Introduction to Unsupervised Learning

Ryan Miller

Introduction

The big five personality traits are arguably the most accepted model of personality in academic psychology:

- Link: https://openpsychometrics.org/tests/IPIP-BFFM/

We'll look at roughly 20,000 responses to a 50-question big five personality test. Since there's not a single outcome variable of interest, so we'll use **unsupervised learning** methods.

Types of Unsupervised Learning

The two most widely used unsupervised learning approaches are **clustering** and **dimension reduction**:

- Clustering groups unlabeled data points based upon their similarities/differences
- Dimension reduction groups or eliminates redundant variables

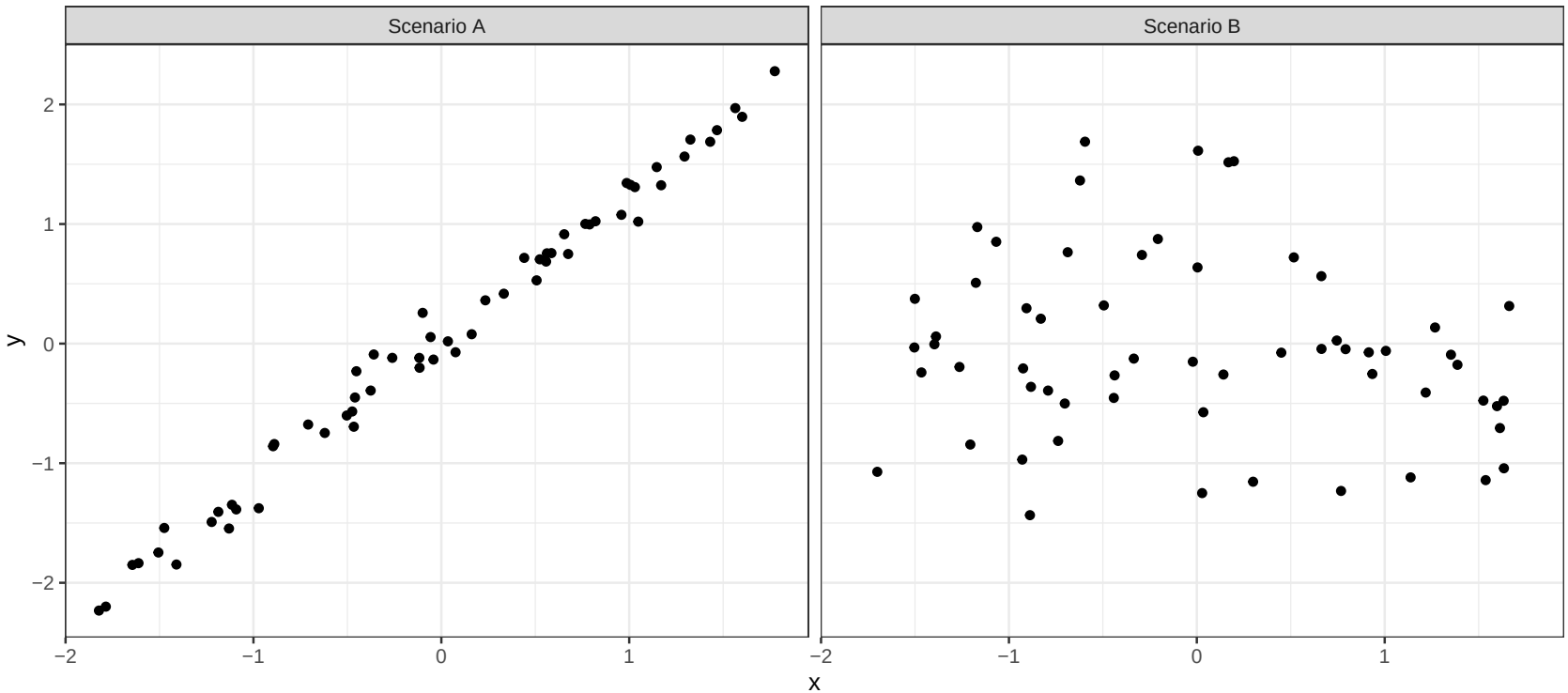
Principal Component Analysis

Principal Component Analysis (PCA) is a dimension reduction approach that creates a new data representation using linear transformations

- Principal components are sequentially constructed
- The first principal component is the direction of maximal variation within the data
  - The second is the direction of maximal variation that is *uncorrelated* with the first
  - Etc.

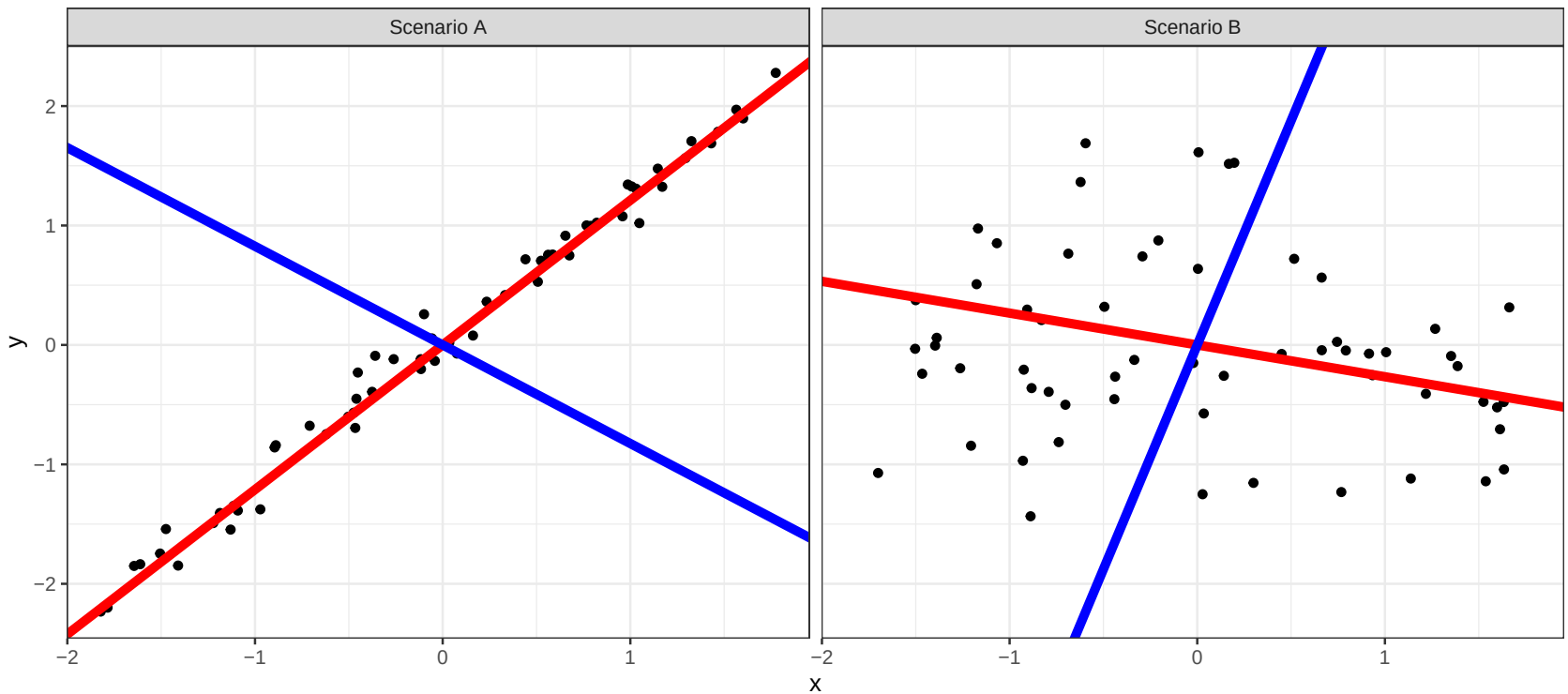
Principal Component Analysis (visual example)

What is the "direction of maximal variation" in each scenario?



Principal Component Analysis (visual example)

Shown below are the first principal component (red) and second principal component (blue)



Principal Component Analysis

Principal components can be expressed by linear combinations of the original variables:

$$PC_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$
$$PC_2 = \phi_{12}X_1 + \phi_{22}X_2 + \ldots + \phi_{p2}X_p$$
$$\ldots$$

- The coefficients, $\phi_{11}, \phi_{21}, \ldots$, are called **loadings**. They describe the contribution of a variable to a principal component.
  - The matrix containing these loadings is sometimes called the **rotation matrix**.

Principal Component Analysis (Big 5 Example)

Below are the top 10 loadings for PC1 (by absolute magnitude). We can see that the PC1 direction is essentially measuring Extroversion:

	PC1
E3	-0.2502747
E5	-0.2317793
E7	-0.2206402
E4	0.2057281
E10	0.2017135
E6	0.1985275
N10	0.1945291
A7	0.1875035
A10	-0.1873902
E1	-0.1835065

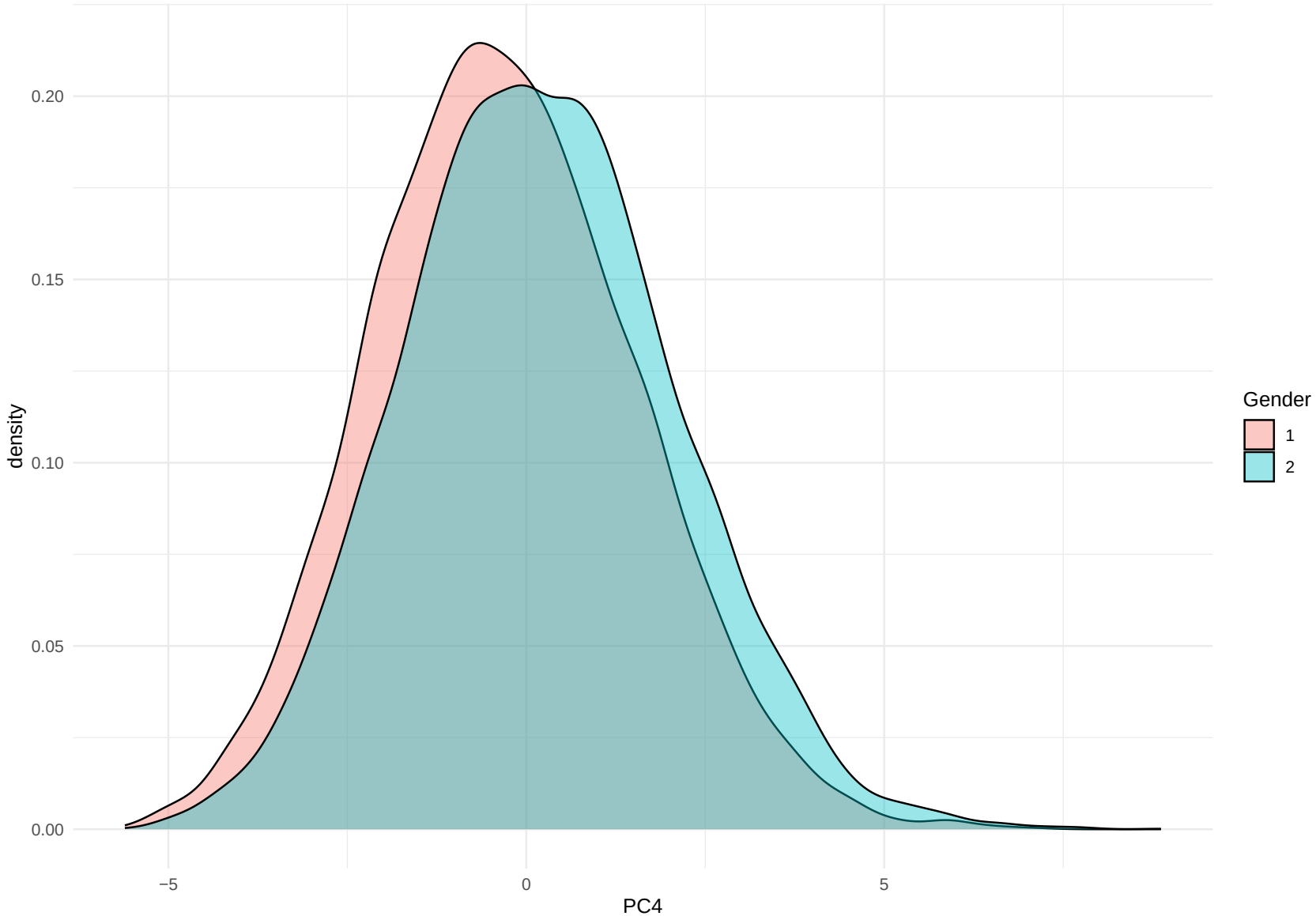
Principal Component Analysis (Big 5 Example)

The number of principal components equals the rank of the data matrix. So, we can solve for up to 50 of them for the Big 5 dataset. Shown below is PC4:

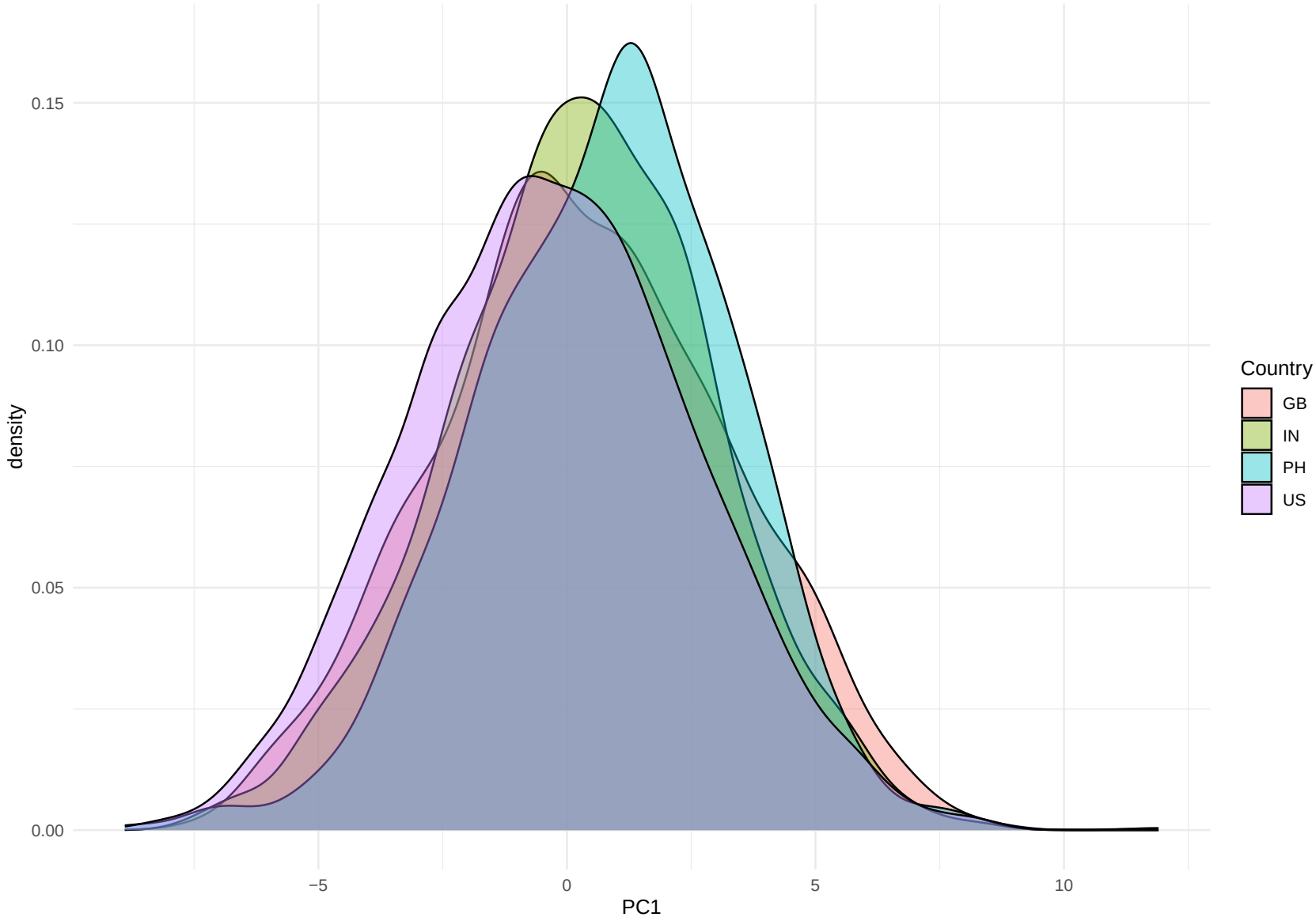
	PC4
O1	-0.3286096
O10	-0.3274160
O8	-0.3236670
O5	-0.2965985
O2	0.2911891
O3	-0.2881393
O7	-0.2692051
O6	0.2546234
O4	0.2510374
O9	-0.2135816

PCA Visualizations

PCA Visualizations



PCA Visualizations



How Many Components?

PCA is a factorization of the covariance matrix, $\mathbf{C} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$, into:

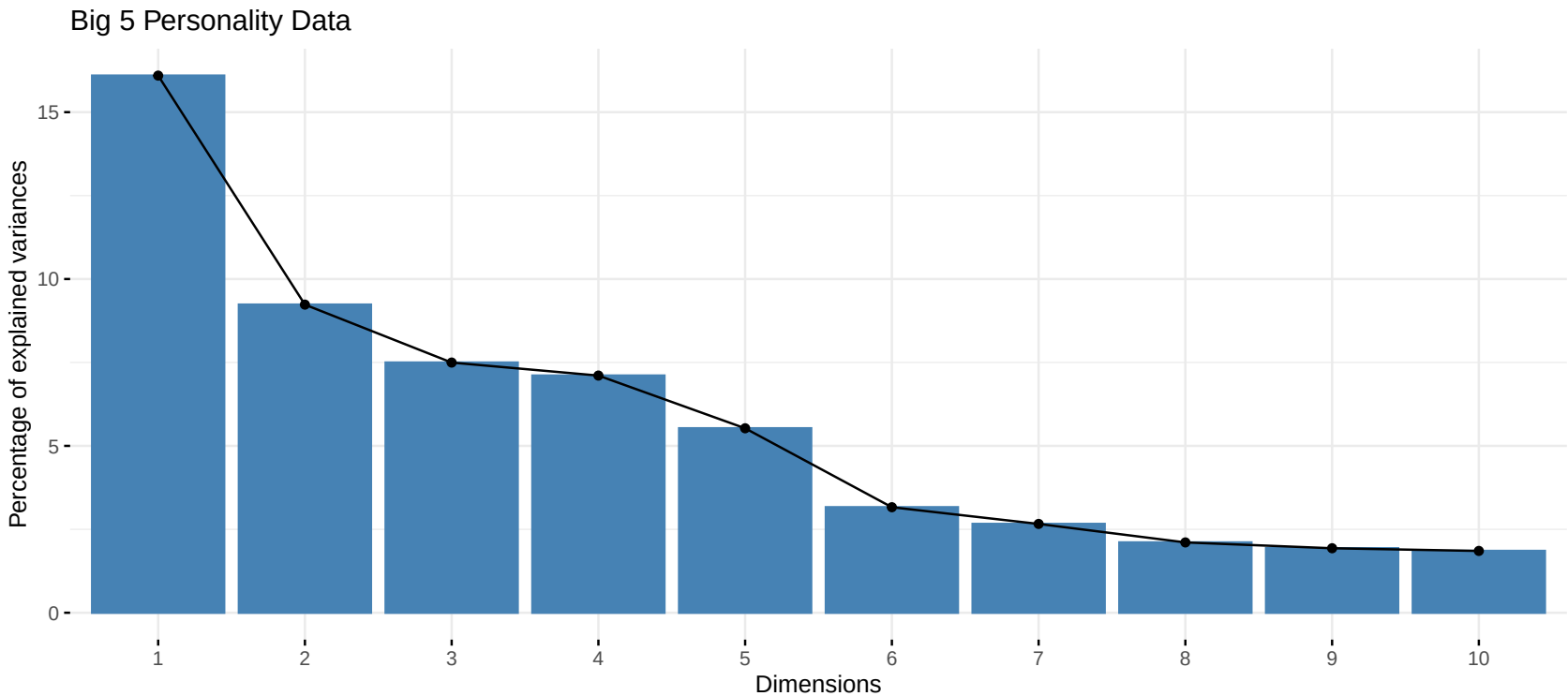
$$\mathbf{C} = \mathbf{V}\mathbf{L}\mathbf{V}^T$$

- $\mathbf{V}$ is a matrix of **eigenvectors**, while $\mathbf{L}$ is a diagonal matrix of **eigenvalues**
- The *columns* of $\mathbf{V}$ contain the *loadings* for each principal component
- The elements of $\mathbf{L}$ relate to the "variance explained" (importance) of each principal component

Note: If the data were standardized, $\mathbf{C}$ is the *correlation matrix*.

How Many Components?

We can use the elements in **L** (the eigenvalues of **C**) to assess the variation captured by each component:



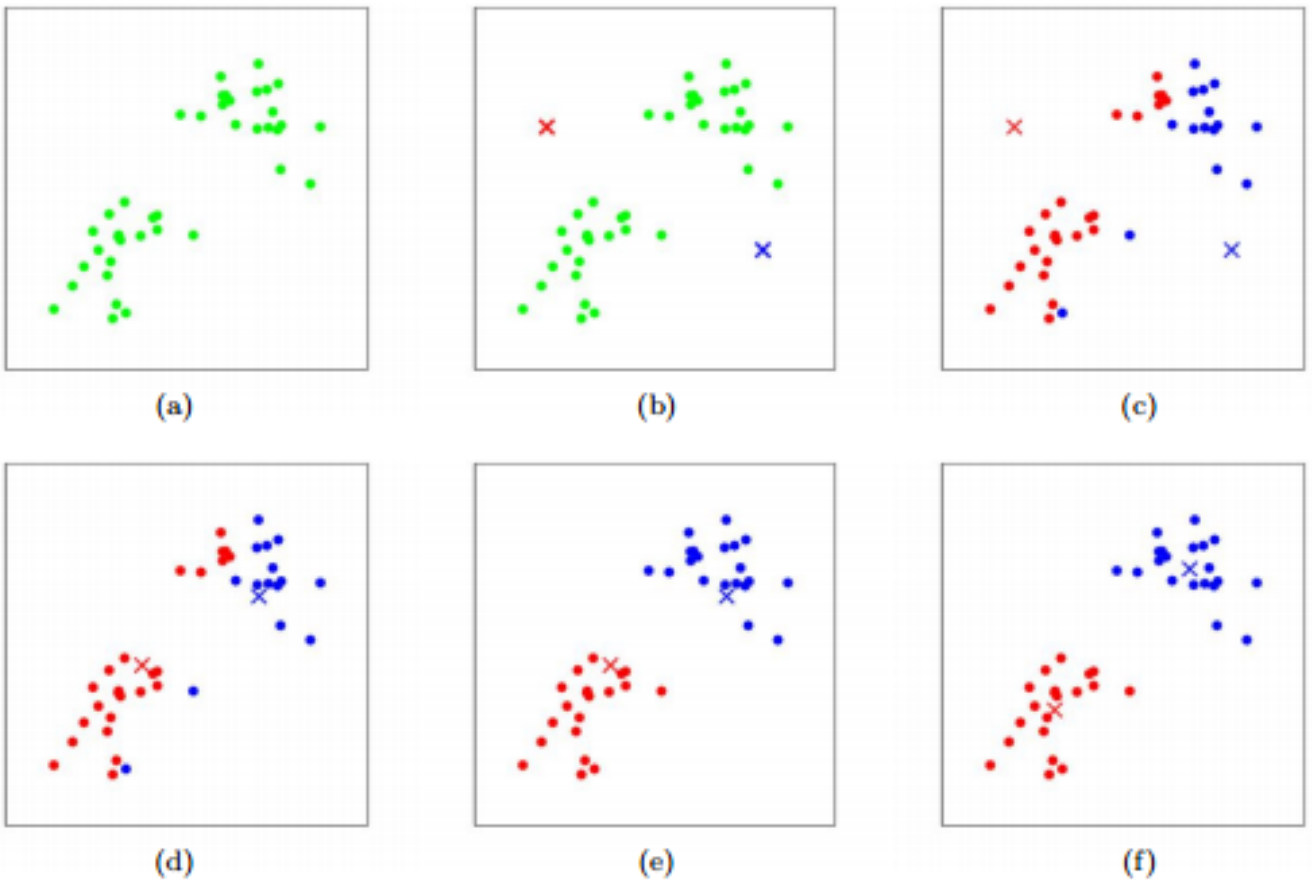
K-Means

- We can view PCA as a strategy for grouping or combining similar/related variables
  - *Clustering* describes methods for grouping combining *similar observations*
- k-means is a clustering algorithm that solves for k distinct cluster centers with a minimum mean distance to their members
  - That is, k-means solves for $\{\mu_1, \mu_2, \ldots, \mu_k\}$ that minimize:

$$\sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

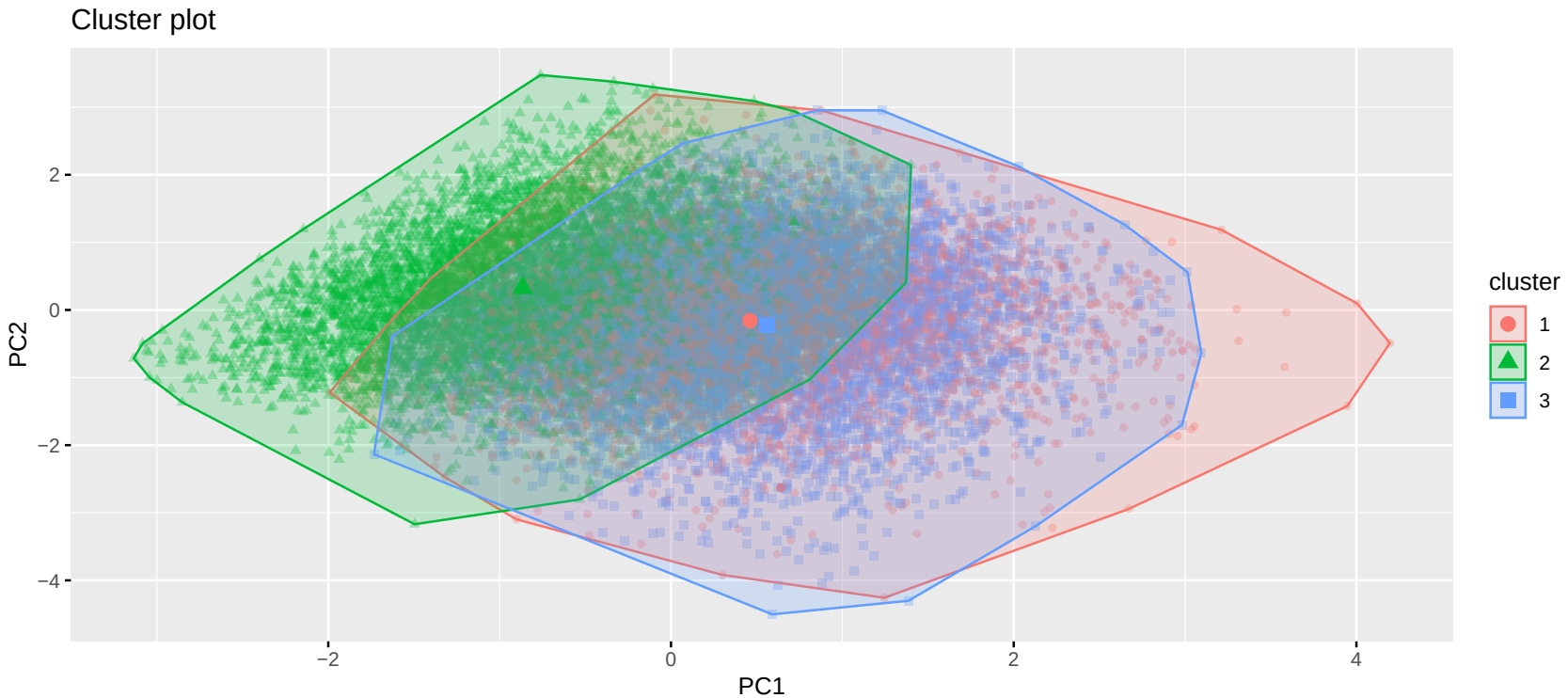
K-means

K-means iterates between two steps: grouping data-points with the nearest centroid and updating the centroids



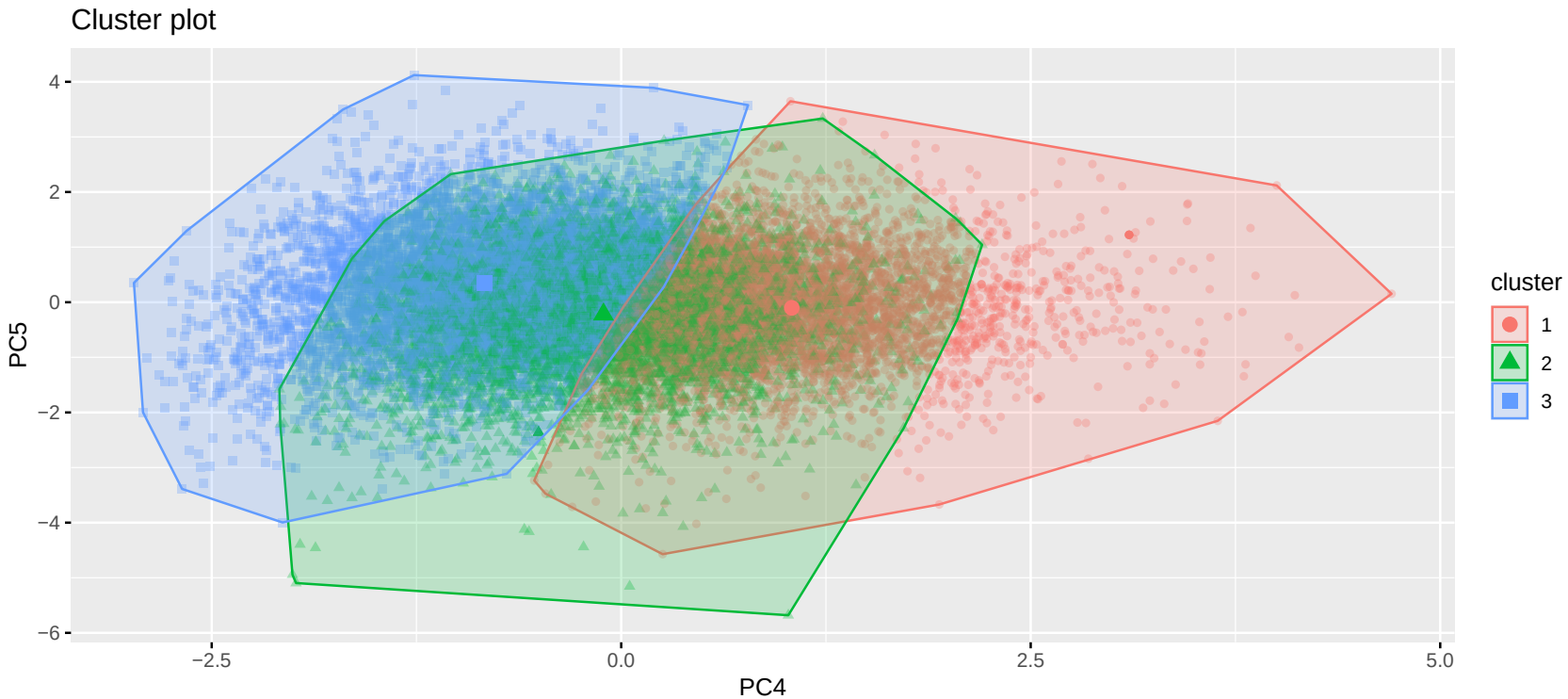
K-means on PC1 - PC5 (Big 5)

K-means applied to standardized scores of PC1 - PC5, shown in the PC1 and PC2 dimensions:



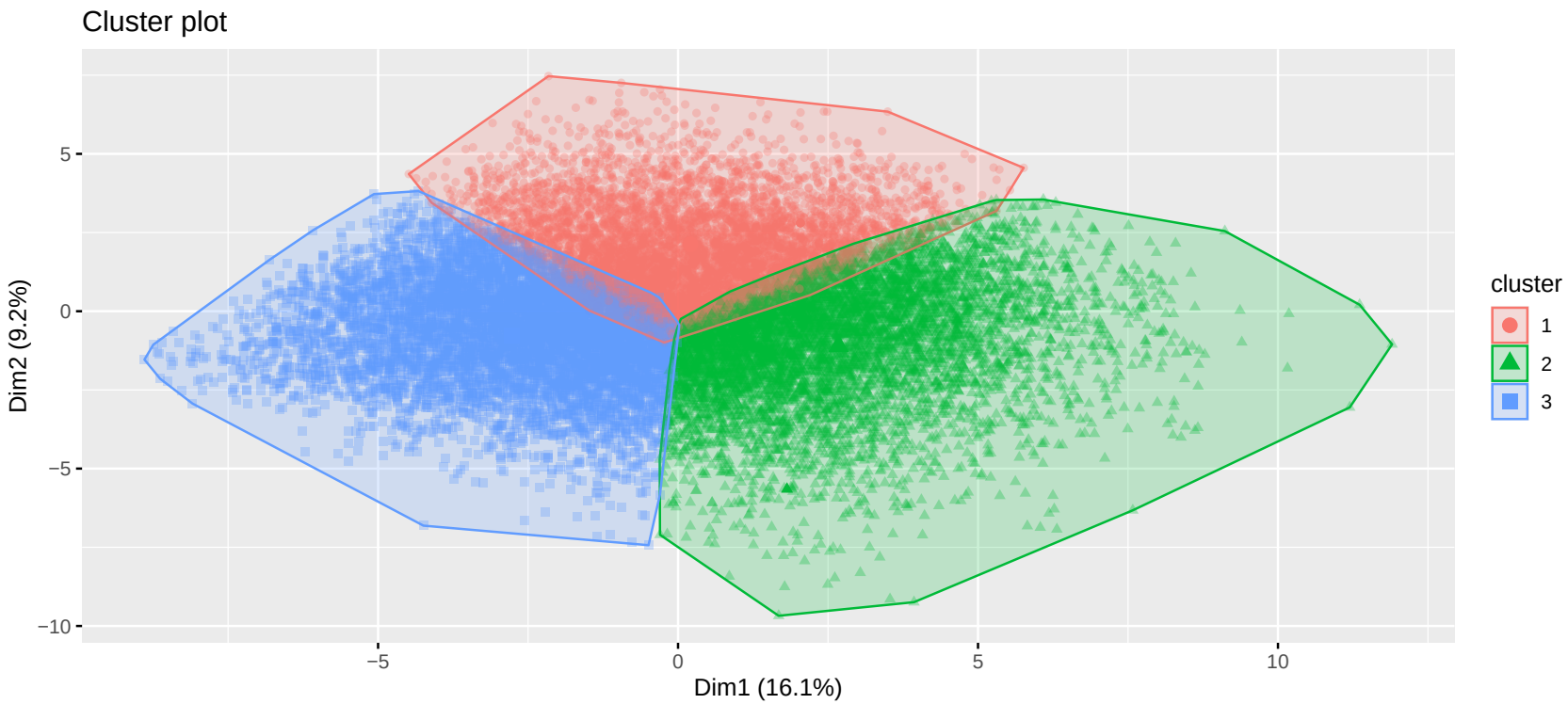
K-means on PC1 - PC5 (Big 5)

K-means applied to standardized scores of PC1 - PC5, clustering shown in the PC4 and PC5 dimensions:



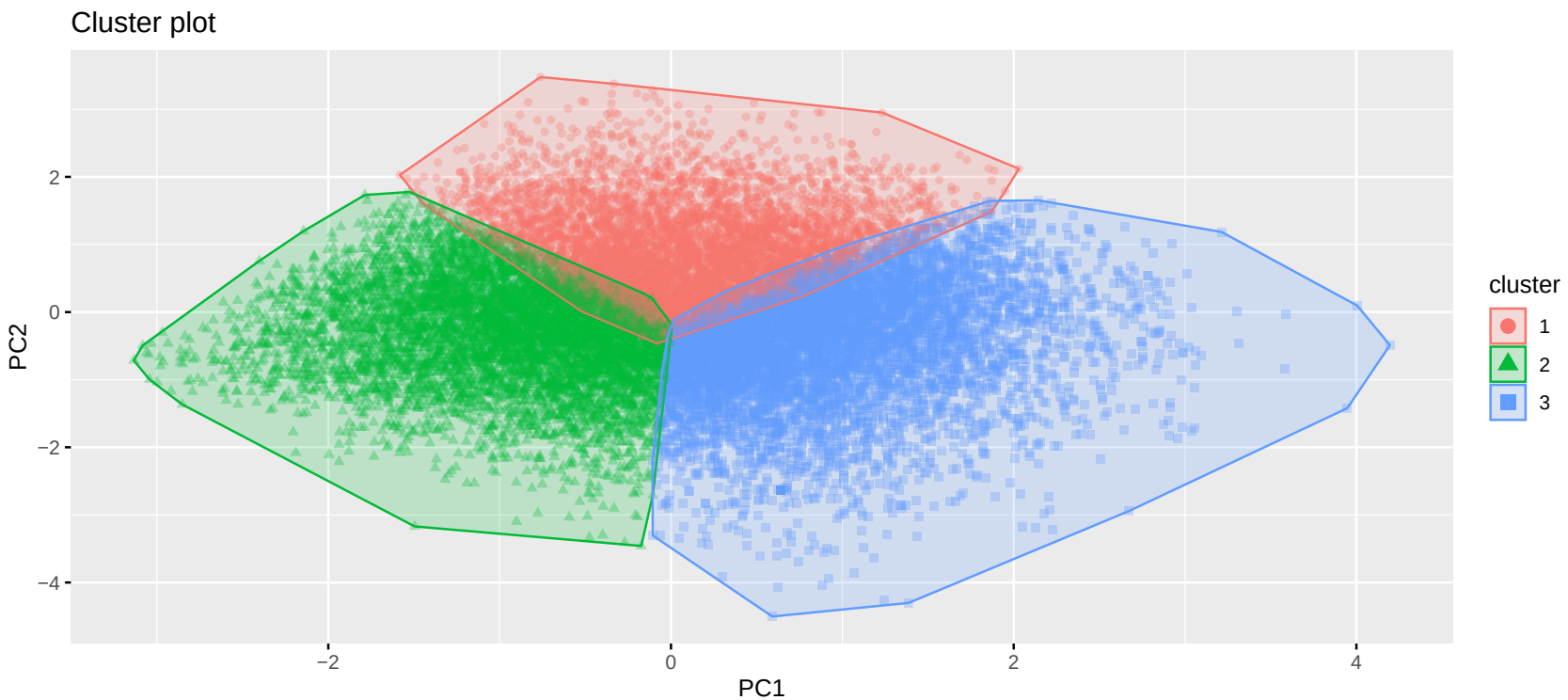
K-means on the Survey Items

K-means applied to the standardized survey items:



K-means on the Survey Items

K-means applied to PC1 - PC40 (no standardization):



DBSCAN

K-means has several downsides:

1. It requires the data analyst pre-specify a number of clusters (ie: value of k)
2. Cluster centroids can be heavily influenced by outliers
3. It does not scale well to high-dimensional data

A popular alternative that overcomes these limitations is the DBSCAN clustering algorithm

DBSCAN

- DBSCAN finds clusters using two parameters: a radius, `eps`, and a minimum number of data-points, `min_samples`
  - The algorithm surrounds each data-point with a hypersphere (or a circle in 2 dimensions)
- These hyperspheres are used to label each data-point as one of three types: *core points*, *border points*, and *noise*
  - Core points contain at least `min_samples` neighbors within their hypersphere
  - Border points contain at least 1 neighbor within their hypersphere
  - Noise points are at least `eps` away from any other data-point
- Cluster membership can then be determined by connected density regions

DBSCAN

The diagram below illustrates DBSCAN:

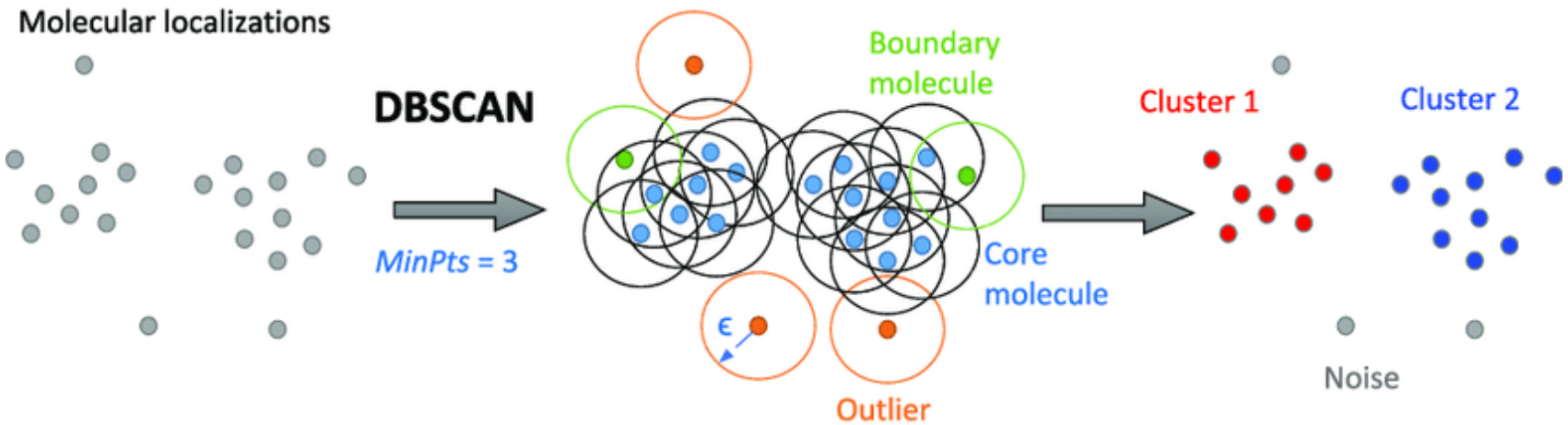


Image Source: A Review of Super-Resolution Single-Molecule Localization Microscopy Cluster Analysis and Quantification_Methods