 & 1 & 0 \end{matrix}\right.$$

?

Why are we doing this rotation?

- In a PCA, each eigenvector has a corresponding eigenvalue
- The **largest eigenvalues** correspond to the **eigenvectors that explain the most variation**
- Percent of variance explained by eigenvector k = eigenvalue k / (sum of all eigenvalues)
- **Largest eigenvalue -> largest axis of variation**

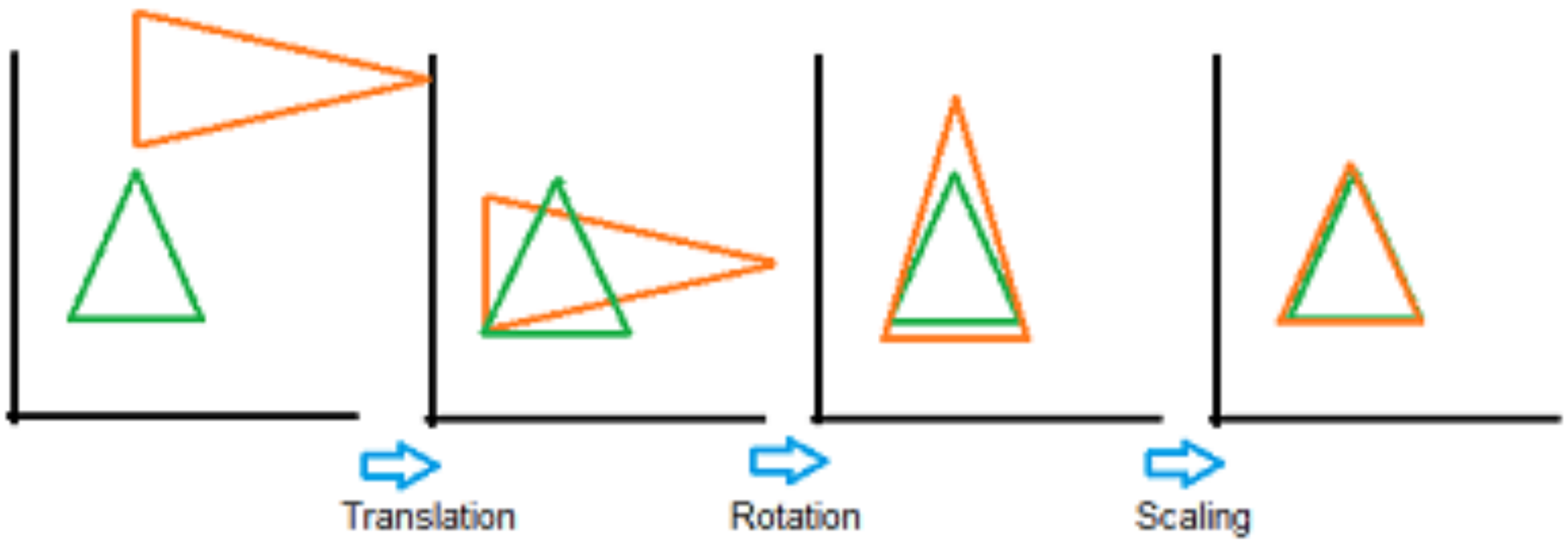
PCA of worldwide human genomes



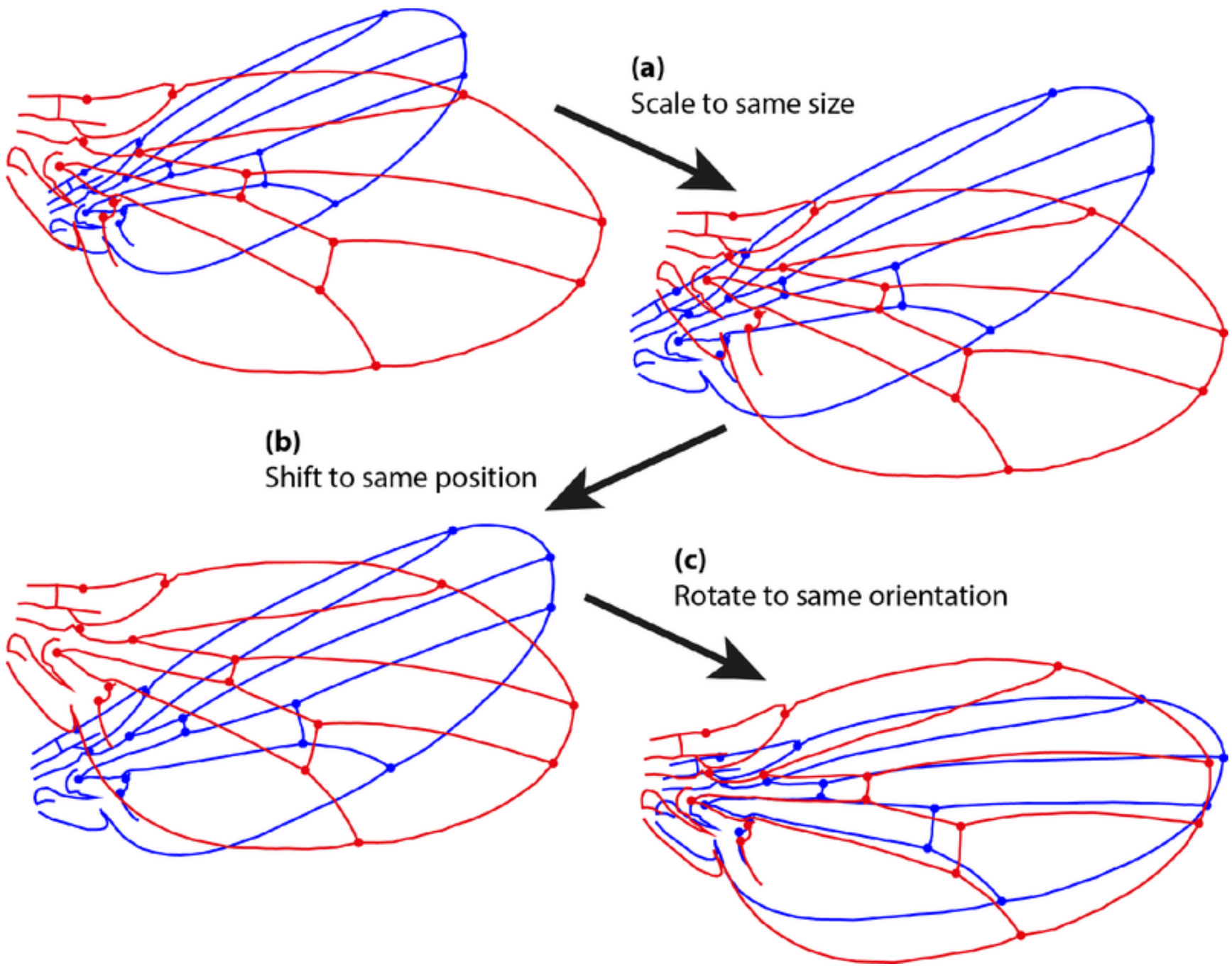
Dealing with missing data: Procrustes transformation

- SNPs in which at least 1 sample has missing data are unusable in a PCA
- Problem: low coverage genomes -> many sites with missing data
- Even bigger problem: combination of many low-coverage genomes -> very few sites with overlap in coverage across all of them
- Solution (Skoglund et al. 2012):
    - For each low-coverage genome, run 1 PCA (with many high-coverage genomes included)
    - Combine loadings from each individual PCA into an overall-PCA, using Procrustes transformation

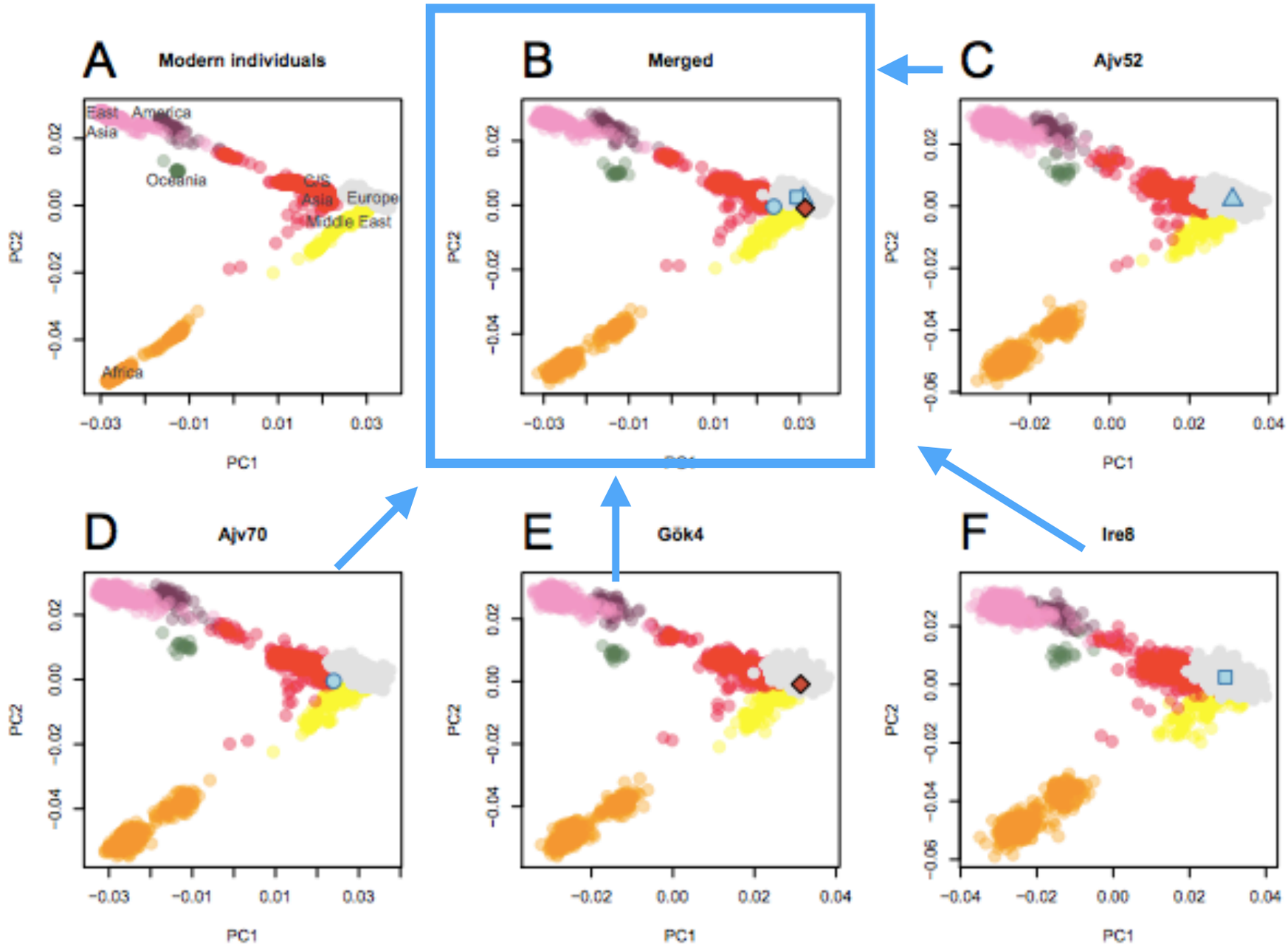
Shape-preserving Procrustes transformation



Shape-preserving Procrustes transformation

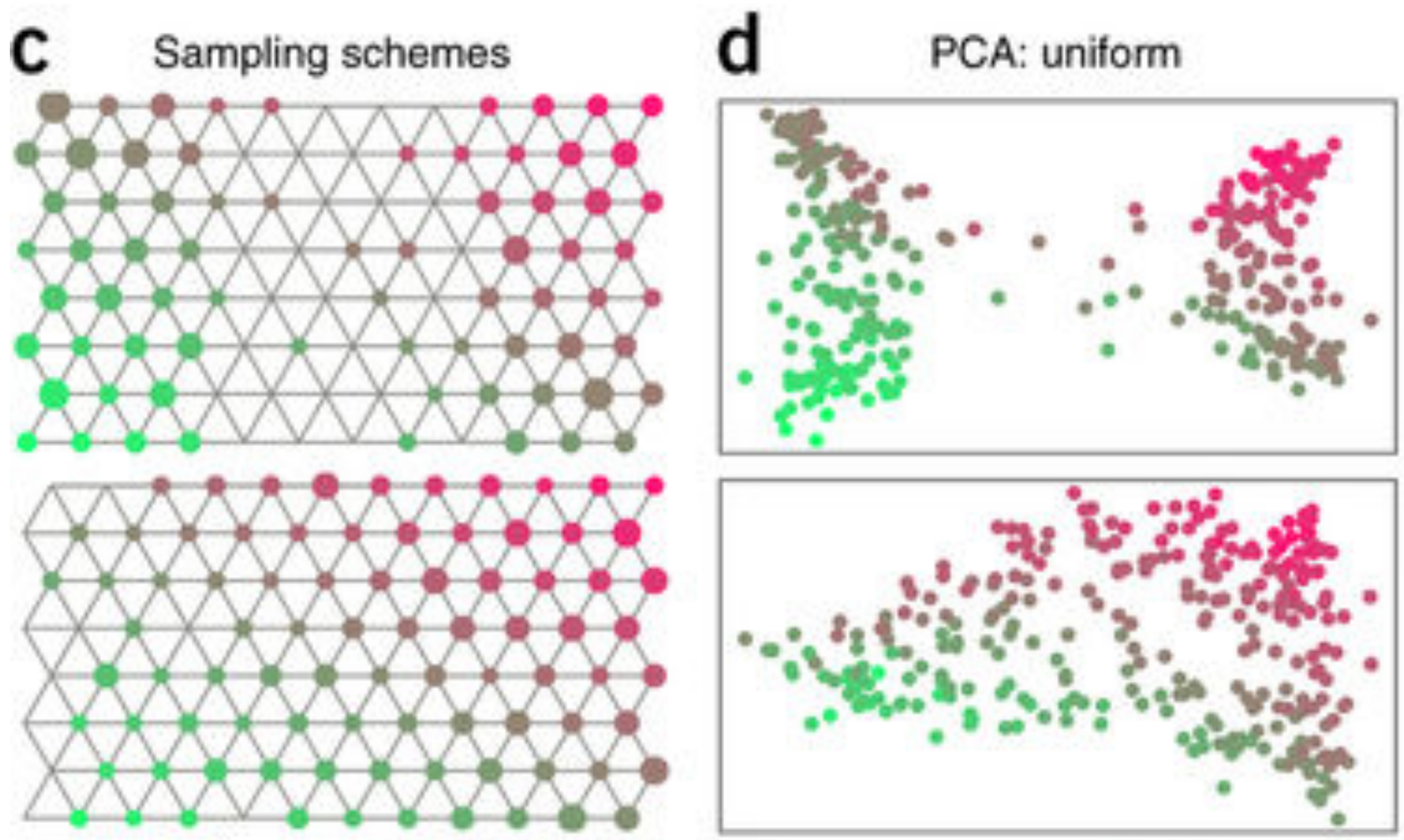


Use a Procrustes transformation using a **high-coverage reference PCA**

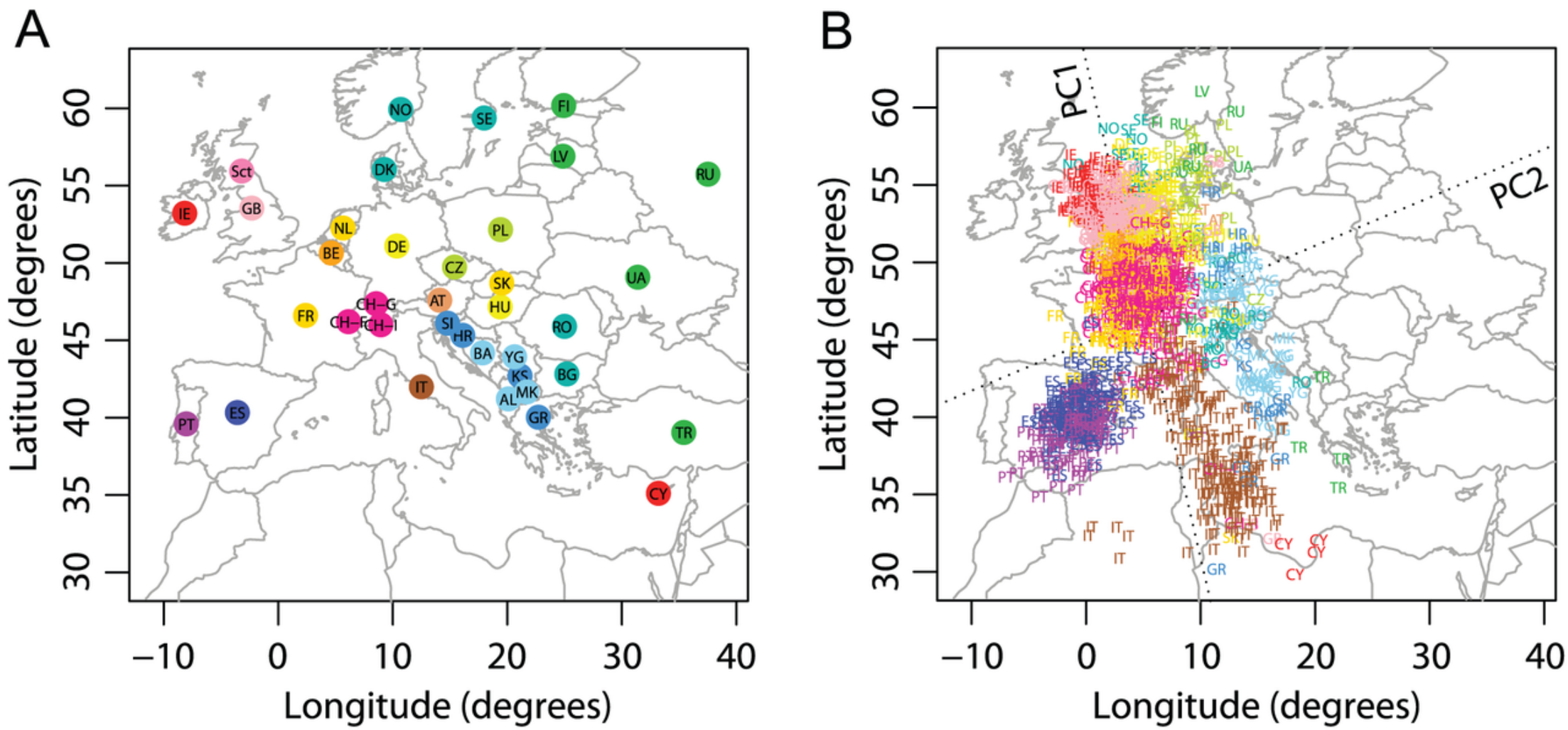


Skoglund et al. 2012

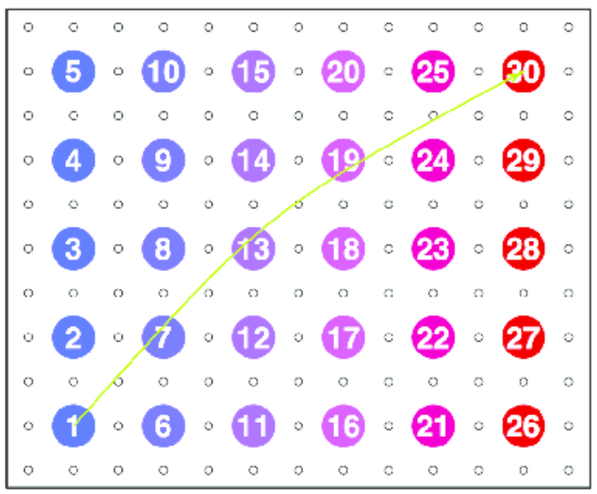
Sampling scheme can be misleading



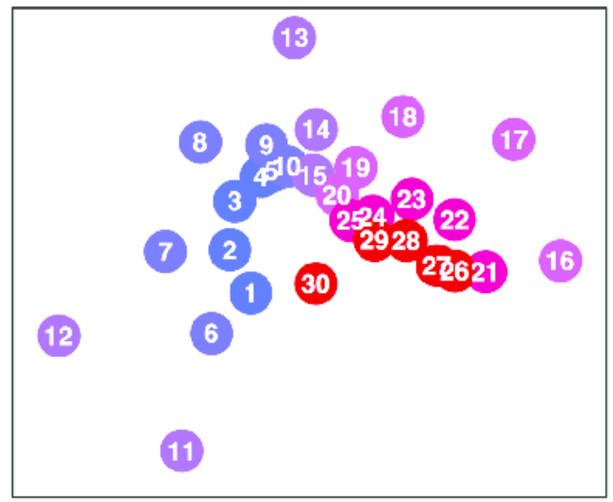
PCA recovers signals of “isolation-by-distance”



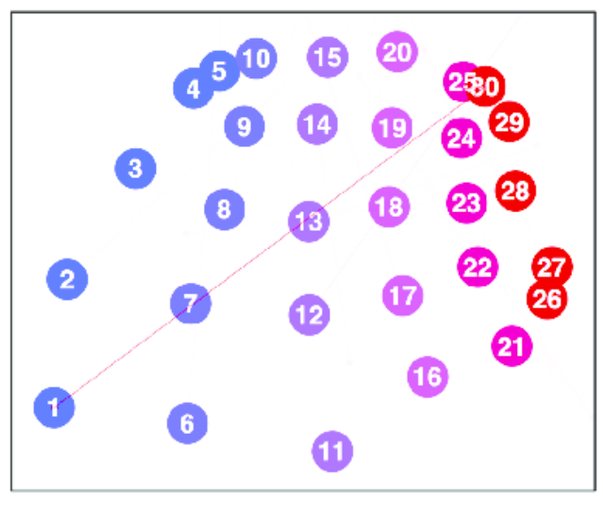
SpaceMix: long-range admixture + isolation-by-distance



(a) simulated lattice with admixture



(b) geogenetic map without admixture inference



(c) geogenetic map with admixture inference

Today

- Exploratory vs. hypothesis-driven analyses
- PCA
- **Latent mixed-membership models ("Structure")**

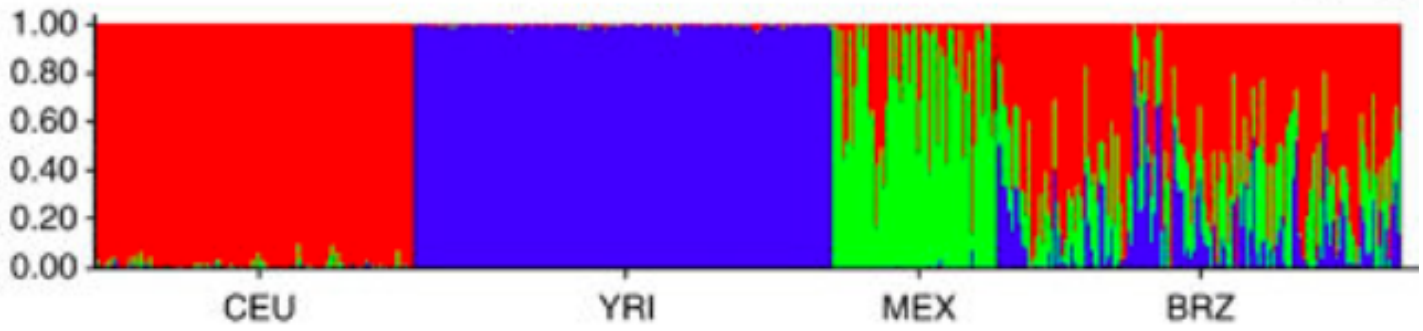
Latent mixed-membership models ("Structure")

Fernando Racimo

Copenhagen, August 2018

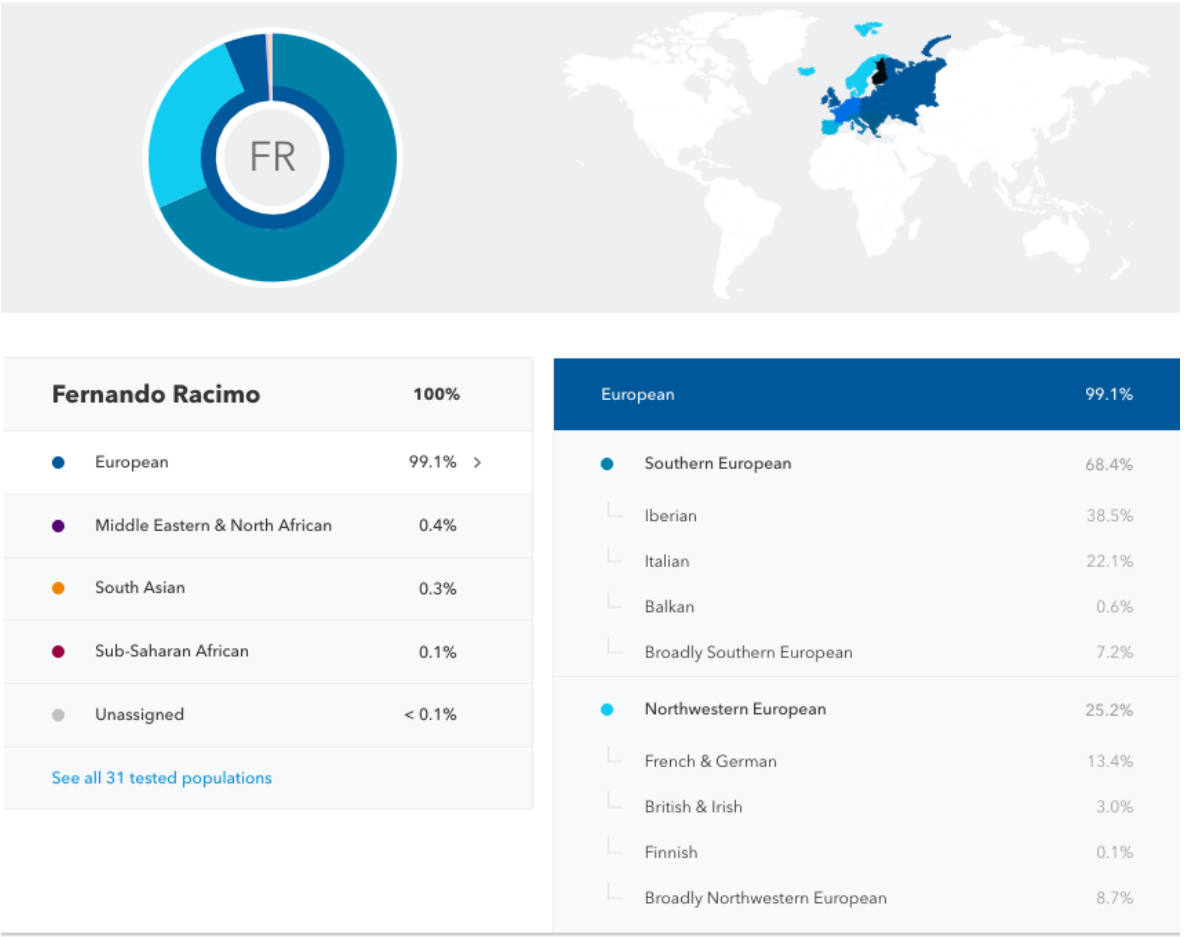
Questions

- Is there population structure in a population?
- Can we identify subpopulation clusters of shared ancestry?
- Are individuals best modeled as mixtures of ancestral populations?
- How much admixture was passed on from each population?



Objectives

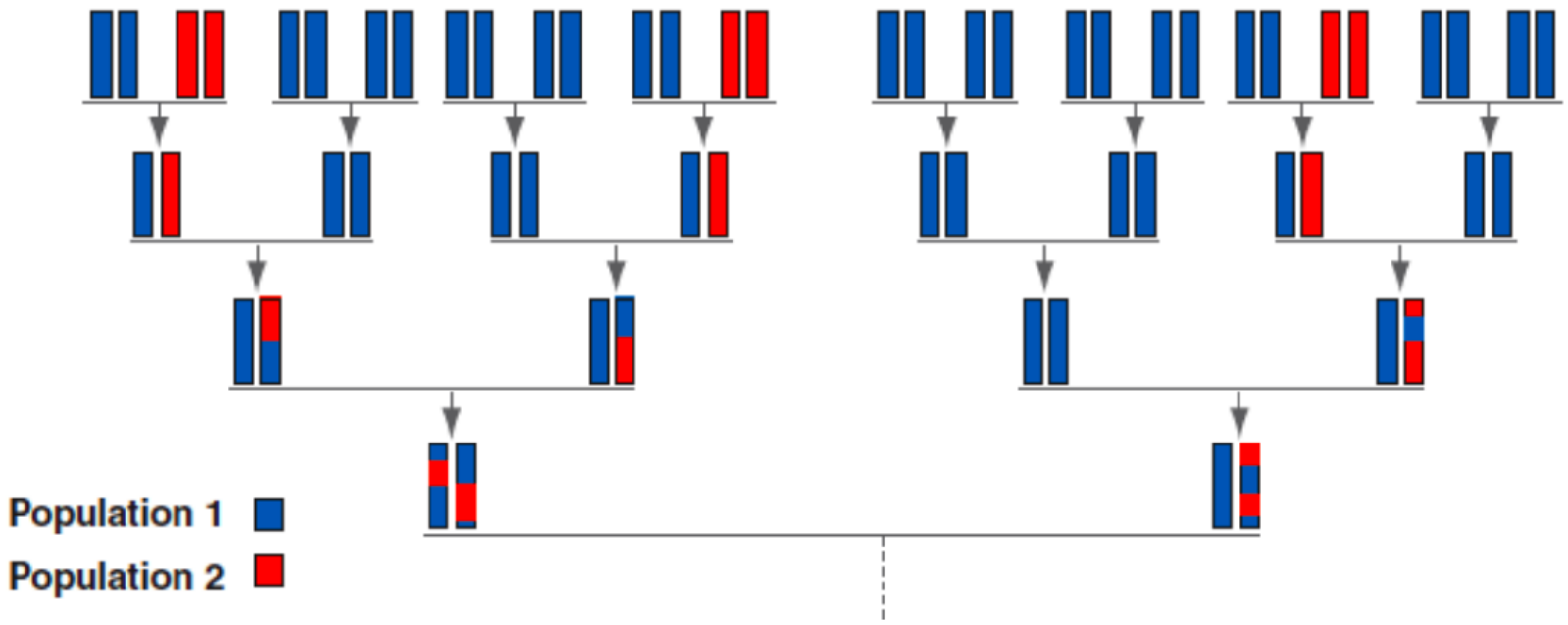
- Learn something about the past genetic history of a population under study
- Learn something about ourselves



The "Structure" model

- The original model was first proposed by Pritchard et al. (2000)
- **Assumption 1**: each individual can be modeled as a mixture of one or more ancestral **"source populations"**
- **Assumption 2**: each locus is independent
- The proportion of genetic matrerial from each source in each individual is called the **"admixture proportion"**
- **Problem 1**: we don't know the identity and number of these source populations
- **Problem 2**: we don't know the admixture proportions
- **Objective**: find best-fitting sources and their proportions

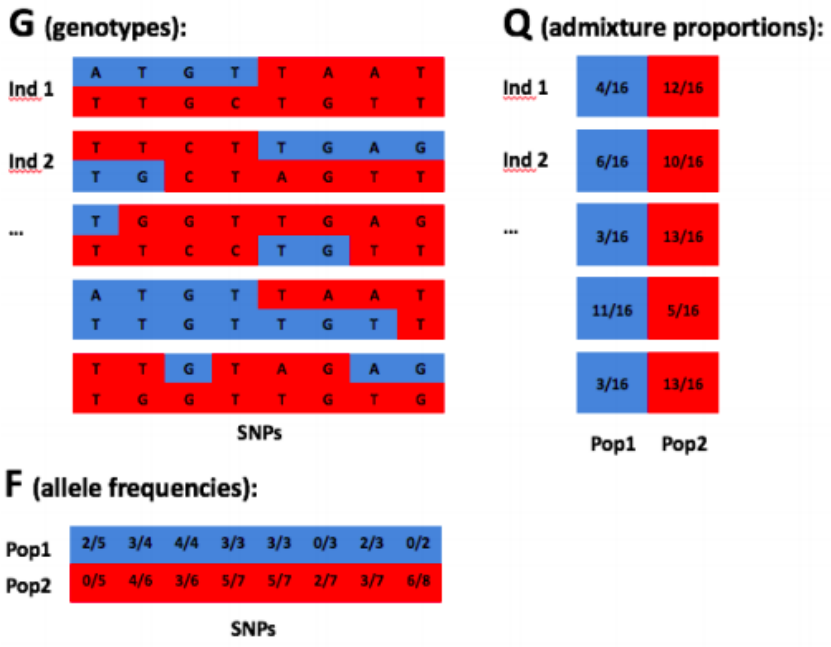
The "Structure" model



[0]Darvasi and Shifman 2005

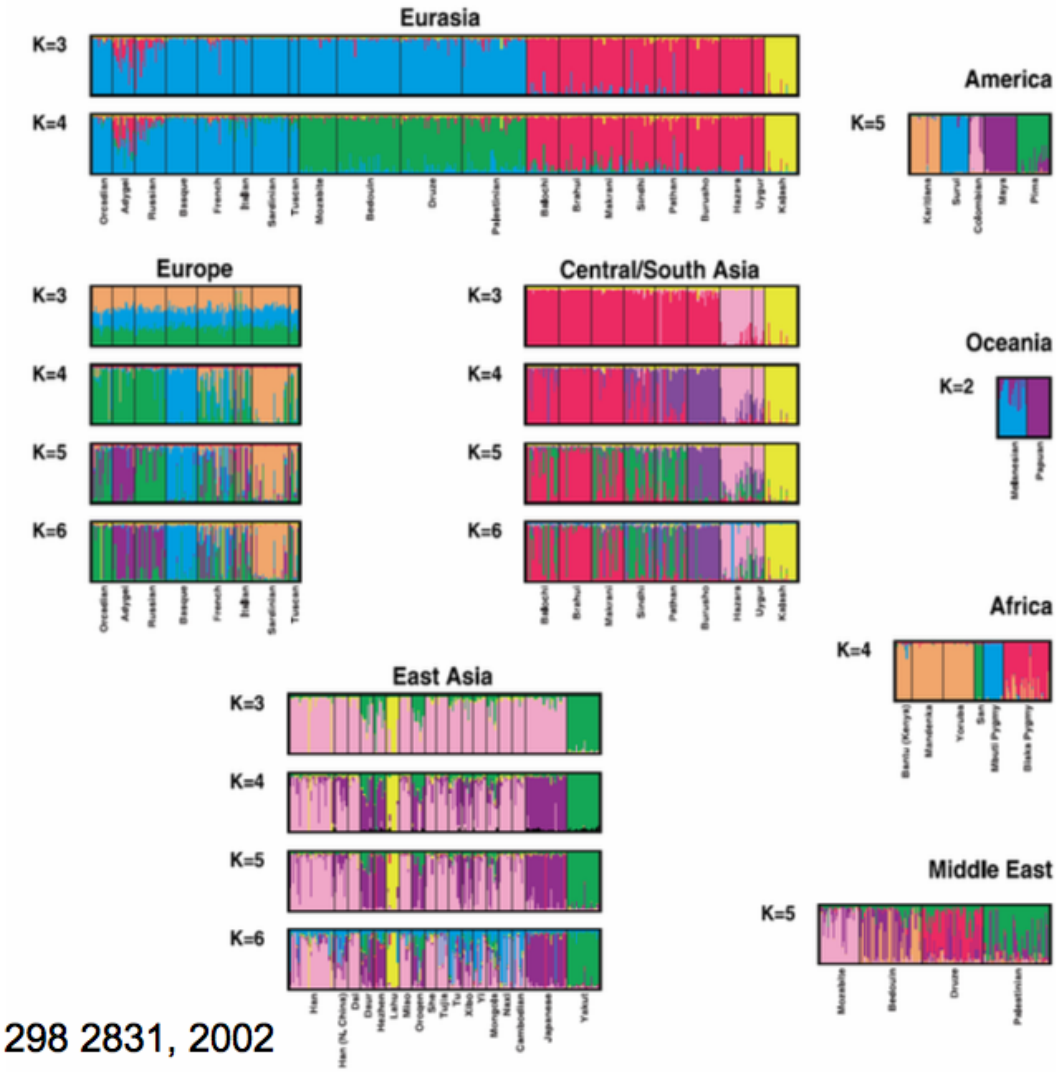
The "Structure" model

- Known: genotypes (G)
- Unknown:
  - admixture proportions (Q)
  - allele frequencies in source populations (F)
- Need to estimate Q and F, given that we know G.
- Objective: Maximize likelihood function: $P[G|Q, F]$



[0] Ida Moltke pers. comm.

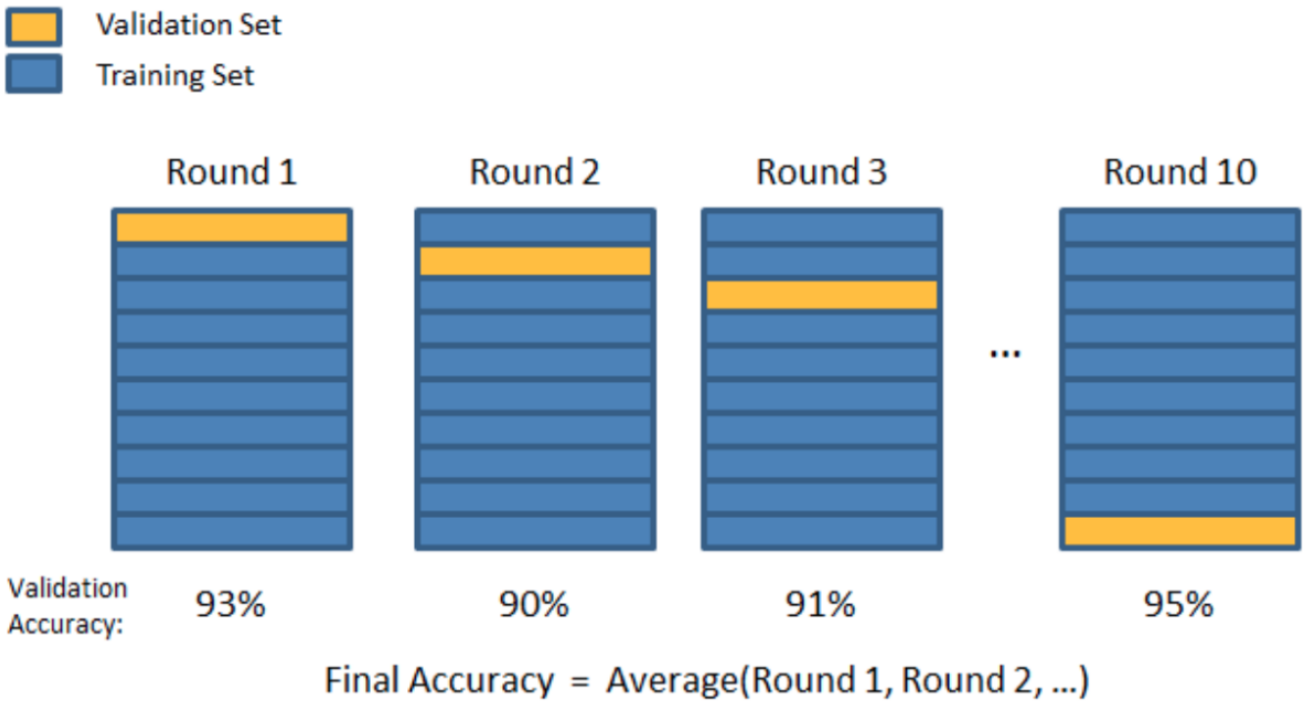
Structure model applied to human populations



Science 298 2831, 2002

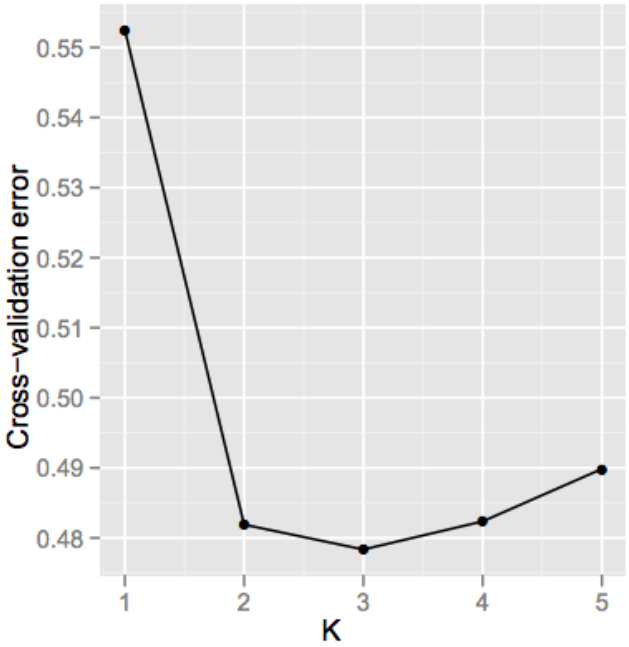
Choosing K

- We can use cross-validation to find a value of K that does not over-fit
- We leave some genotypes out and predict them based on their estimated ancestries
- **Important**: well-fitting parameter $\neq$ biologically meaningful parameter



Choosing K

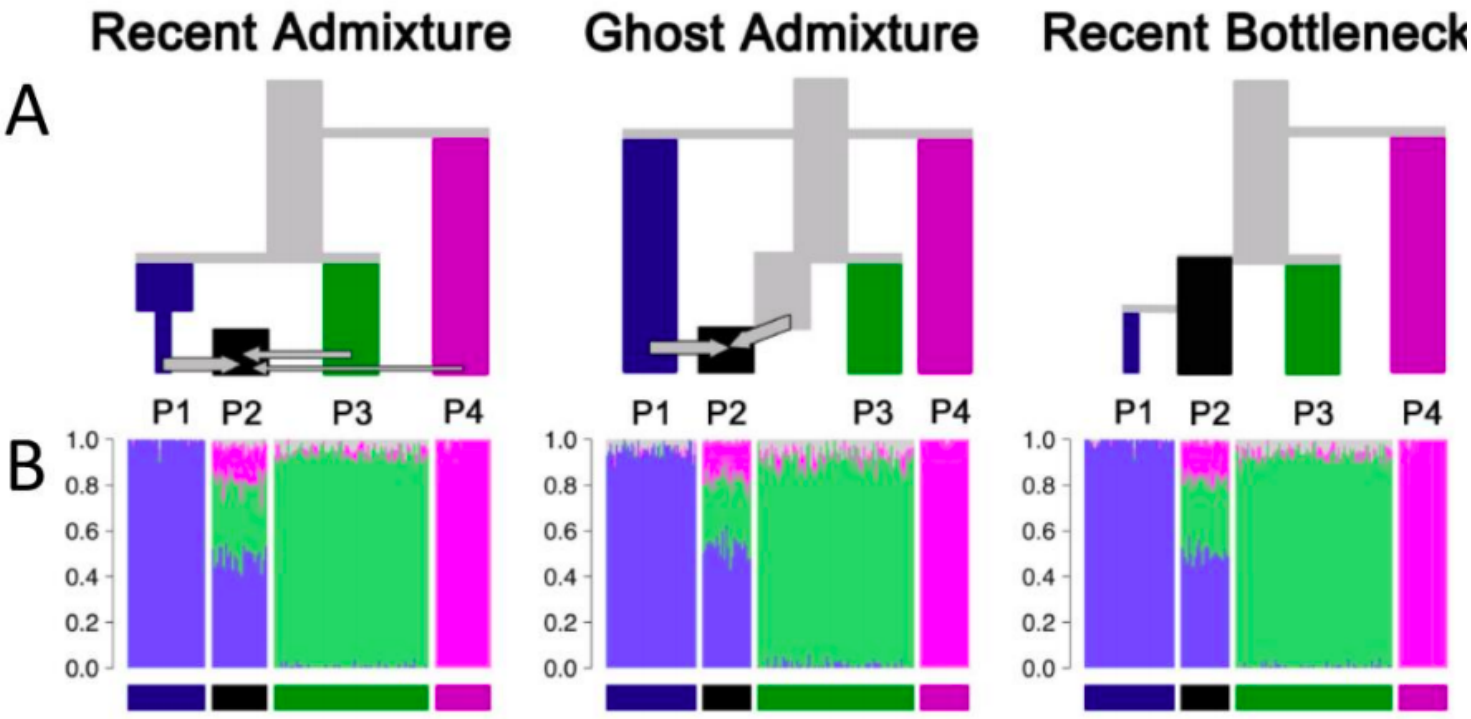
- We can use cross-validation to find a value of K that does not over-fit
- We leave some genotypes out and predict them based on their estimated ancestries
- **Important**: well-fitting parameter $\neq$ biologically meaningful parameter



[0] Alexander et al. 2011

Over-interpreting Structure results

- Structure does not necessarily pick up admixture events![1]
- "Source populations" need not be real populations that ever existed!
- A population that is highly drifted will be assigned its own cluster at high enough K



[1] Falush et al. 2016

Variations on a theme...

- Structure (Pritchard et al. 2000): original model; uses Bayesian priors to obtain posterior estimates of Q and F
- Admixture (Alexander et al. 2011): faster than Structure; uses a maximum likelihood model rather than a Bayesian model; uses cross-validation to choose K
- fastStructure (Raj et al. 2014): faster than Structure; uses variational inference to choose K; can detect weak structure
- ngsAdmix (Skotte et al. 2013): can work with genotype likelihoods; better for low coverage data
- Ohana (Cheng et al. 2016): uses Gaussian approximation to model drift in each ancestry component; can detect selection by testing for local deviations from genome-wide model