

Principal Component Analysis

E. Fersini

Why Feature Reduction?

Why even think about Feature Reduction?

- Naive theoretical view:

More features

=> More information

=> More discrimination power.

- In practice:

many reasons why this is not the case!

- We need to identify features with a strong
discriminative/predictive power



Why Feature Reduction?

Chiwawa or Muffin?



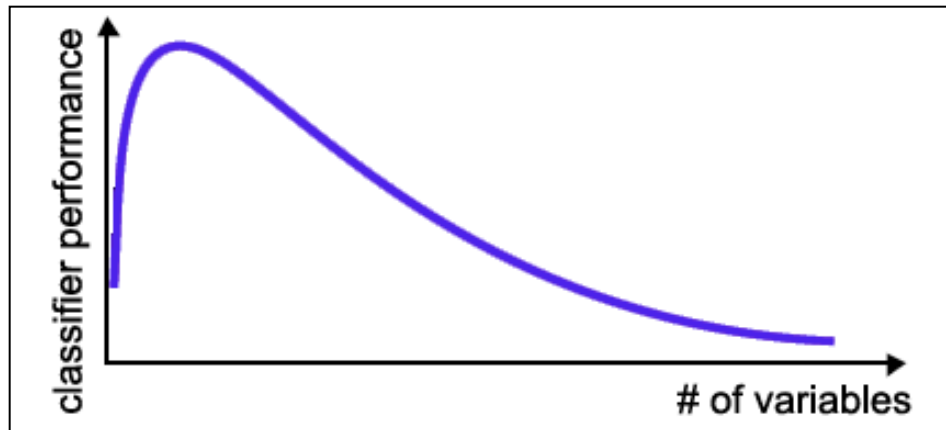
Why Feature Reduction?

- Many explored domains have hundreds to tens of thousands of variables/features with many irrelevant and redundant ones!
- In domains with many features the underlying probability distribution can be very complex and very hard to estimate (e.g. dependencies between variables)!
- Irrelevant and redundant features can “confuse” the models!
- Limited training data!
- Limited computational resources!
- **Curse of dimensionality!**

Curse of dimensionality

- The required number of samples (to achieve the same accuracy) grows **exponentially** with the number of variables!
- In practice: number of training examples is fixed!

=> the **model's** performance usually will degrade for a large number of features!

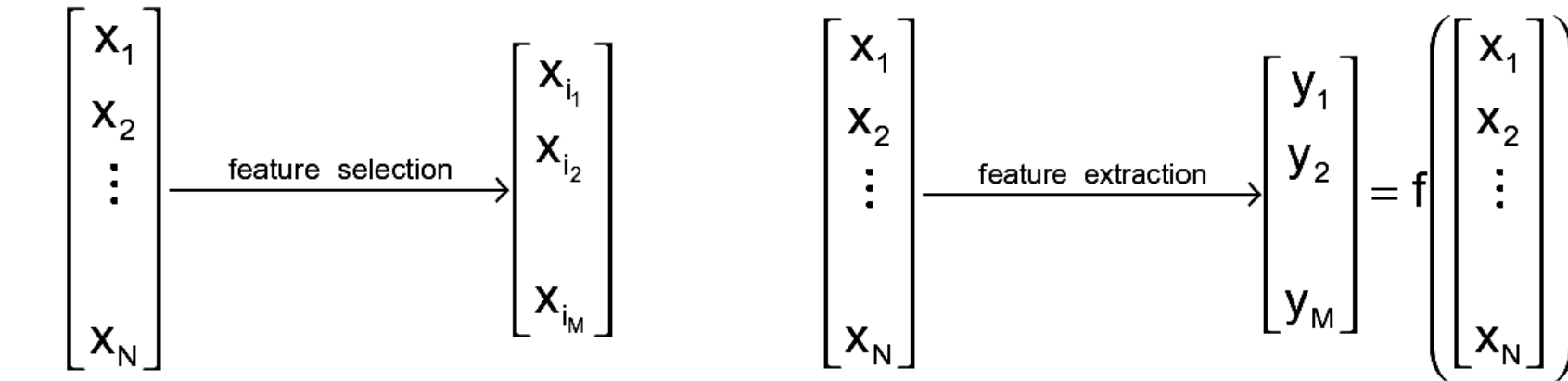


In many cases the information that is lost by discarding variables is made up for by a more accurate training in the lower-dimensional space!

Feature Selection vs Feature Extraction

- **Two general approaches for dimensionality reduction**

- Feature extraction: Transforming the existing features into a lower dimensional space
- Feature selection: Selecting a subset of the existing features without a transformation

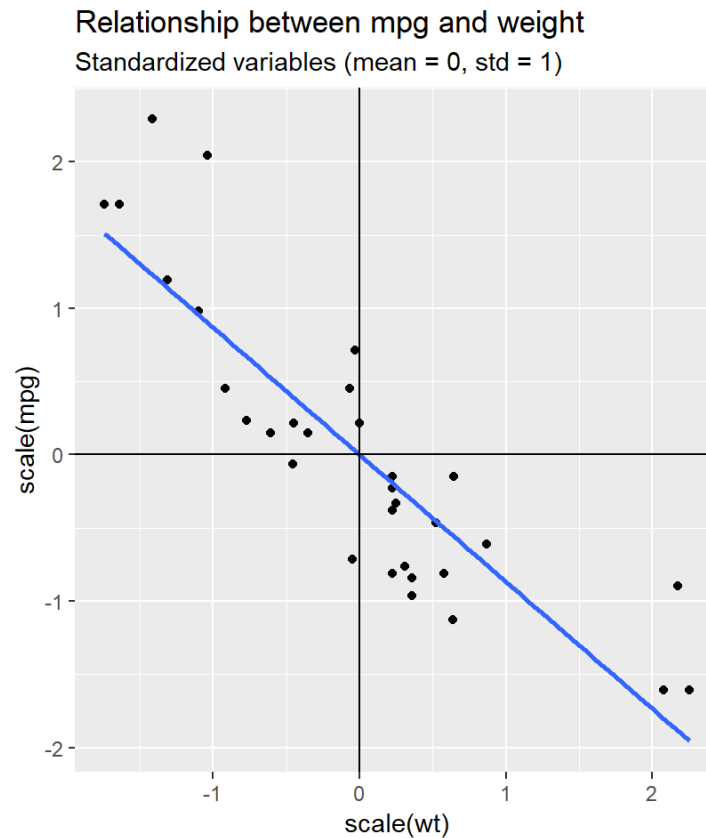


Feature Extraction

- Given a feature set $\mathbf{x}=\{\mathbf{x}_i \mid i=1\dots N\}$ find a mapping $\mathbf{y}=f(\mathbf{x}):R^N\rightarrow R^M$ with $M<N$, such that the transformed feature vector \mathbf{y}_i preserves (most of) the information or structure in R^N .
- An optimal mapping $\mathbf{y}=f(\mathbf{x})$ will be one that results in no increase in the minimum probability of error
 - **Methods:**
 - Linear Transformations: **Principal Component Analysis**

Principal Component Analysis

- Intuition:



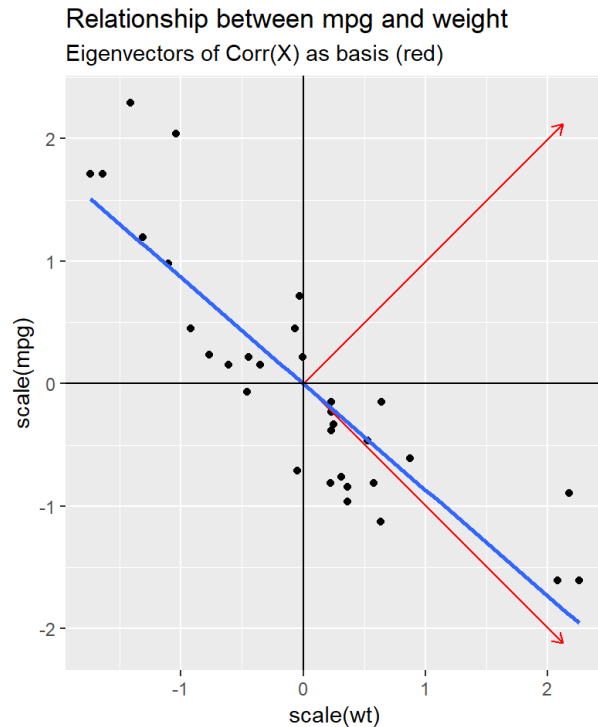
Linear relationship between weight (wt) and miles per gallon (mpg).

Most of the points fit rather tightly around the blue line, without deviating to much above or below.

So, if we had to summarise our data with one dimension only, it would be quite reasonable to pick a dimension somewhat similar to the blue line.

Principal Component Analysis

- To end up with:



One possible new coordinate system (red): the axis are equal to the eigenvectors of the covariance matrix of our dataset.

Now imagine we could **rotate** our coordinate axis so that in the new coordinate system the covariance matrix looks like this:

$$\text{Cov}(x, y) = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}$$

The two variables are **uncorrelate** given our new basis.

Principal Component Analysis

- PCA is a type of **linear transformation** that fits the dataset to a new coordinate system:
 - most **significant variance** is found on the **first coordinate**
 - each **subsequent coordinate** is orthogonal to the last and has **less variance**.
- In practice:
 - we transform a set of x correlated variables to a set of p uncorrelated principal components.

Principal Component Analysis

1. Form the data matrix \mathbf{X} containing your data;

\mathbf{X} is of size $K \times N$

2. Calculate the covariance matrix \mathbf{S} , based on \mathbf{X} ;

Capture redundancy into the dataset

3. Solve $\mathbf{S}\underline{\mathbf{e}} = \lambda\underline{\mathbf{e}}$ for the eigenvectors $\underline{\mathbf{e}}$ and eigenvalues λ

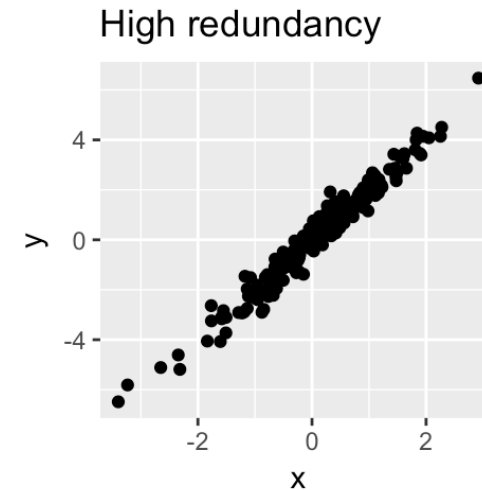
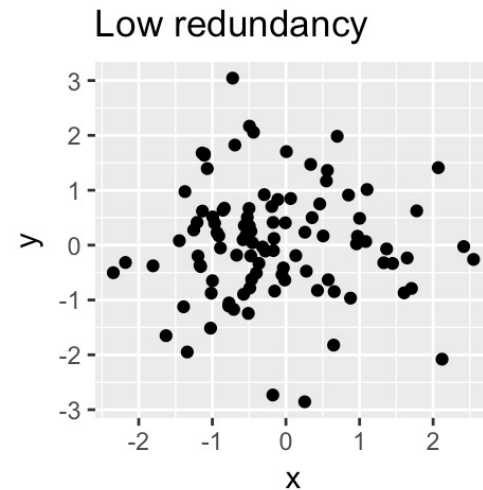
- **eigenvector** is a direction
- **eigenvalue** tells us how much variance there is in the data in that direction

4. Solve $\mathbf{P} = \mathbf{X}\underline{\mathbf{e}}$ to calculate the principal components (N PCs)

Transform some large number of variables into a smaller number of uncorrelated variables called principal components (PCs).

Understanding PCA

- Technically speaking, the **amount of variance** retained by each principal component is measured by the so-called **eigenvalue**.
- The PCA method is particularly useful when the variables within the data set are highly correlated.
- Correlation indicates that there is **redundancy** in the data. Due to this redundancy, PCA can be used to reduce the original variables into a smaller number of new variables (= principal components) explaining most of the variance in the original variables.



PCA in practice

- The dataset describes athletes' performance during two sporting events (Decstar and OlympicG). It contains 27 individuals described by 13 variables.

name	100m	Long.jump	//	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58		63.19	291.7	1	8217	Decastar
CLAY	10.76	7.4		60.15	301.5	2	8122	Decastar
Macey	10.89	7.47		58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74		55.39	278.05	5	8343	OlympicG
\\								
Zsivoczky	10.91	7.14		63.45	269.54	6	8287	OlympicG
Hernu	10.97	7.19		57.76	264.35	7	8237	OlympicG
Pogorelov	10.95	7.31		53.45	287.63	11	8084	OlympicG
Schoenbeck	10.9	7.3		60.89	278.82	12	8077	OlympicG
Barras	11.14	6.99		64.55	267.09	13	8067	OlympicG
KARPOV	11.02	7.3		50.31	300.2	3	8099	Decastar
WARNERS	11.11	7.6		51.77	278.1	6	8030	Decastar
Nool	10.8	7.53		61.33	276.33	8	8235	OlympicG
Drews	10.87	7.38		51.53	274.21	19	7926	OlympicG

PCA - Standardization

- In PCA, **variables** are often **scaled** (i.e. standardized). This is particularly recommended when variables are measured in different scales (e.g: kilometers, centimeters, ...);
 - otherwise, the PCA outputs obtained will be severely affected.
- The goal is to make the variables **comparable**. Generally variables are scaled to have standard deviation equal to one and mean equal to zero.
- When scaling variables, the data can be transformed as follow:

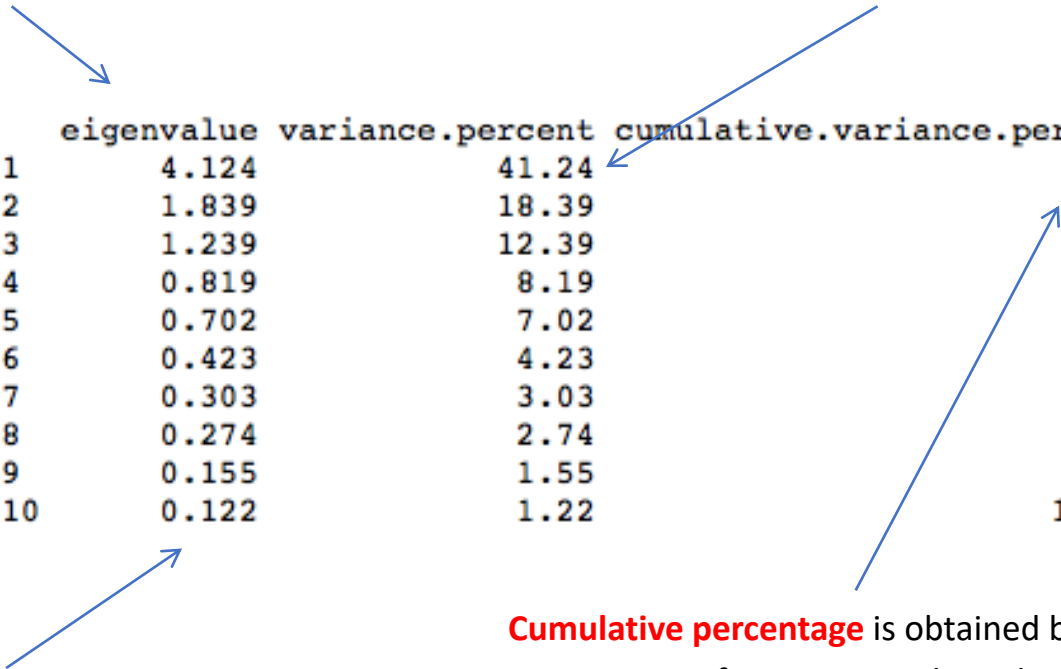
$$\frac{x_i - \text{mean}(x)}{\text{sd}(x)}$$

- Where $\text{mean}(x)$ is the mean of x values, and $\text{sd}(x)$ is the standard deviation (SD).

PCA - Eigenvalues

Proportion of variation explained by each eigenvalue.

Variance percentage: 41.24% of the variation is explained by this first eigenvalue.



##	eigenvalue	variance.percent	cumulative.variance.percent
## Dim.1	4.124	41.24	41.2
## Dim.2	1.839	18.39	59.6
## Dim.3	1.239	12.39	72.0
## Dim.4	0.819	8.19	80.2
## Dim.5	0.702	7.02	87.2
## Dim.6	0.423	4.23	91.5
## Dim.7	0.303	3.03	94.5
## Dim.8	0.274	2.74	97.2
## Dim.9	0.155	1.55	98.8
## Dim.10	0.122	1.22	100.0

The **sum** of all the **eigenvalues** give a total variance of 10.

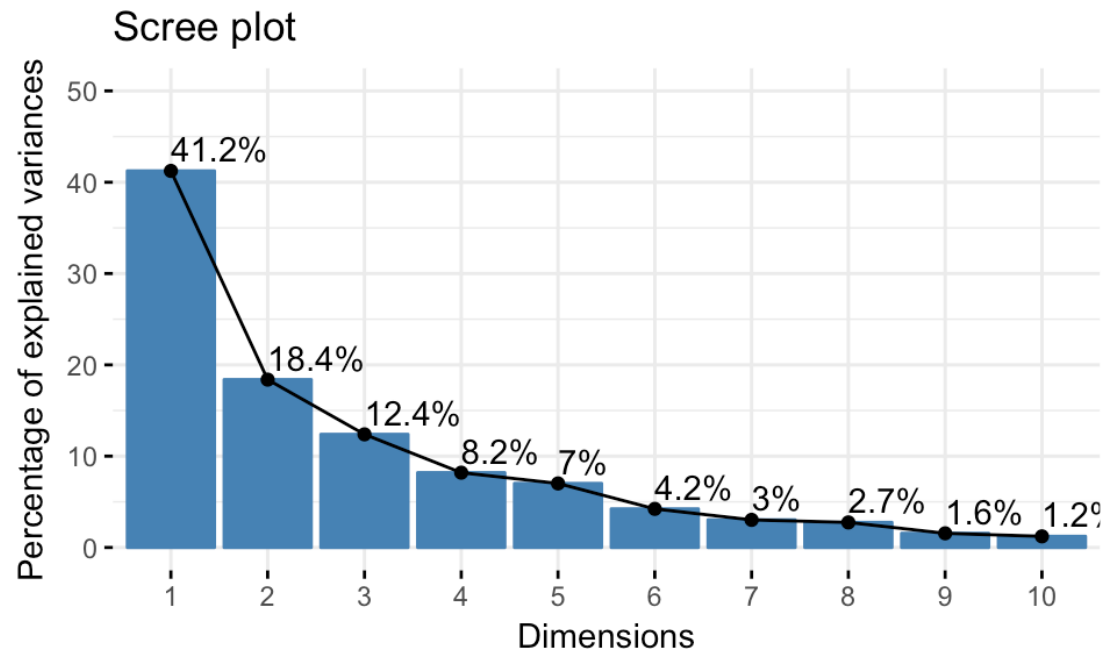
Cumulative percentage is obtained by adding the successive proportions of variation explained to obtain the running total. This means that about 59.627% of the variation is explained by the first two eigenvalues together.

PCA - Eigenvalues

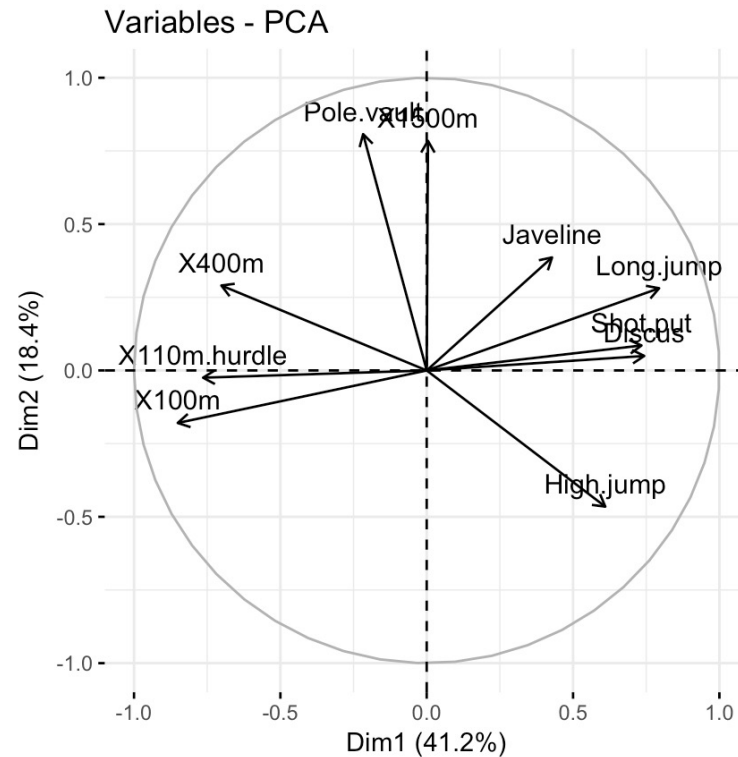
- Eigenvalues can be used to determine the number of principal components to retain after PCA (Kaiser 1961):
 - An **eigenvalue > 1** indicates that PCs account for more variance than accounted by one of the original variables in standardized data. This is commonly used as a cutoff point for which PCs are retained. This holds true only when the data are standardized.
 - You can **also limit the number of component** to that number that accounts for a certain fraction of the total variance. For example, if you are satisfied with 70% of the total variance explained then use the number of components to achieve that.

PCA - Eigenvalues

- In our example, the first three principal components explain 72% of the variation. This is an acceptably large percentage.
- An **alternative method** to determine the **number of principal components** is to look at a Scree Plot, which is the plot of eigenvalues ordered from largest to the smallest.

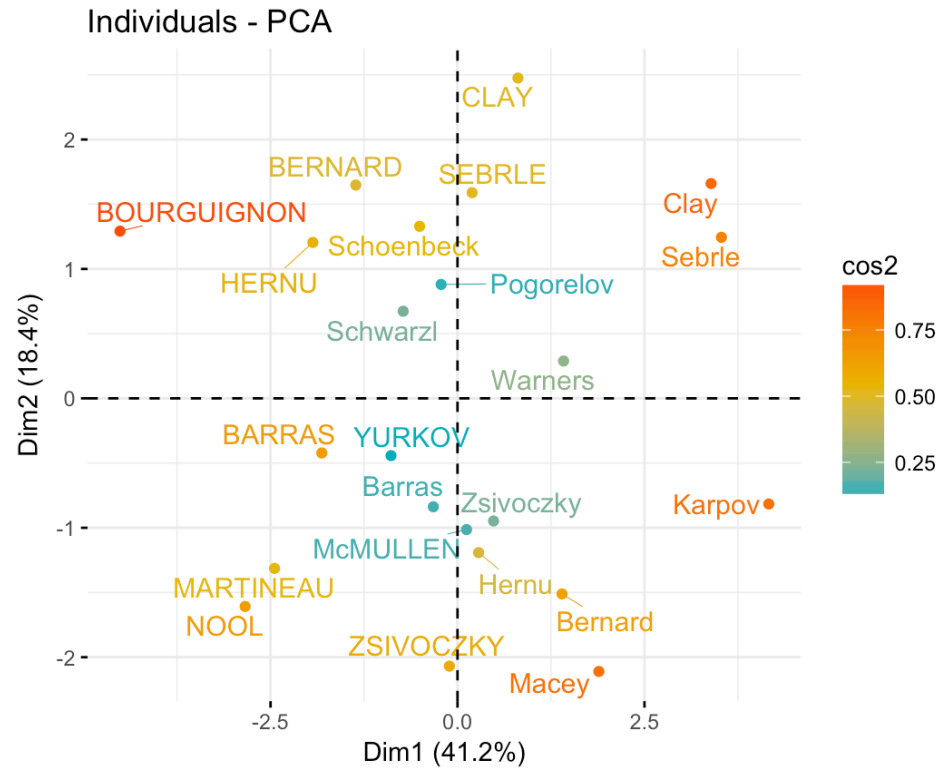


PCA - variables



- **Positively correlated** variables are grouped together.
- **Negatively correlated** variables are positioned on opposite sides of the plot origin (opposed quadrants).
- The **distance** between variables and the origin measures the quality of the variables. Variables that are away from the origin are well represented.

PCA - individuals



- A **high \cos^2** indicates a good representation of the individual on the principal component.
- A **low \cos^2** indicates that the individual is not perfectly represented by the PCs.