Data, Inference & Applied Machine Learning

Course: 18-785

Patrick McSharry
patrick@mcsharry.net
www.mcsharry.net
Twitter: @patrickmcsharry

Fall 2024

ICT Center of Excellence
Carnegie Mellon University

Week 11

Course outline

Week	Description
1	Regression
2	Linear models
3	Nonlinear models
4	Supervised learning
5	Unsupervised learning
6	Ensemble approaches

Applied Machine Learning

WEEK 11A

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Multiple sources of information	10
2	Discussion	Noise reduction	10
3	Case study	Fetal Electrocardiogram	10
4	Analysis	Signal separation	20
5	Demo	Techniques for dimensionality reduction	20
6	Q&A	Questions and feedback	10

Unsupervised Learning

- Unsupervised learning refers to the analysis of a set of N features, which can be stored in the columns of a matrix X such that

$$X = [x_1 \mid \ldots, \mid x_N].$$

- Unsupervised learning requires us to analyse X without access to labels or responses y.
- This can be achieved by detecting patterns from within the X matrix itself.

Unsupervised learning goals

- To model the underlying structure or distribution in the entire dataset X.
- X is usually larger for unsupervised than supervised learning.
- In the case of a few dimensions, it may be possible to model p(X) non-parametrically at all values of X.
- In general, we seek low-dimensional manifolds within the X-space that represent high data density.
- Usually not possible to measure accuracy as there is no correct answer.

Poll

- Name some unsupervised methods

- **Slido.com**
- **#81797**

Unsupervised methods

- Filtering
- PCA
- ICA
- Similarity
- Novelty detection
- Hierarchical clustering
- k-means
- Gaussian mixture models
- Hidden Markov Models

Association Rules

- Association rules learning is probably the most widely used machine learning technique.
- From point-of-sale (POS) systems to web page usage mining, this method is often employed to examine large databases of transactions.
- It aims to detect the interesting relationships between variables in the database.
- The objective is to better understand human behavior through this analysis.

Retail

- The retail industry wishes to present its customers with offers on merchandise it believes will be of interest.
- In order to do that, though, it needs to know what you’ve bought previously and what other customers, similar to you, have bought.
- Brands such as Tesco and Target thrive on basket analysis to see what you’ve purchased previously.

Association Rule example

- For example, the rule {onions, potatoes}→ {burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat.
- Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing, product placements and cross-selling.

Apriori

- Agrawal (1995) introduced the Apriori algorithm to overcome the curse of dimensionality and succeed with a small number of passes over the data.
- Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as *candidate generation*), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

Example

- Consider the following database, where each row is a transaction and each cell is an individual item of the transaction:
- [alpha beta epsilon]
- [alpha beta theta]
- [alpha beta epsilon]
- [alpha beta theta]
- The association rules that can be determined from this database are the following:

  100% of sets with alpha also contain beta

  50% of sets with alpha, beta also have epsilon

  50% of sets with alpha, beta also have theta

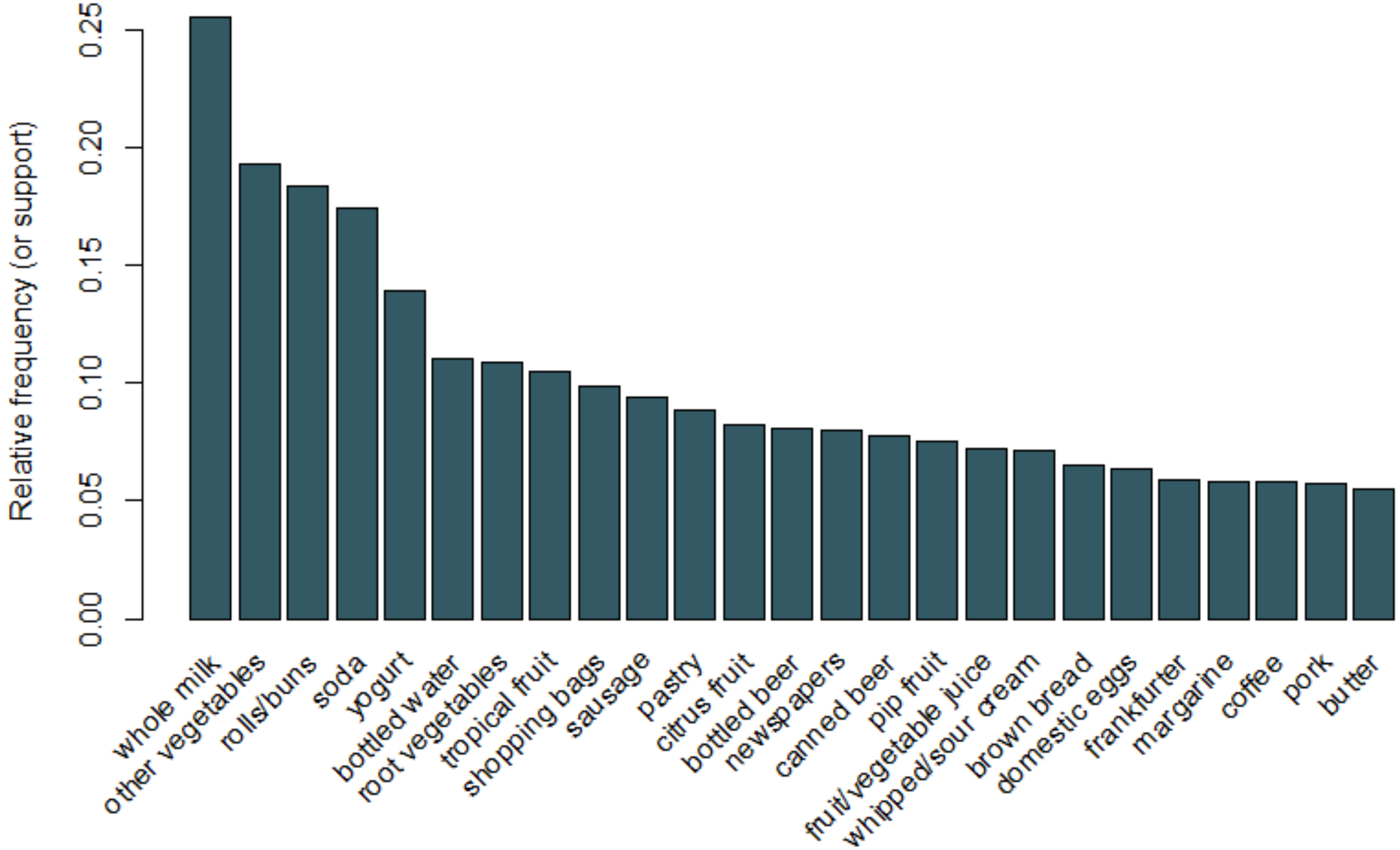
Beer and diapers

- A purported survey of behavior of supermarket shoppers discovered that customers (presumably young men) who buy diapers tend also to buy beer.
- This anecdote became popular as an example of how unexpected association rules might be found from everyday data.
- According to Daniel Powers: In 1992, Thomas Blischok, manager of a retail consulting group at Teradata, and his staff prepared an analysis of 1.2 million market baskets from about 25 Osco Drug stores. Database queries were developed to identify affinities. The analysis "did discover that between 5:00 and 7:00 p.m. that consumers bought beer and diapers". Osco managers did NOT exploit the beer and diapers relationship by moving the products closer together on the shelves.

Market Basket Analysis (MBA)

- **Support**: percentage of transactions that contain all of the items in a basket
- **Confidence**: probability that a transaction that contains the items on the left hand side of the rule also contains the item on the right hand side
- **Lift**: probability of all of the items in a rule occurring together (support) divided by the product of the probabilities of the items on the left and right hand side occurring as if there was no association between them.

Support



Source: Lynsey McColl
https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/

MBA rules

Rule	Support	Confidence	Lift
{instant food products, soda} → {hamburger meat}	0.001	0.632	19.00
{soda, popcorn} → {salty snacks}	0.001	0.632	16.70
{flour, baking powder} → {sugar}	0.001	0.556	16.41
{ham, processed cheese} → {white bread}	0.002	0.633	15.05
{whole milk, instant food products} → {hamburger meat}	0.002	0.500	15.04

Source: Lynsey McColl
https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/

Quiz

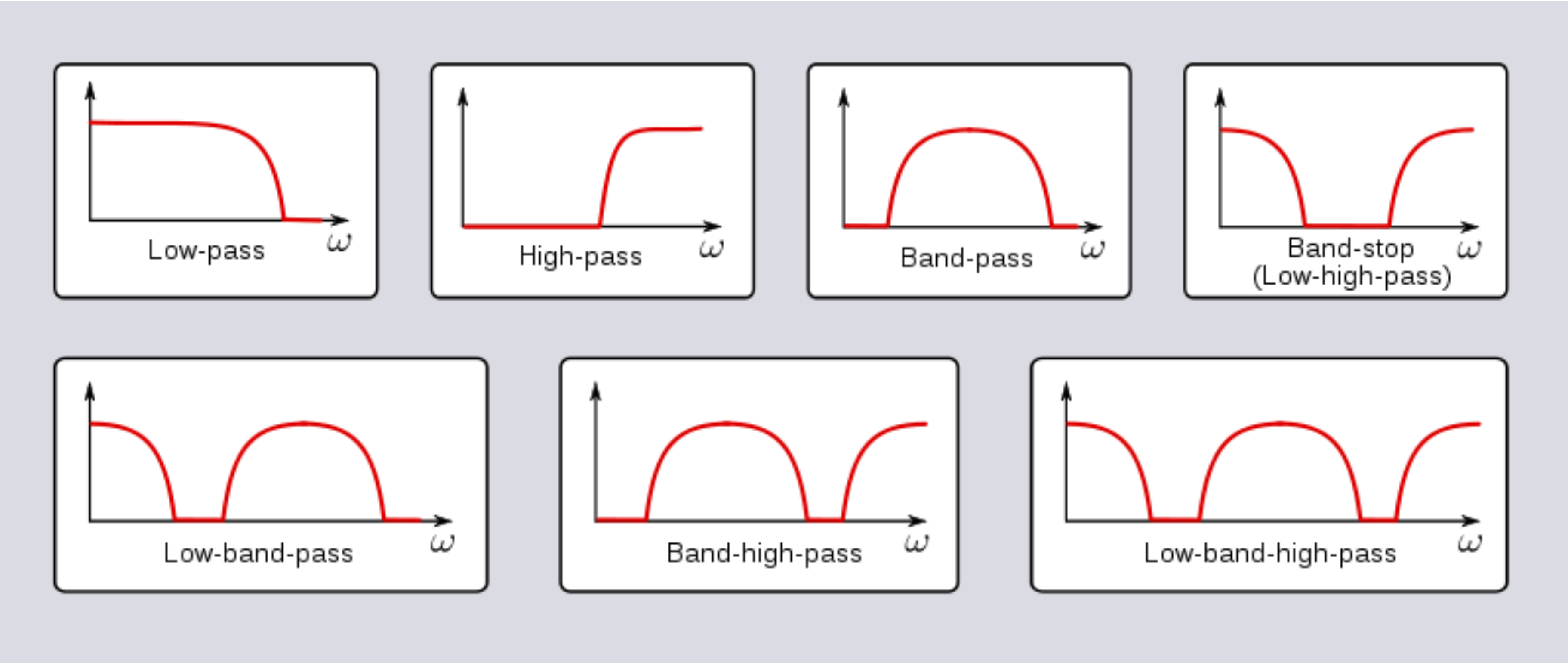
- Which of the following is a technique for data cleansing prior to constructing a predictive model?
  a) Data collection
  b) Data visualisation
  c) Data filtering
  d) Data streaming

- **Slido.com**
- **#81797**

Filtering

- Filtering refers to an algorithm that is applied to a signal to remove noise or unwanted components
- Examples of noise include measurement errors, truncation effects, senor malfunction
- Unwanted components could be seasonality effects, interference effects, or understood fluctuations due to external processes

Spectral filters



Source: wikipedia

Principal Component Analysis

- PCA performs an orthogonal transformation to provide linearly uncorrelated variables called principal components (PCs).
- The PCs are ordered in terms of the amount of variance captured with the first PC explaining the maximum variance.
- PCA also provides an eigenvalue spectrum where each value indicates the amount of variance represented by successive PCs.

PCA

- Construct an NxM data matrix **X** corresponding to N measurements and M variables
- The covariance matrix **C** of the data matrix **X** is subjected to an eigenvalue decomposition:

  $$\mathbf{C} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T$$

- where **V** is orthogonal ($\mathbf{V}\mathbf{V}^T=\mathbf{I}$) and $\boldsymbol{\Sigma}^2$ is a positive definite diagonal matrix
- Columns of **V**, denoted by $\mathbf{v}_m$, are the Mx1 eigenvectors of the MxM covariance matrix **C**
- Eigenvalues $\sigma_m{}^2$ on diagonal of $\boldsymbol{\Sigma}^2$ represent the variance associated with each eigenvector $\mathbf{v}_m$

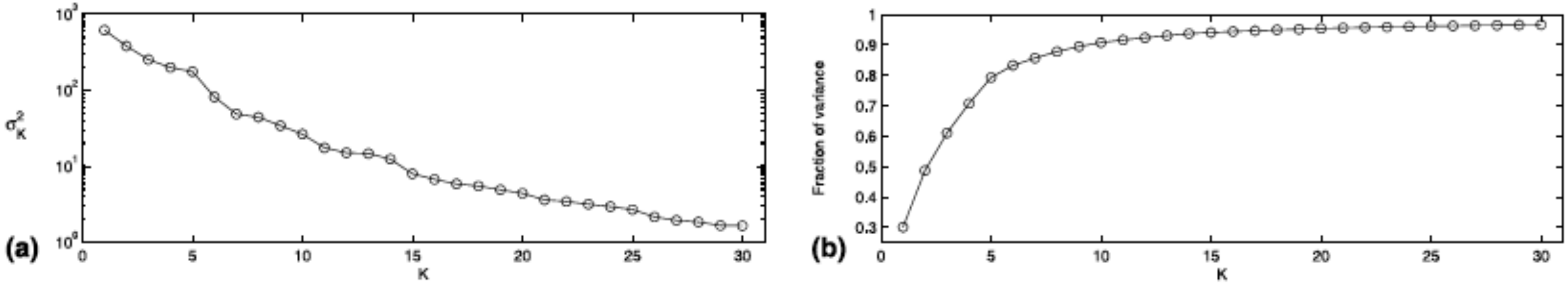
PCA

- The PCA of $\mathbf{X}$ can therefore be given as

  $$\mathbf{Y} = \mathbf{XV}$$

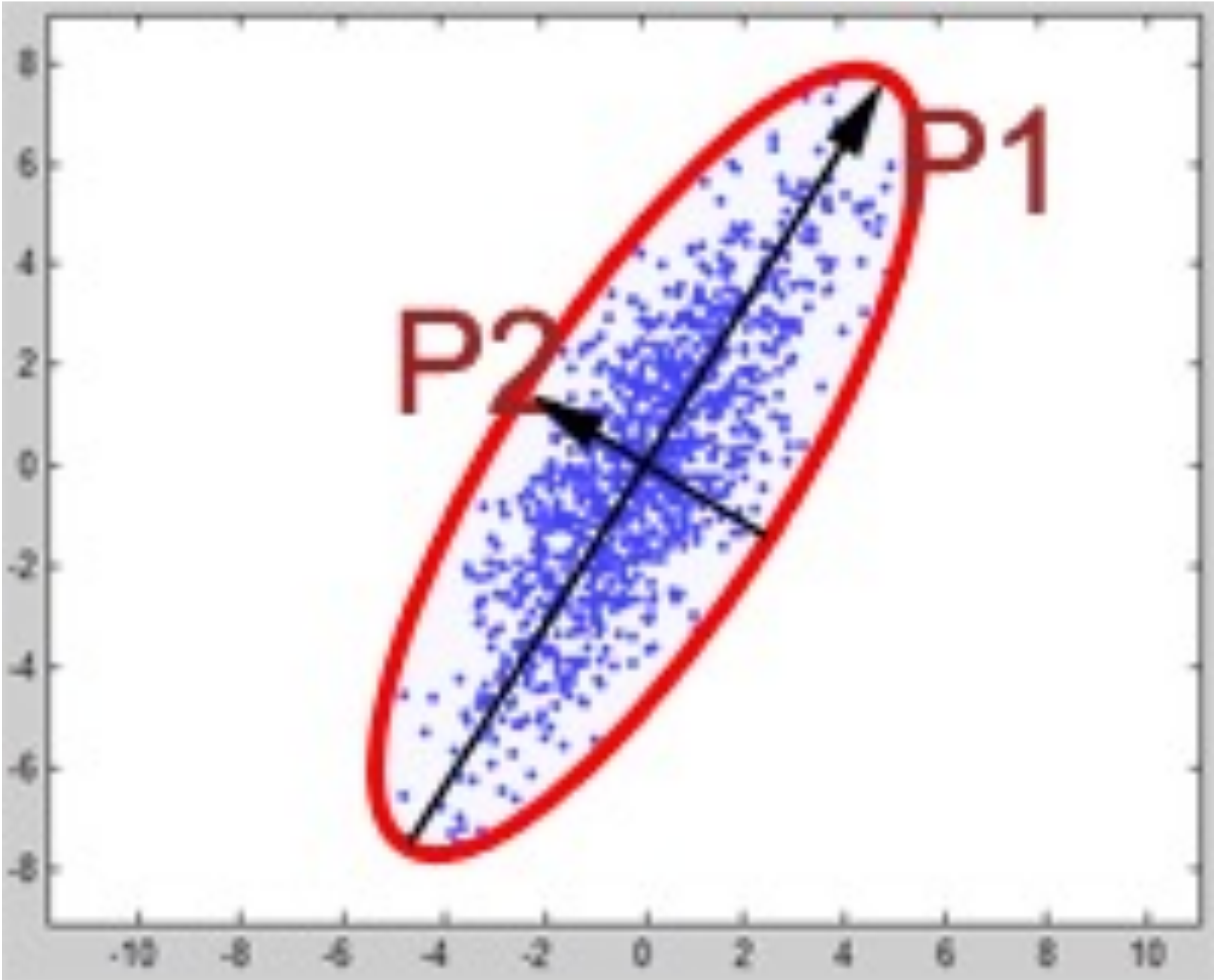
- where the vectors of weights or *loadings* $\mathbf{v}_m$ map each row vector $\mathbf{x}_m$ of $\mathbf{X}$ to a new vector of principal component *scores* $\mathbf{y}_m$
- The new variables in the columns of $\mathbf{Y}$ successively capture the maximum possible variance from the data matrix $\mathbf{X}$

PCA – noise reduction



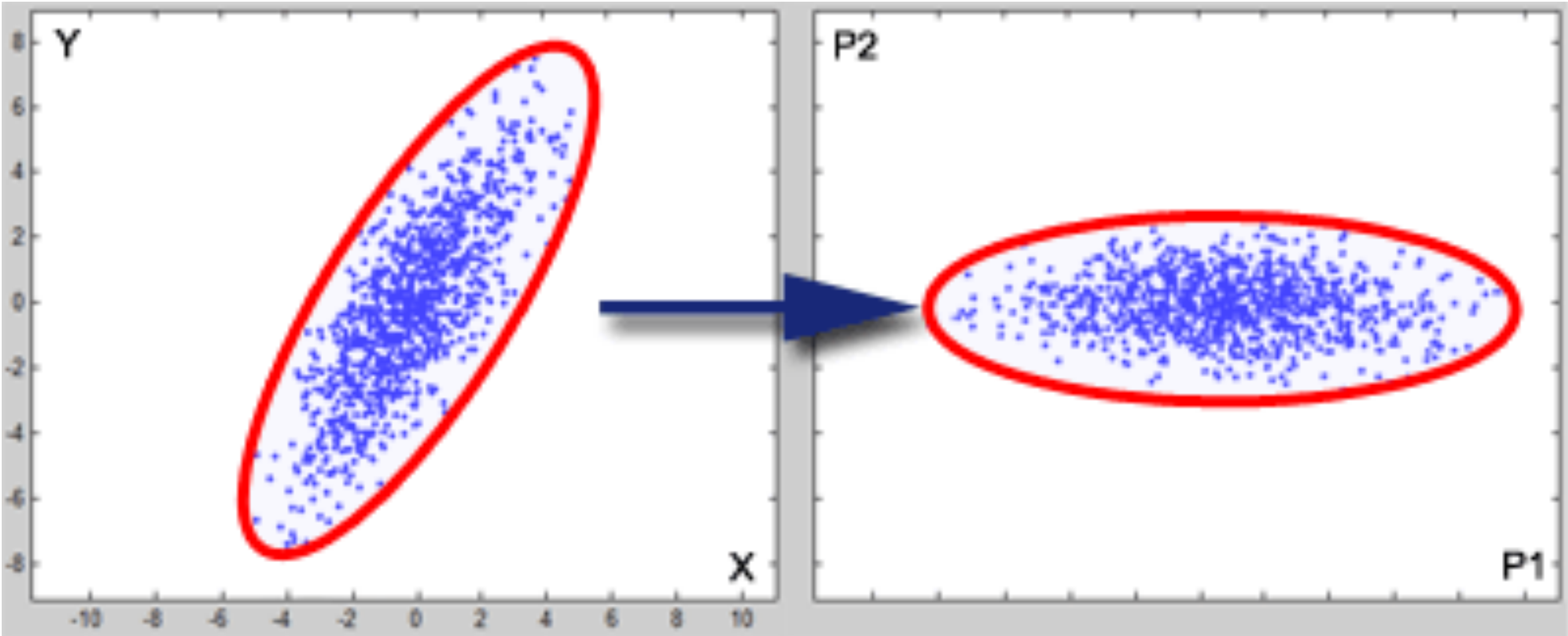
- PCA can be used for noise reduction where the contribution of higher components is deleted.
- Options are to include certain number of components or require fraction of variance.
- This assumes that important signals are related to high levels of variance and that noise corresponds to low levels of variance.

PCA: capturing variance



Source: www.peltarion.com

PCA: rotating axes



Source: www.peltarion.com

Poll

- PCA assumes that the signal and noise can be separated using an orthogonal set of basis functions. Is this always a valid assumption?
  - Yes
  - No
- **Slido.com**
- **#81797**

Nonnegative matrix factorization

- Nonnegative matrix factorization (*NMF*) is a dimension-reduction technique based on a low-rank approximation of the feature space.
- Besides providing a reduction in the number of features, NMF guarantees that the features are nonnegative, producing additive models that respect, for example, the nonnegativity of physical quantities.
- Given a nonnegative m-by-n matrix X and a positive integer $k < \min(m,n)$, NMF finds nonnegative m-by-k and k-by-n matrices W and H, respectively, that minimize the norm of the difference $X - WH$.
- W and H are thus approximate nonnegative factors of X such that $X \sim WH$.

Nonlinear systems and filtering

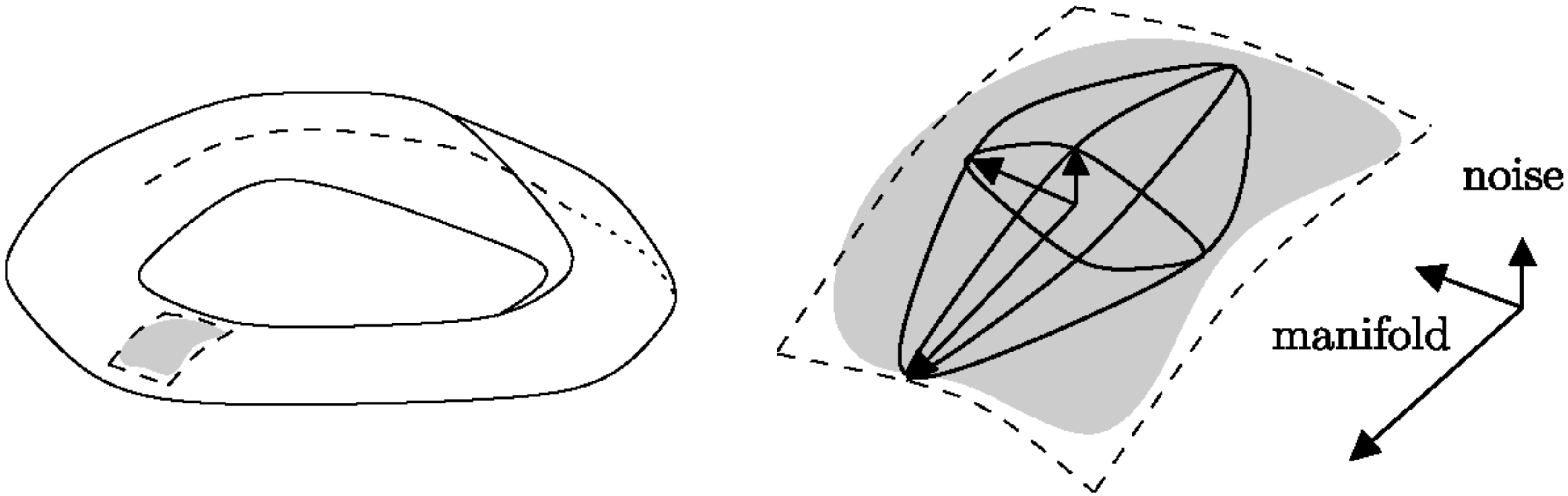
- Unfortunately nonlinear systems generate activity at all frequencies.
- This implies that it is usually not possible to separate the signal and noise using spectral approaches.
- It may be necessary to model either or both the signal and noise in parallel and use an iterative procedure.

Maternal and fetal ECG

- One example of an application of filtering is the extraction of a fetal ECG from the recording of an ECG from the mother's abdomen.
- The signal may be written as

$$s(t) = s_{maternal}(t) + s_{fetal}(t) + \varepsilon(t).$$

Nonlinear Noise Reduction (NNR)



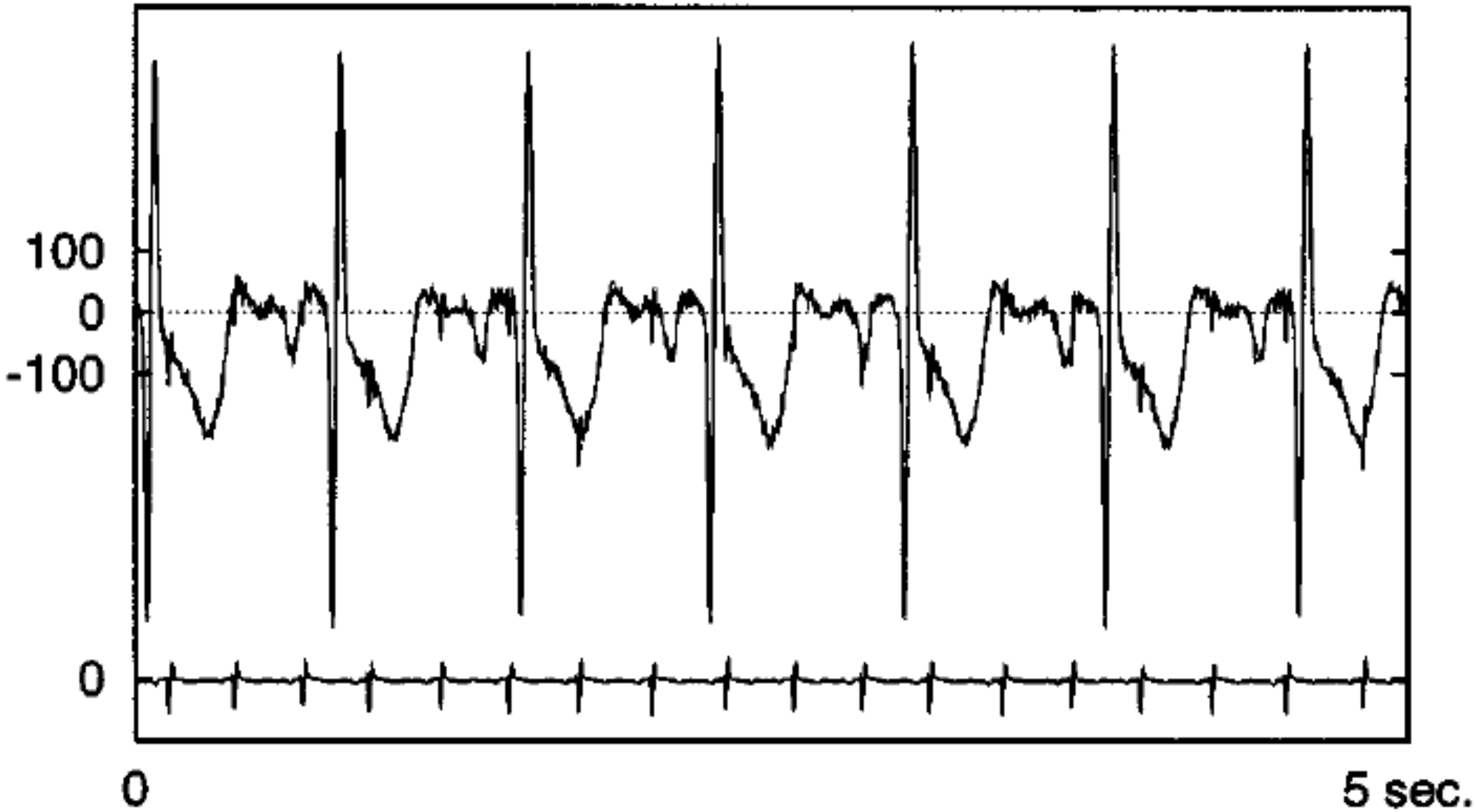
•Noise reduction using local projection

•For each point to be corrected, a neighbourhood is formed (grey) and approximated by an ellipsoid

•A 2D manifold embedded in a 3D space could be cleaned by projecting onto the first two principal components

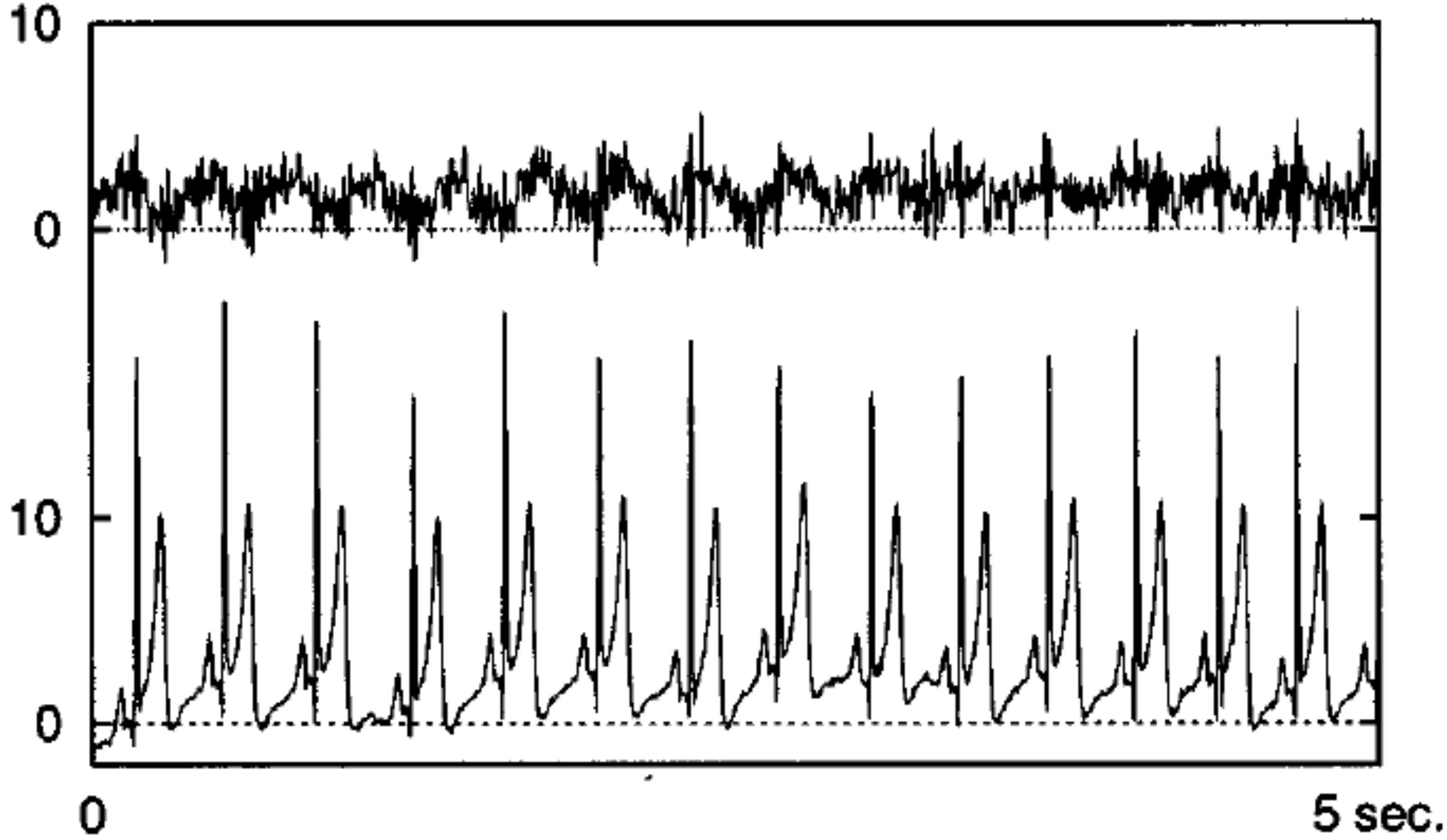
•Schreiber et al. [www.mpipks-dresden.mpg.de/~tisean]

Maternal and Fetal ECG

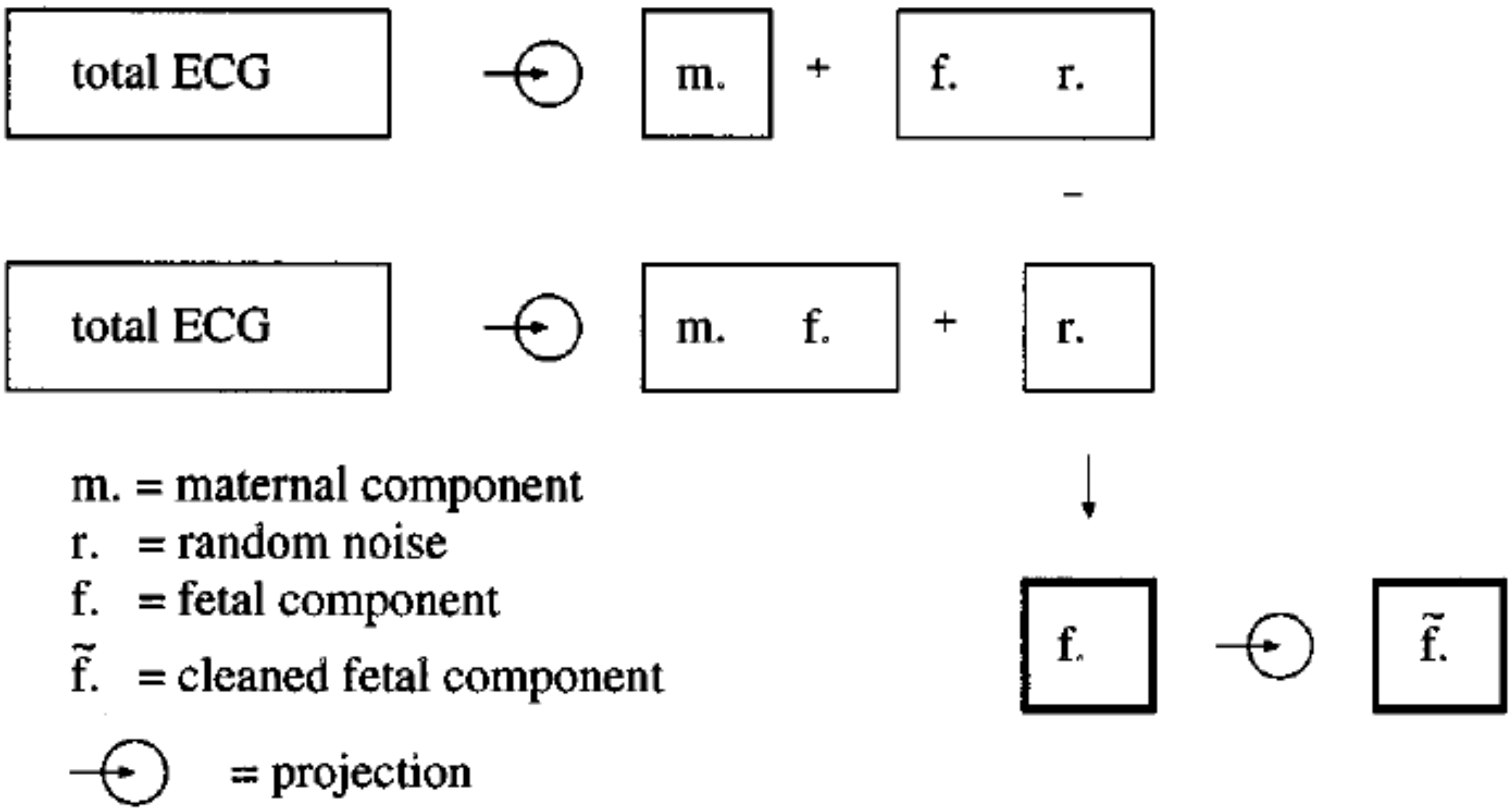


Schreiber & Kaplan (1996). Signal separation by nonlinear projections: The fetal electrocardiogram. Physical Review E 53(5): R4326: R4329

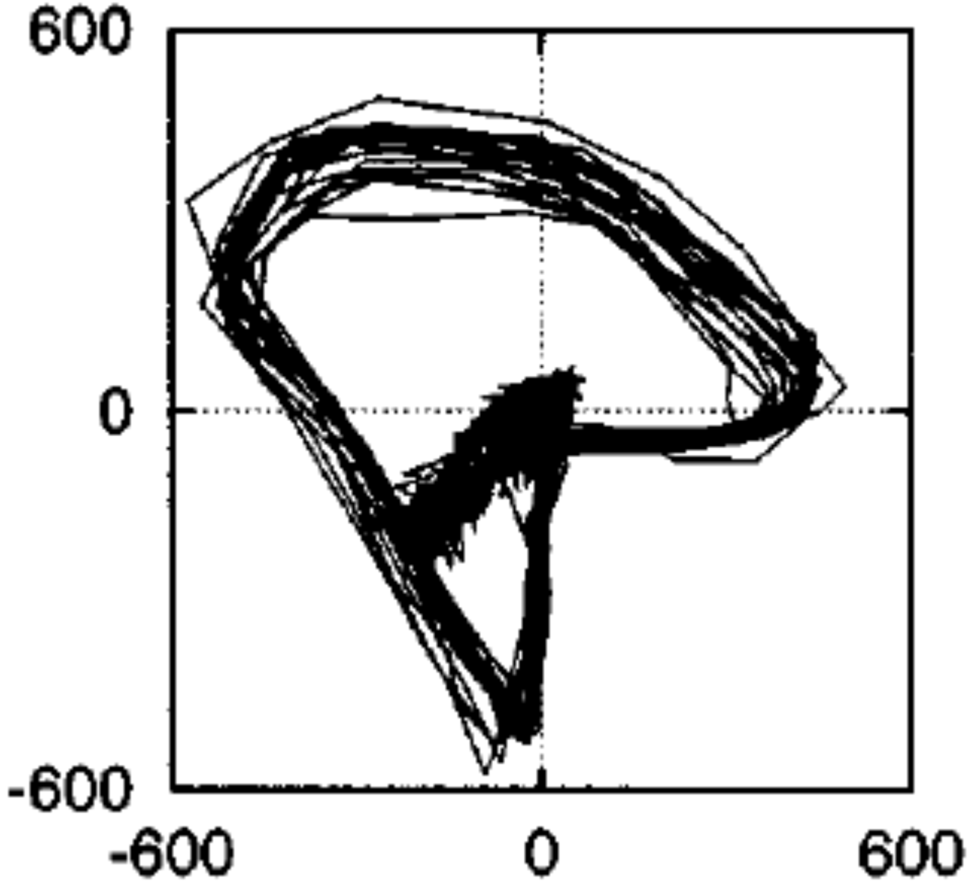
Wiener filtering



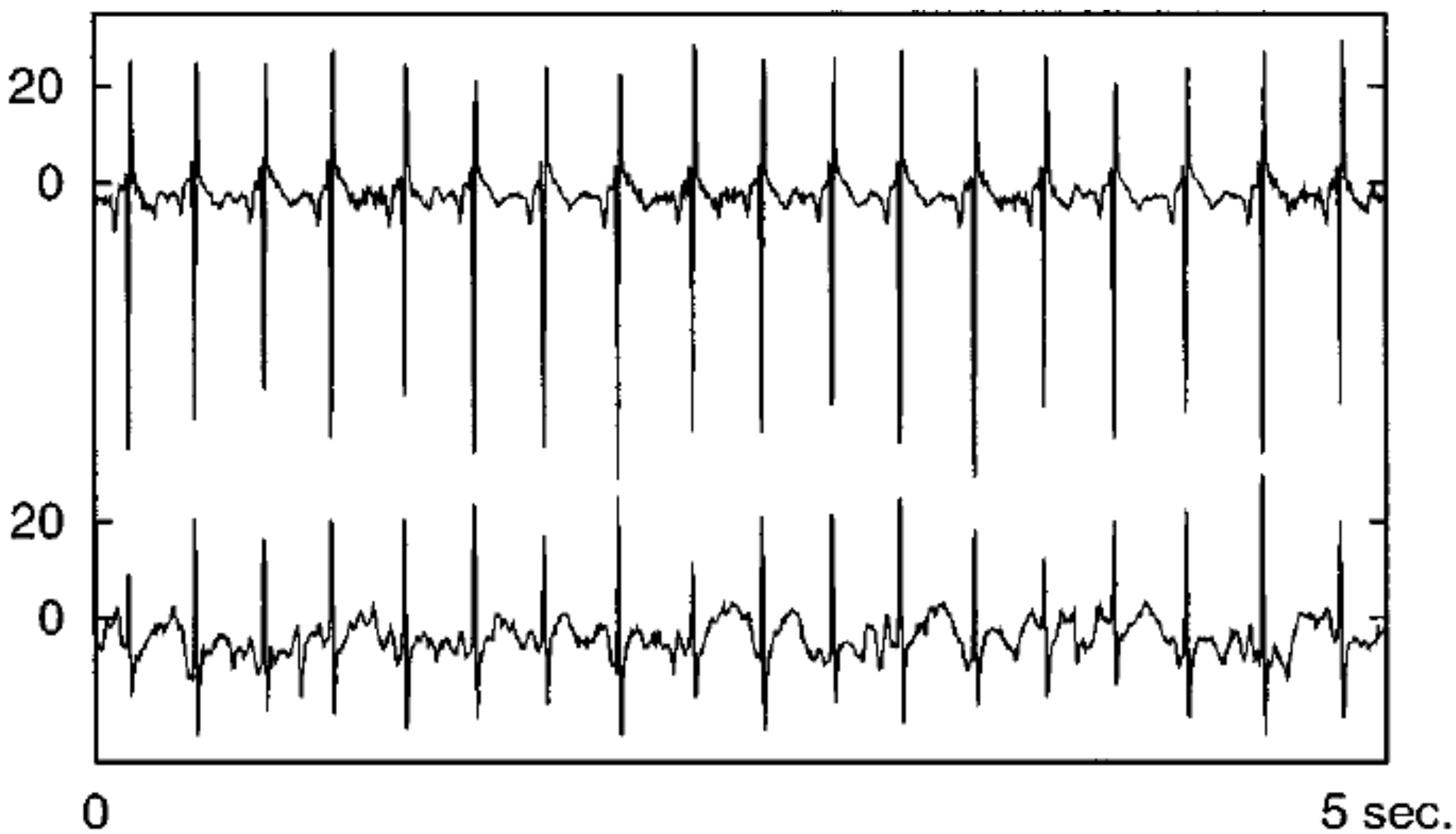
Flow chart


total ECG
m.
+
f. r.
−
total ECG
m. f.
+
r.
f.
f̃.
m. = maternal component
r. = random noise
f. = fetal component
f̃. = cleaned fetal component
= projection

Reconstruction of ECG



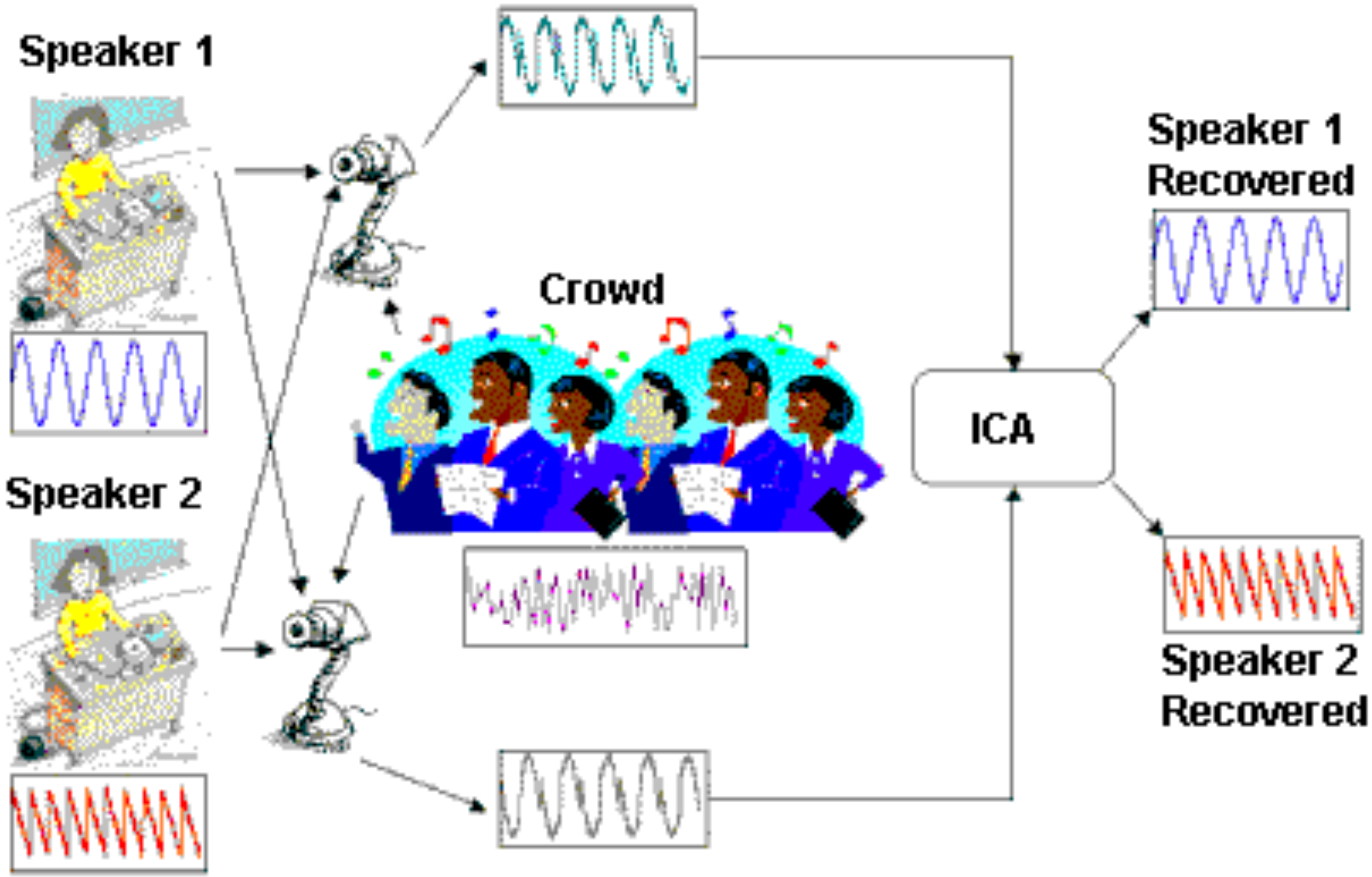
Original and extracted Fetal ECG



Independent component analysis

- Independent component analysis (ICA) is a technique for separating an observed dataset into independent sub-parts.
- ICA finds direction of maximum independence wheresas PCA finds direction of maximum variance.
- Suppose that data matrix **Y** results from the mixing of independent source signals in **X** according to **Y = BX**.
- ICA attempts to find a de-mixing matrix **W** such that **X = WY**.
- Iterative methods are used based on mutual information, entropy and kurtosis.

ICA application to speech



Source: www.statsoft.com

NNR versus ICA

- NNR had better results as measured by correlation ρ
- ICA had better results as measured by noise reduction factor χ
- NNR is better at recovering the morphology of the ECG and is less likely to distort the shape of the P, QRS and T waves
- ICA is better at recovering specific points on the ECG such as the R-peak, which is necessary for obtaining RR intervals

McSharry & Clifford (2004). A comparison of nonlinear noise reduction and independent component analysis using a realistic dynamical model of the electrocardiogram. SPIE.

Q&A

Applied Machine Learning

WEEK 11B

Today’s Lecture

No.	Activity	Description	Time
1	Challenge	Clustering and segmentation	10
2	Discussion	Eurovision	10
3	Case study	Customer Segmentation	10
4	Analysis	Dendrograms and K-means	20
5	Demo	Techniques for clustering	20
6	Q&A	Questions and feedback	10

Clustering

- Cluster analysis, also called segmentation analysis or taxonomy analysis, creates groups or clusters of data.
- Clusters are formed in such a way that objects in the same cluster are very similar and objects in different clusters are very distinct.
- Appropriate measures of similarity depend on the application being considered.

Poll: Odd one out?

A

Slido.com
#47142

B

D

C

Similarity

- Key to the goals of cluster analysis is the notion of the degree of similarity between the individual objects being clustered.
- A clustering method attempts to group the objects based on the definition of similarity provided.
- Objects in the same cluster should be similar.
- Objects belonging to different clusters should be dissimilar.

Proximity matrices

- Proximity can be based on judgment.
- Participants could be asked to judge by how much certain objects differ from one another.
- Dissimilarities can then be computed by averaging over the collection of such judgments.
- This type of data can be represented by an NxN matrix D, where N is the number of objects, and each element d_{ij} records the proximity between the ith and jth objects.
- This matrix is then provided as input to the clustering algorithm.

Dissimilarity and distance

- Most algorithms assume symmetric dissimilarity matrices.
- If the original matrix D is not symmetric, it can be replaced by $(D+D^T)/2$.
- Subjectively judged dissimilarities are seldom distances in the strict sense, since the triangle inequality $d_{ij} \leq d_{ik}+d_{jk}$, for all k in $\{1, . . . ,N\}$ does not hold.
- Therefore any algorithms that require distances cannot be used with such data.

Distances

- You can normalize the values in the data set before calculating the distance information.
- In a real world data set, variables can be measured against different scales.
- For example, one variable can measure Intelligence Quotient (IQ) test scores and another variable can measure head circumference.
- These discrepancies can distort the proximity calculations.
- Using the *zscore* function, you can convert all the values in the data set to use the same proportional scale.

Hierarchical clustering

- Hierarchical Clustering groups data over a variety of scales by creating a cluster tree or dendrogram.
- The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level.
- This allows you to decide the level or scale of clustering that is most appropriate for your application.

Hierarchical Clustering Algorithm

- Find the similarity or dissimilarity between every pair of objects in the data set:

  Distances between objects are computed using *pdist*.

- Group the objects into a binary, hierarchical cluster tree:

  The *linkage* function uses the distance information to determine the proximity of objects to each other and a *dendrogram* is formed.

- Determine where to cut the hierarchical tree into clusters:

  The *cluster* function to prune branches off the bottom of the hierarchical tree, and assign all the objects below each cut to a single cluster.

Eurovision Song Contest

Ireland won a record 7 times, **Luxembourg**, **France** and the **United Kingdom** 5 times. **Sweden** and the Netherlands won 4 times.
ABBA is the most successful Eurovision Song Contest winner.

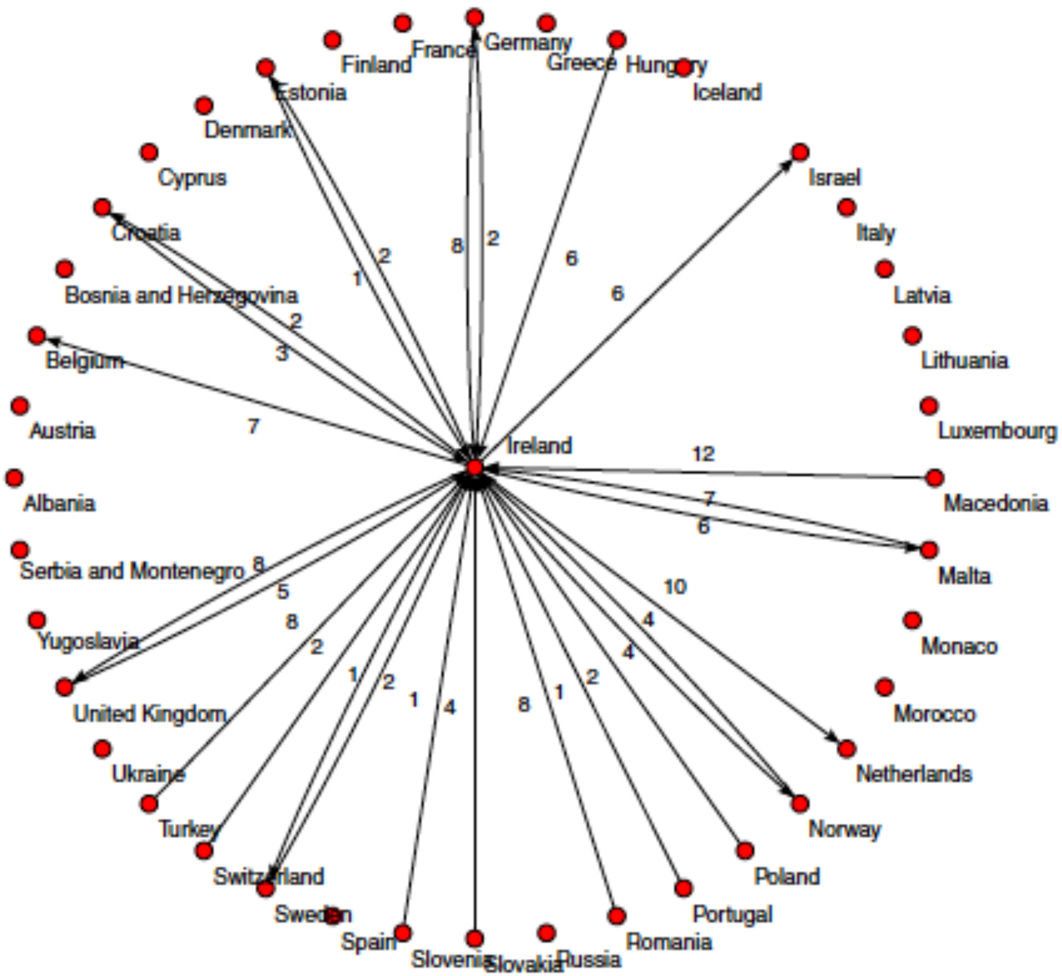
Eurovision Song Contest

- Is there any evidence of cliques in the Eurovision contest?
- Analysis of voting behavior in the contest over a period from 1992 to 2003 provides some answers.
- To measure the voting activity of each country, the average number of points assigned to each other entrant in the years in which they both compete was calculated.
- The closeness of each pair of countries can then be measured by comparing these data series using Pearson's correlation coefficient ρ_{ij}.
- If two particular countries assign exactly the same number of points to each other participating country, and thus possess identical data series, their Pearson coefficient will approach one.
- These correlation coefficients are then transformed to produce a distance:

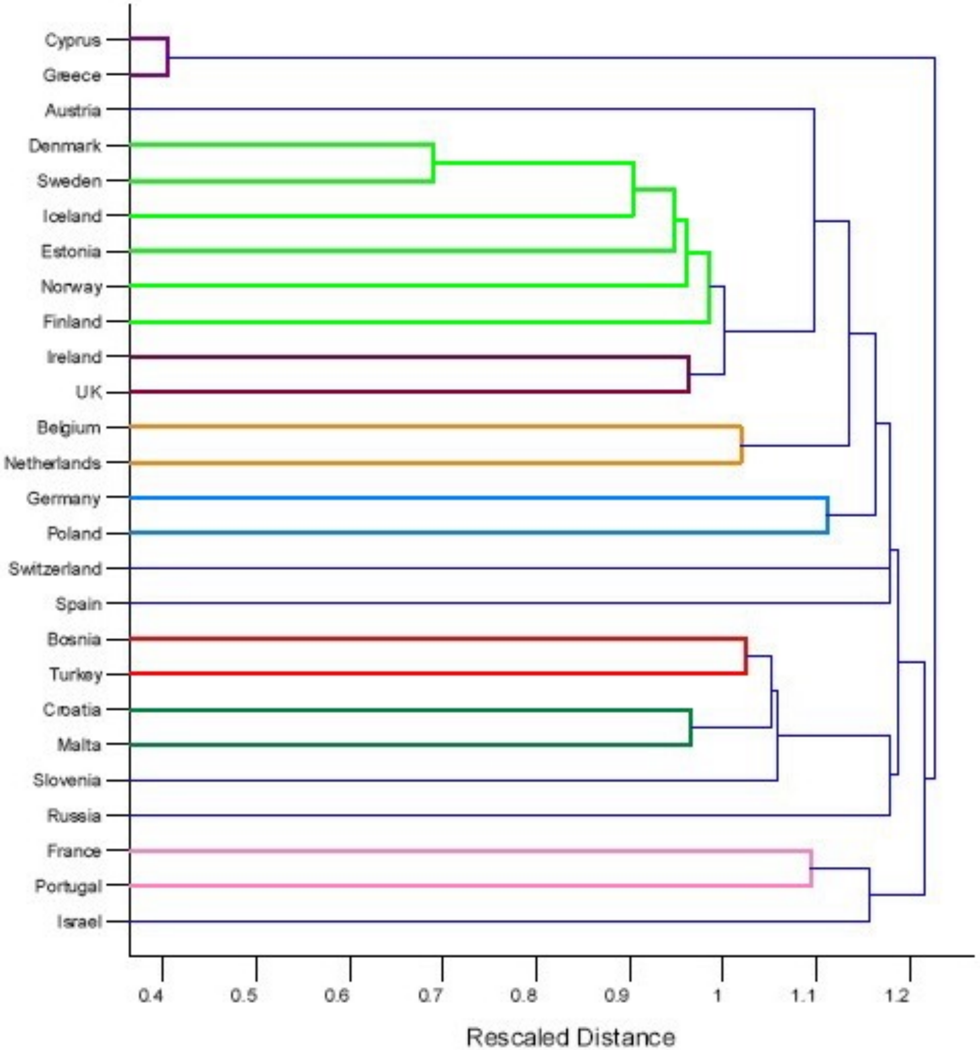
  $d_{ij} = (2(1 - \rho_{ij}))^{1/2}$.
- The most closely related countries have distances close to 0, while the least correlated countries have distances close to 2.

Daniel Fenn et al. (2006). How does Europe Make Its Mind Up? Physica A 360: 576-598.

Ireland in 1998 Eurovision



Eurovision Dendrogram



Dendrogram construction

- Greece and Cyprus have the smallest distance and so they are combined first.
- The next smallest distance is between Denmark and Sweden and so they form the next cluster.
- Once two countries A and B have been combined into a cluster, they are considered to be at the same distance from another country C, which is equal to the shorter of the distances AC and BC.
- This construction is then generalized up for clusters with more than two countries.
- The distance between any two clusters is the shortest distance between any two countries in the two clusters.
- Progressively more countries and clusters are combined in this way, with some countries combining with existing clusters, until all the countries are united into a single cluster.

Linkage

- In order to take the NxN matrix of distances and create a dendrogram or agglomerative hierarchical cluster tree, we need to select a linkage method.

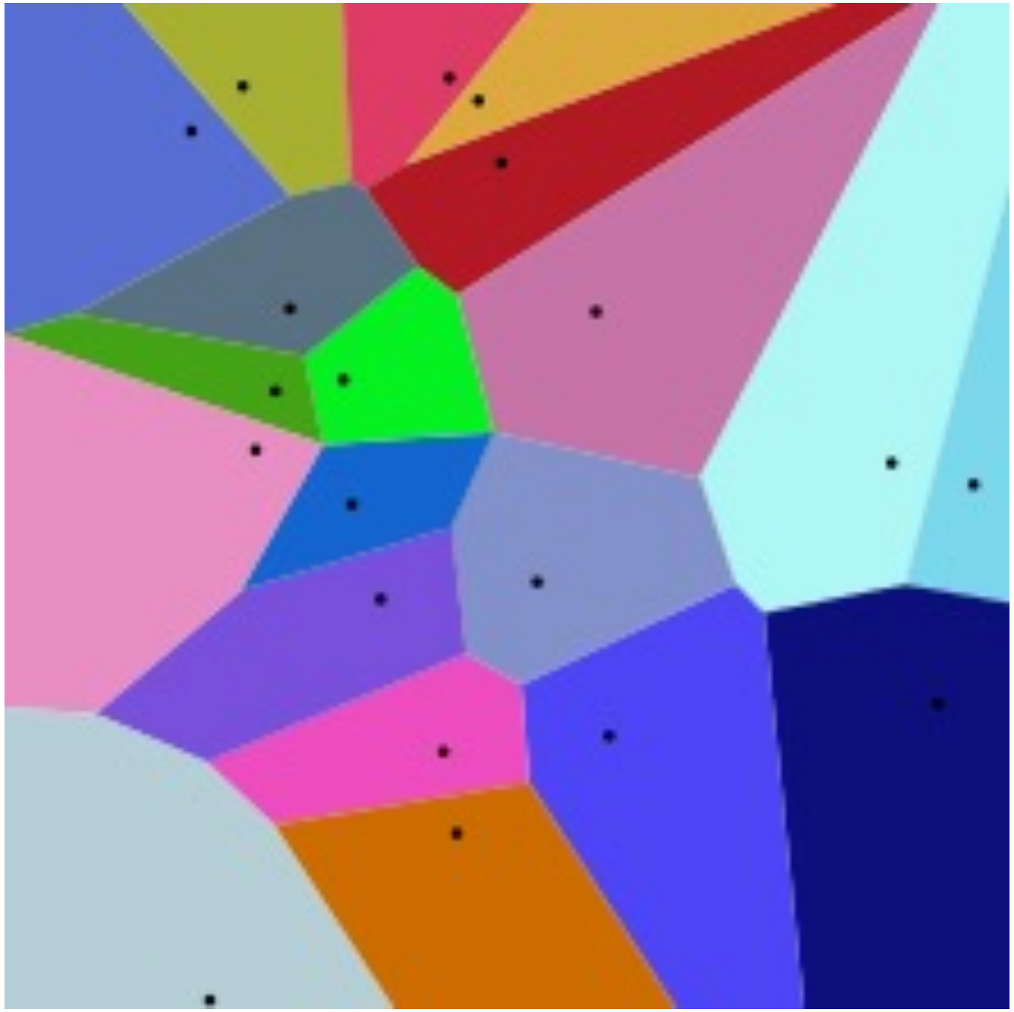
- In the Eurovision example, the shortest distance was used to link the countries.

Method	Description
‘single’	Shortest distance
‘complete’	Furthest distance
‘average’	Average distance

Eurovision Conclusions

- Despite the British tendency to feel distant from Europe, the analysis shows that the U.K. is remarkably compatible, or ‘in tune’, with other European countries.
- Equally surprising is the finding that some other core countries, most notably France, are significantly ‘out of tune’ with the rest of Europe.
- In addition, the analysis provides evidence of a widely-held belief that there are unofficial cliques of countries.
- However these cliques are not always the expected ones, nor can their existence be explained solely on the grounds of geographical proximity.

Voronoi Diagram



Source: wikipedia

k-means

- *k*-means clustering is a partitioning method.
- The idea is to partition the N-by-M data matrix X into k mutually exclusive clusters, and return the index of the cluster to which it has assigned each observation.
- Unlike hierarchical clustering, k-means clustering operates on actual observations (rather than the larger set of dissimilarity measures), and creates a single level of clusters.
- The distinctions mean that k-means clustering is often more suitable than hierarchical clustering for large amounts of data.

k-means

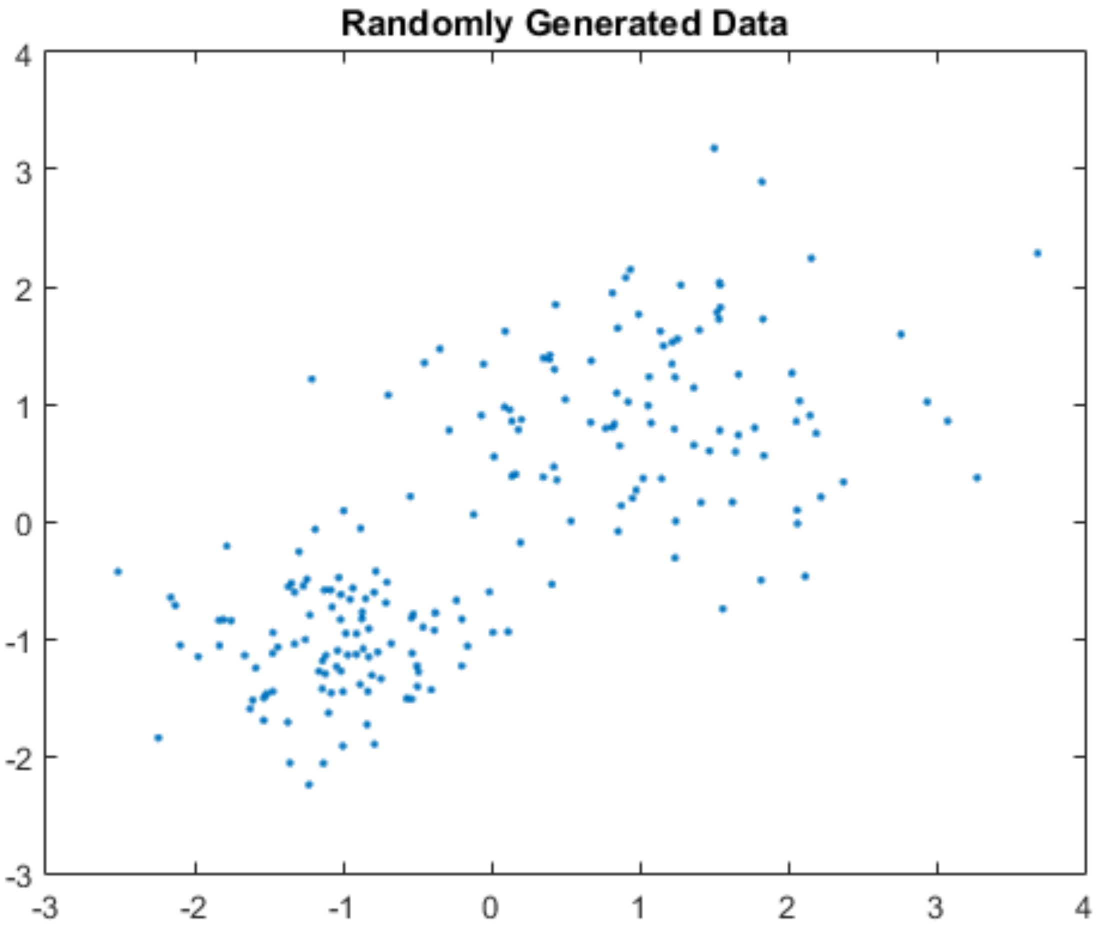
- k-means treats each observation (row in the dataset X) as an object having a location in space.
- It finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible.
- Various different distance measures are available, depending on the kind of data being clustered.
- Each cluster in the partition is defined by its member objects and by its centroid, or center.
- The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized.
- K-means computes cluster centroids differently for each distance measure, to minimize the sum with respect to the measure specified.

k-means

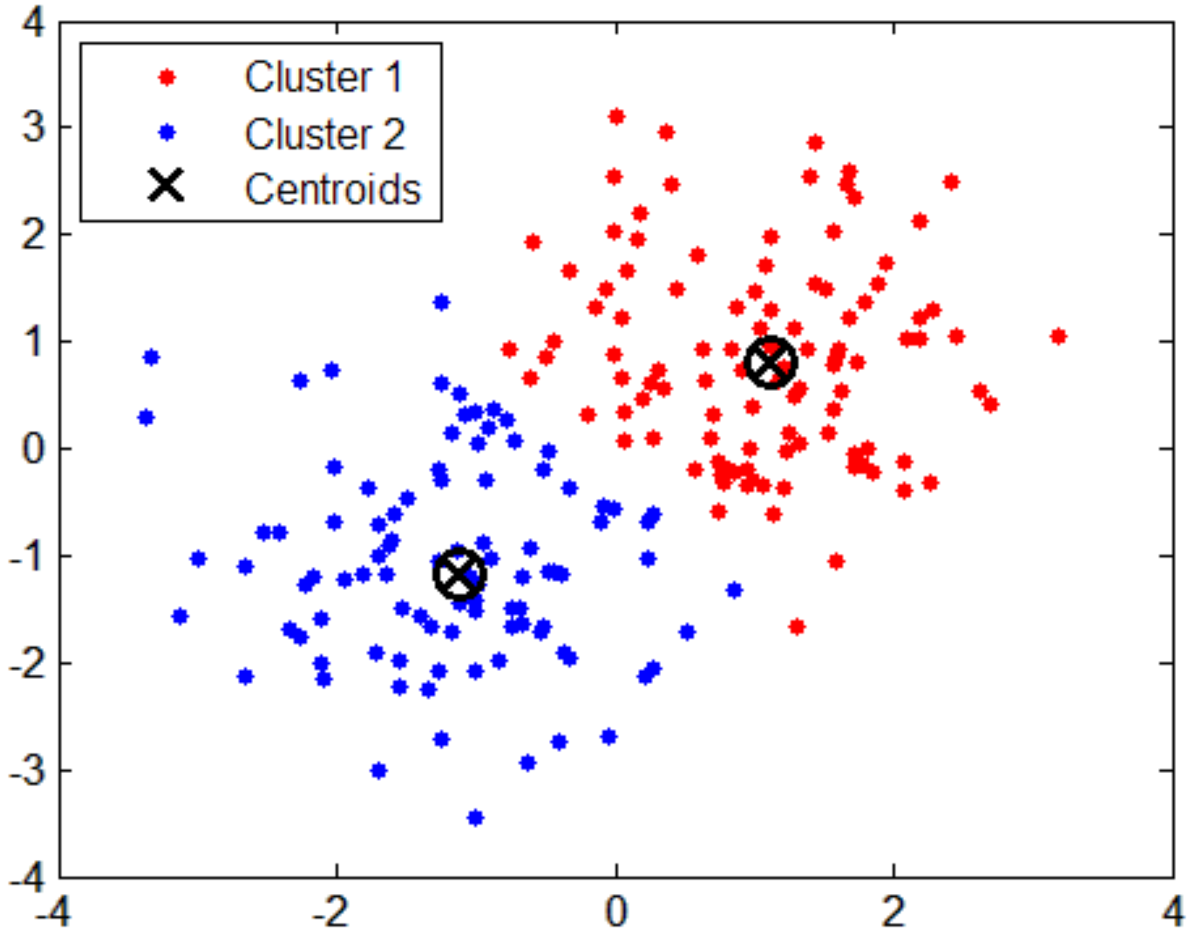
- k-means uses iterative partitioning to minimize the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances.
- Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_N$), where each observation is a M-dimensional real vector, k-means clustering aims to partition the N observations into k ($\leq N$) sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares where $\boldsymbol{\mu}_i$ is the mean of points in S_i:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} \| \boldsymbol{x} - \boldsymbol{\mu}_i \|^2$$

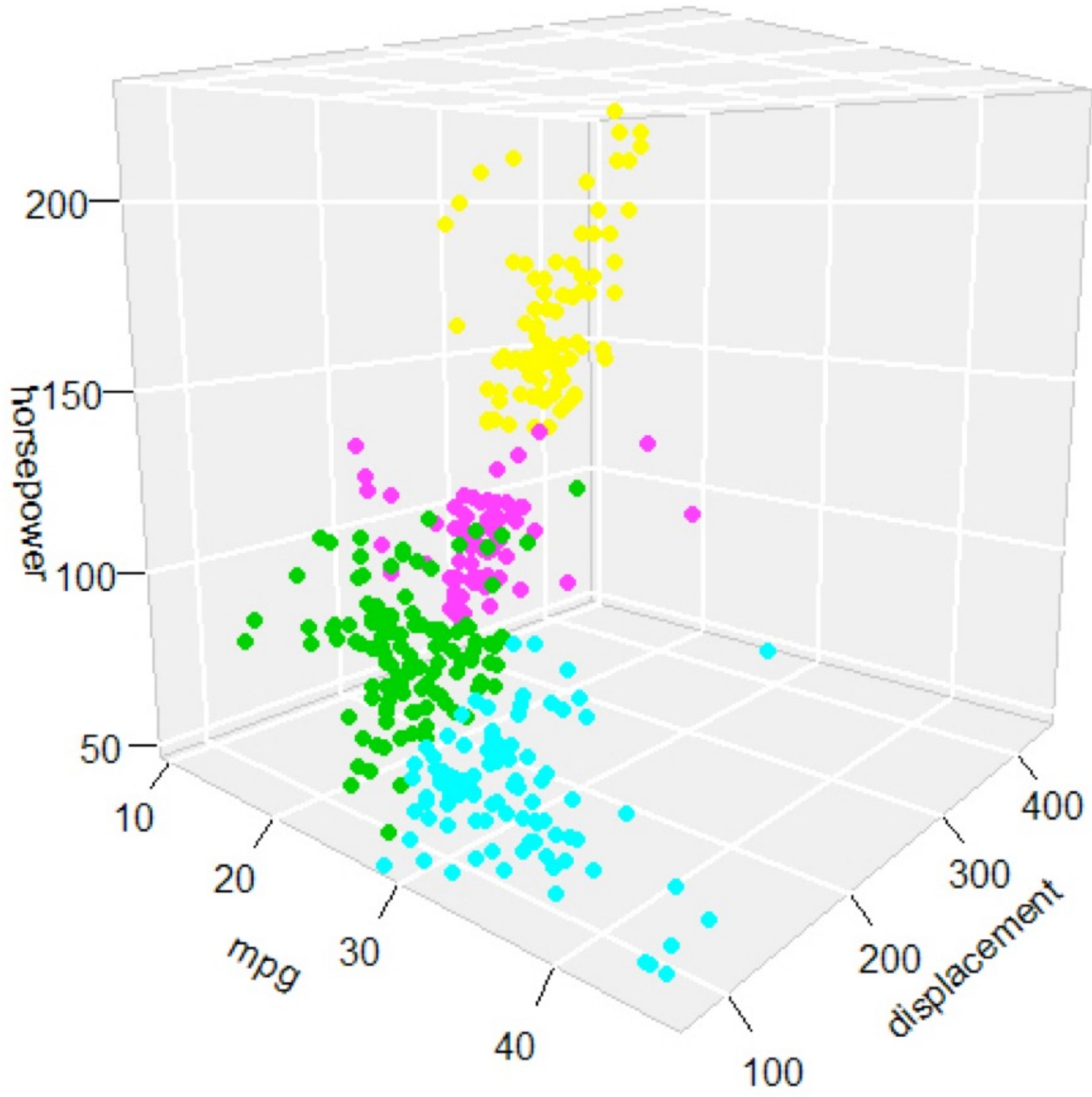
The Data



k-means clustering



Clustering of Horsepower, MPG, and Displacement

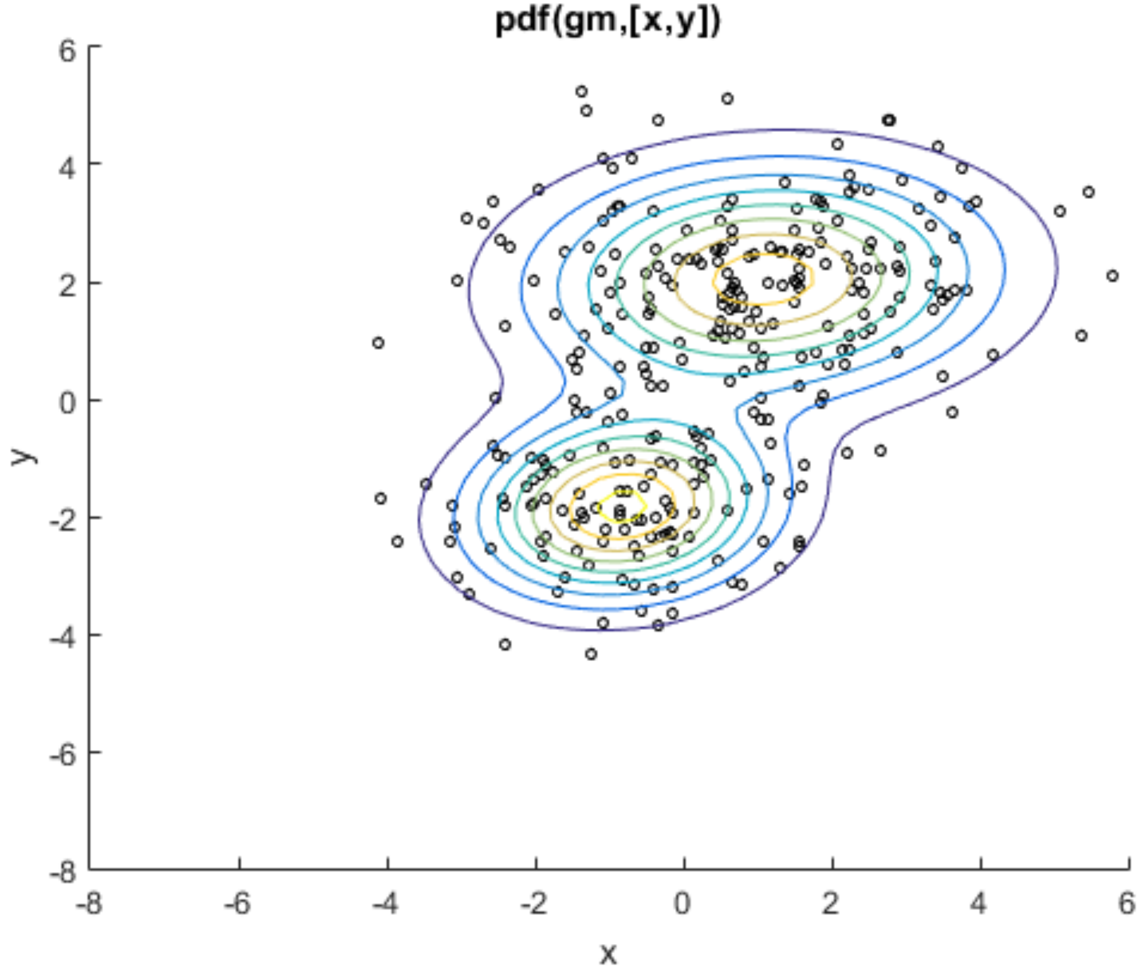


https://blogs.oracle.com/bigdata/k-means-clustering-machine-learning

Gaussian Mixture Models

- Gaussian Mixture Models form clusters by representing the probability density function of observed variables as a mixture of multivariate normal densities.
- Gaussian mixture modeling uses an iterative algorithm that converges to a local optimum.
- Gaussian mixture distributions can be used for clustering data, as the multivariate normal components of the fitted model can represent clusters.
- The posterior probabilities for each point indicate that each data point has some probability of belonging to each cluster.
- Gaussian mixture modeling may be more appropriate than *k*-means clustering when clusters have different sizes and correlation within them.

Gaussian Mixture Model



Self organizing maps

- A self organizing map (SOM) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional discretized representation of the data.
- A SOM is also known as a Kahonen map.
- SOMs are different from other ANNs in the sense that they use a neighborhood function to preserve the topological properties of the input space.
- SOMs with a small number of nodes behave in a way that is similar to *k*-means.

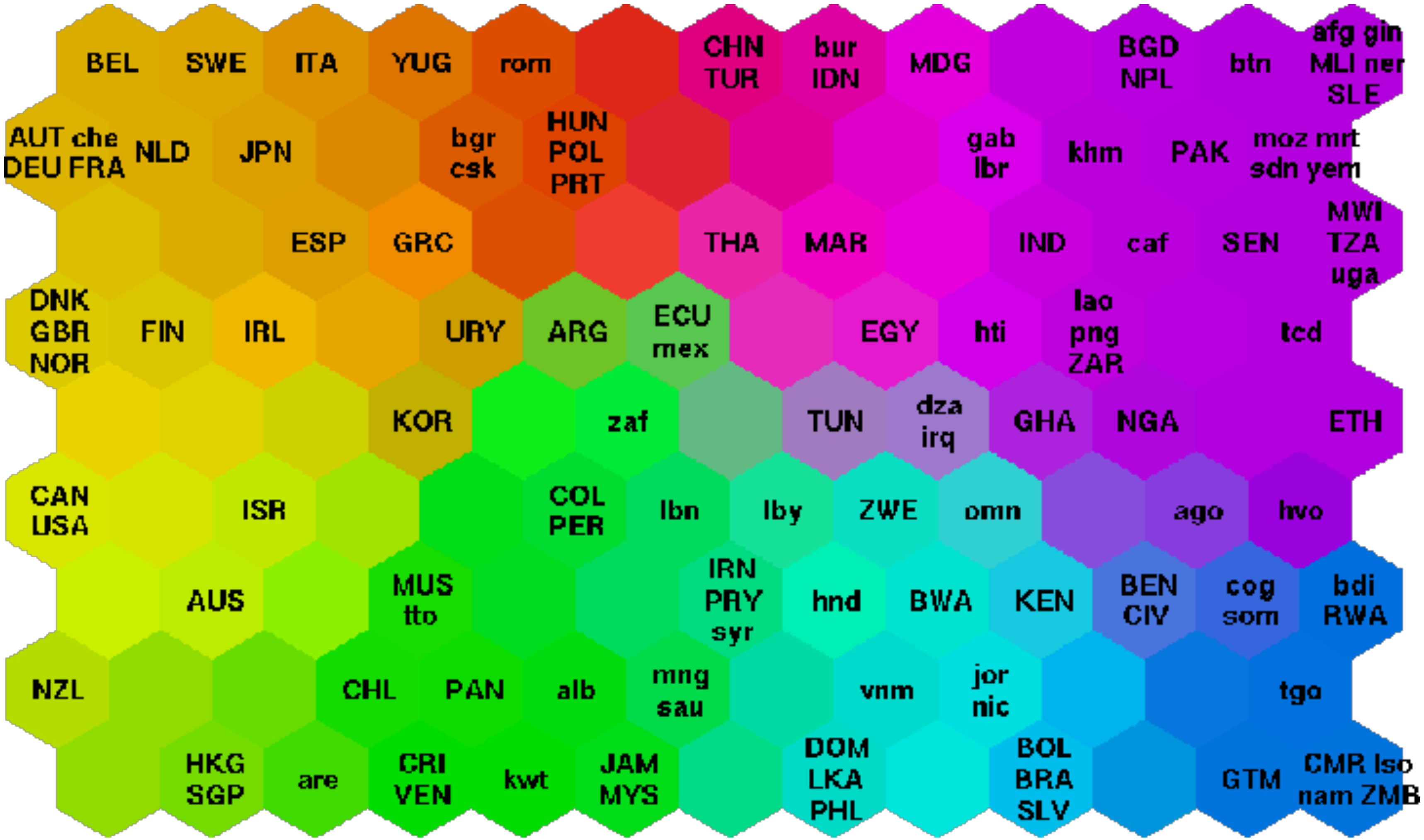
SOM interpretation

- SOM provides a topology preserving mapping from the high dimensional space to map units.
- Map units, or neurons, usually form a two-dimensional lattice and thus the mapping transforms from a high dimensional space onto a plane.
- The property of topology preserving means that the mapping preserves the relative distance between the points.
- Points that are near each other in the input space are mapped to nearby map units in the SOM.
- The SOM can thus serve as a cluster analyzing tool of high-dimensional data.

Visualisation

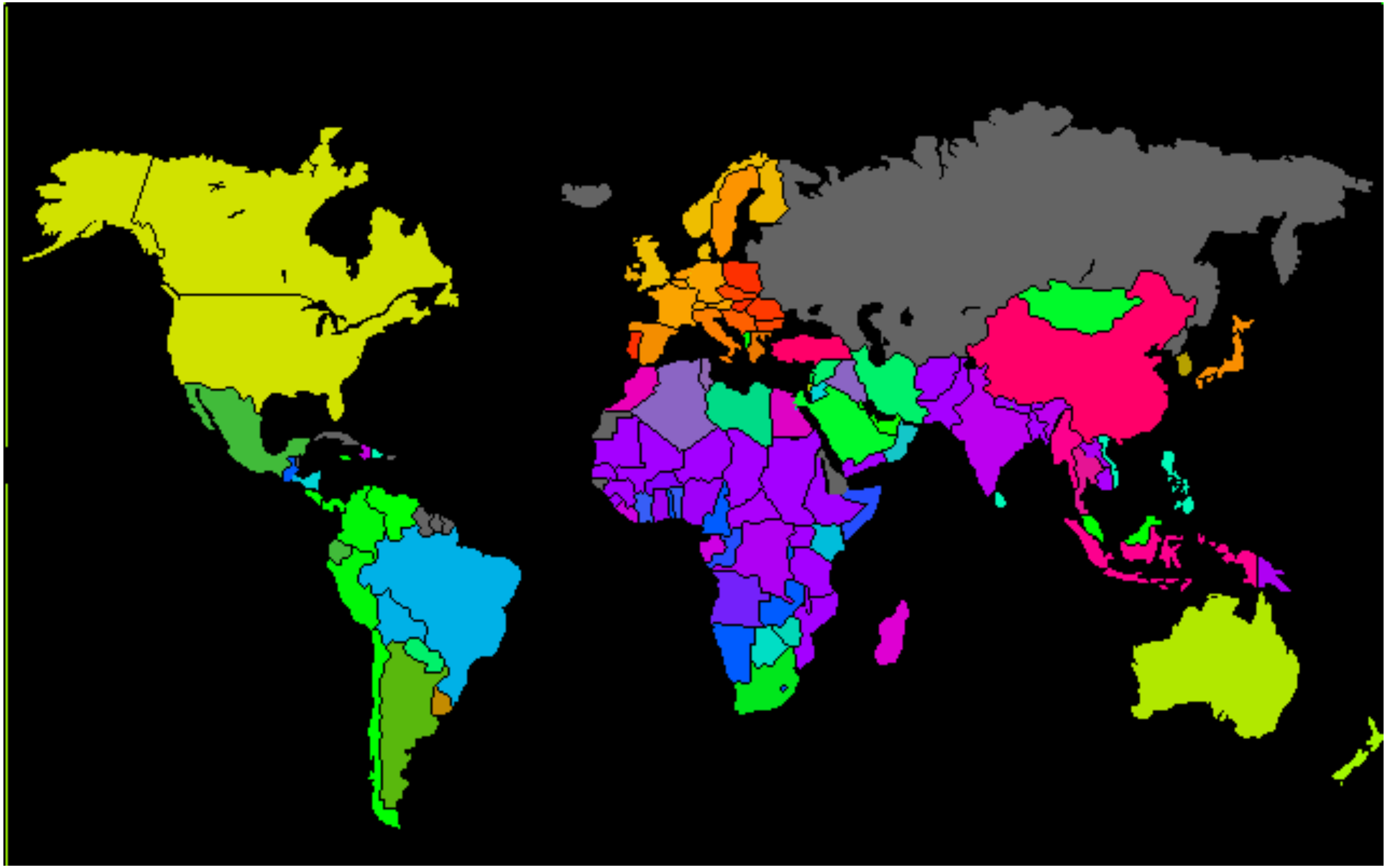
- SOM provides a data visualization technique which helps to understand high dimensional data by reducing the dimensions of data to a 2D map.
- SOM also represents clustering concept by grouping similar data together.
- Therefore it can be said that SOM reduces data dimensions and displays similarities among data.

World Bank data
(39 quality of life indicators)

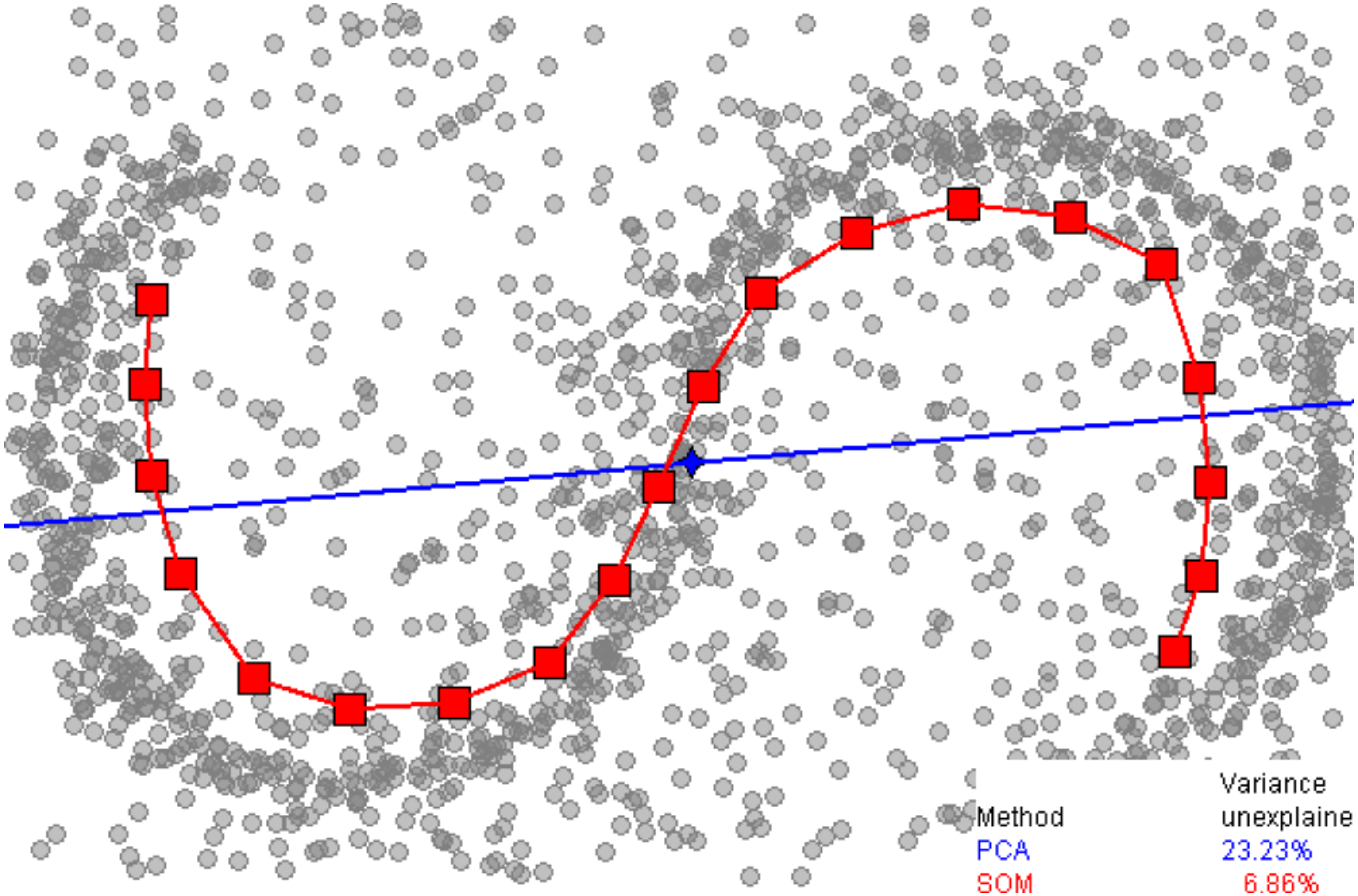


http://www.cis.hut.fi/research/som-research/worldmap.html

World map using SOM



SOM versus PCA

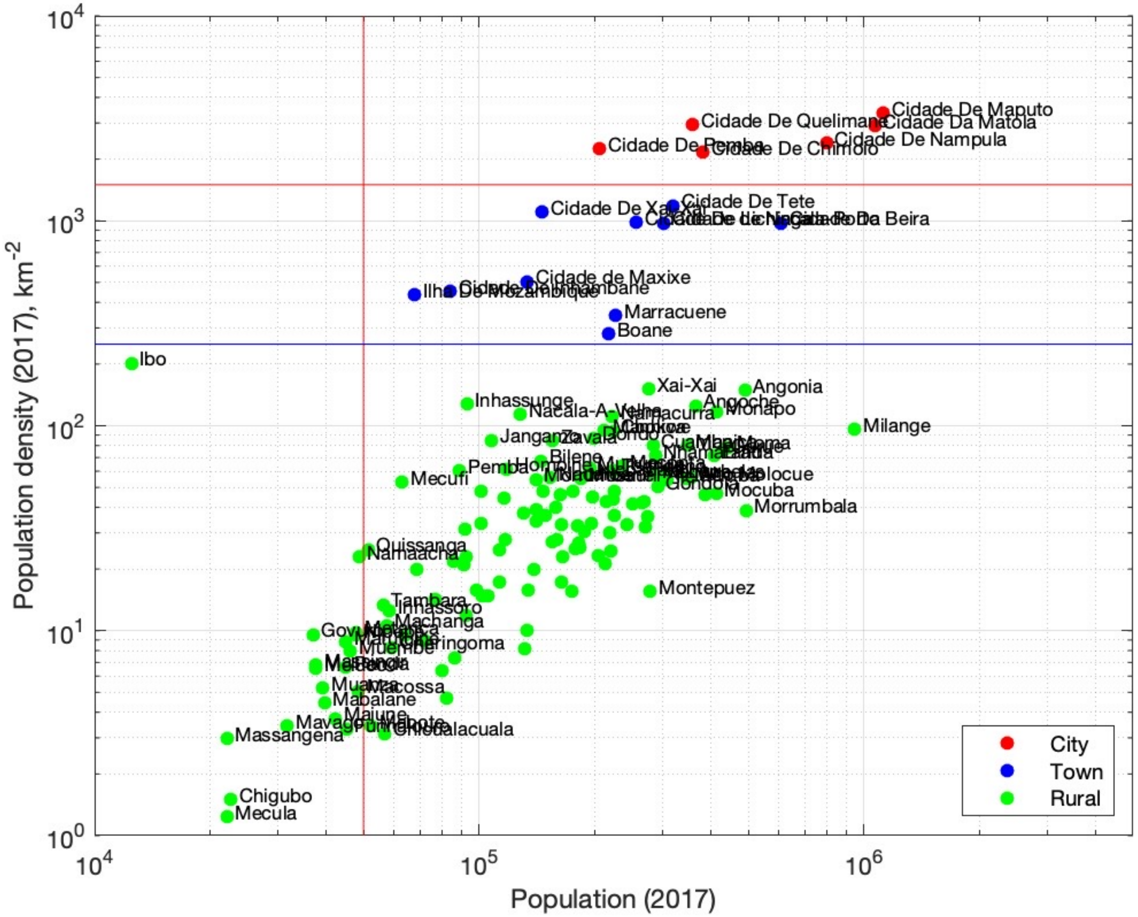


Poll

- What variables would you use to cluster districts by urbanization level (rural, town, city)?

- **Slido.com**
- **#47142**

Mozambique Urbanization



[Dijkstra et al. (2020)](#) discuss how national definitions of urban and rural areas differ significantly from one country to another.

They suggest a simple classification approach using only population and population density

- City: Population >=50000 and Population Density >1500 km^{-2}
- Town: Population >=5000 and Population Density >= 250 km^{-2}
- Rural: otherwise

Source: McSharry et al. (2020). Investing in More Inclusive Growth in Mozambique: spatial approach. World Bank Country Economic Memorandum.

Customer Segmentation

- Customer segmentation allows a company to target specific groups of customers effectively and better allocate marketing resources.
- Traditional segmentation focuses on identifying customer groups based on demographics and attributes such as attitude and psychological profiles.
- Value-based segmentation, on the other hand, looks at groups of customers in terms of the revenue they generate and the costs of establishing and maintaining relationships with them.
- Segmentation can help to predict churn and target sales by treating each segment differently.

Customer Churn

- Churn refers to when a customer ceases his or her relationship with a company.
- Businesses treat a customer as churned when a particular amount of time elapses without generating revenue.
- The cost of customer churn includes both lost revenue and the marketing costs involved in replacing those customers with new ones.
- Churn rate is a measure of the number of customers that have churned over a specific period of time.

Targeted Sales

- A deeper understanding about what your customers are buying offers opportunities for cross-selling.
- For example, a builders merchant sells a large amount of bricks but fails to sell the sand and cement (missed cross-selling opportunity).
- Customer segmentation allows identification of customers who regularly buy certain products.
- The merchant can target them with relevant offers encouraging them to increase spending.

Poll

- What variables would you use to segment customers?

- **Slido.com**
- **#47142**

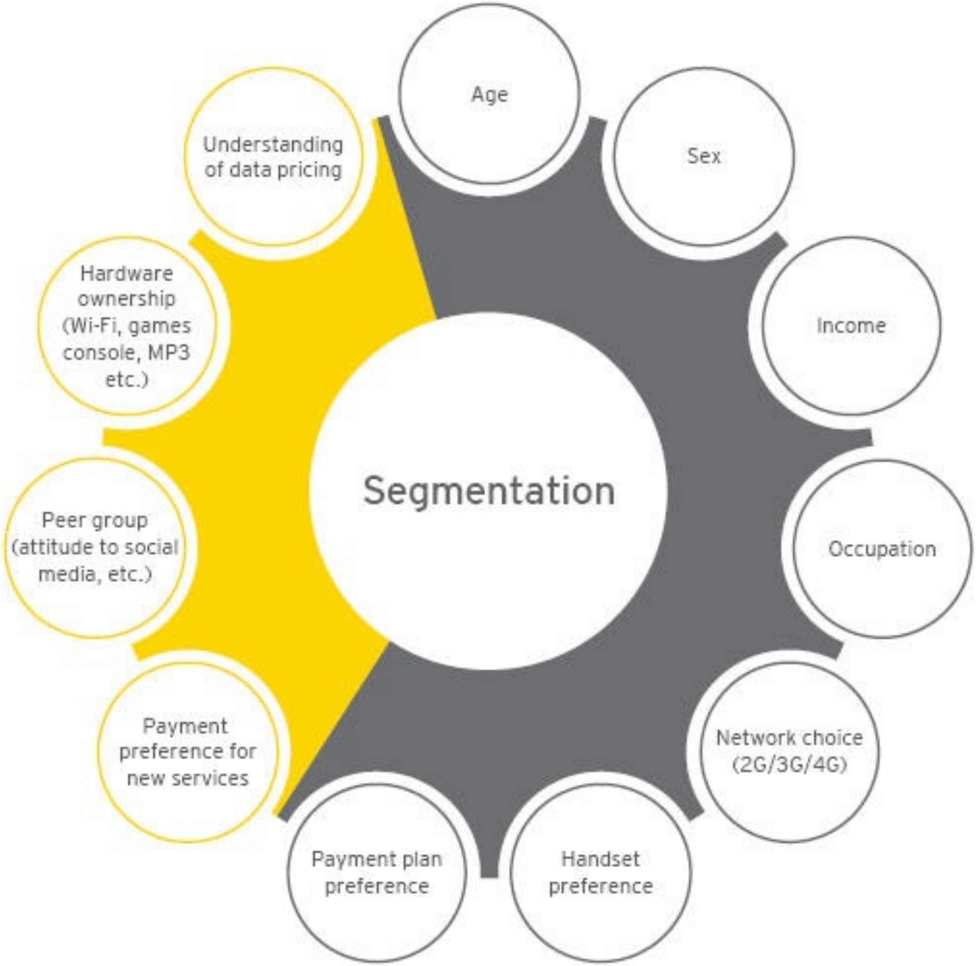
Features for segmentation

- An example of features for customer segmentation may include the following:
- Age
- Gender
- Geographic location
- Socio-economic group
- Income
- Spending

Examples of segments

Feature	Segment A	Segment B
Gender	Male	Female
Age	Retired	Teenage
Location	Rural	Urban
Income	High disposable income	Low disposable income
Free time	High	High

Segmentation for Mobile services



“The mobile maze”. Ernst & Young: www.ey.com

Segmenting customer attributes

- Young urban consumers are the highest users of mobile services.
- Prepaid 3G smartphone customers are a high-value segment.
- 36 to 45-year-olds have high potential.
- Behavioral and attitudinal factors significantly impact service usage.
- Segmentation will need to evolve.

E&Y's Global Consumer Banking Survey

- Goes beyond traditional measurements of age, income and geography.
- It also defines customers based on the products they own, the channels they prefer and their reasons for trust.
- Survey included 32,000 customers across 43 countries to evaluate 31 banking experience elements.
- This research provides a unique view of today's customer, regardless of location, and reveals opportunities for banks to more effectively invest resources and craft strategies to strengthen their customer relationships.
- Grouped survey respondents into eight global segments. Each segment is defined by its members' common behaviors and characteristics, including similar preferences in what they want from their primary financial service provider.
- The segments vary by size, assets and willingness to pay more for key benefits.

Eight segments in Banking

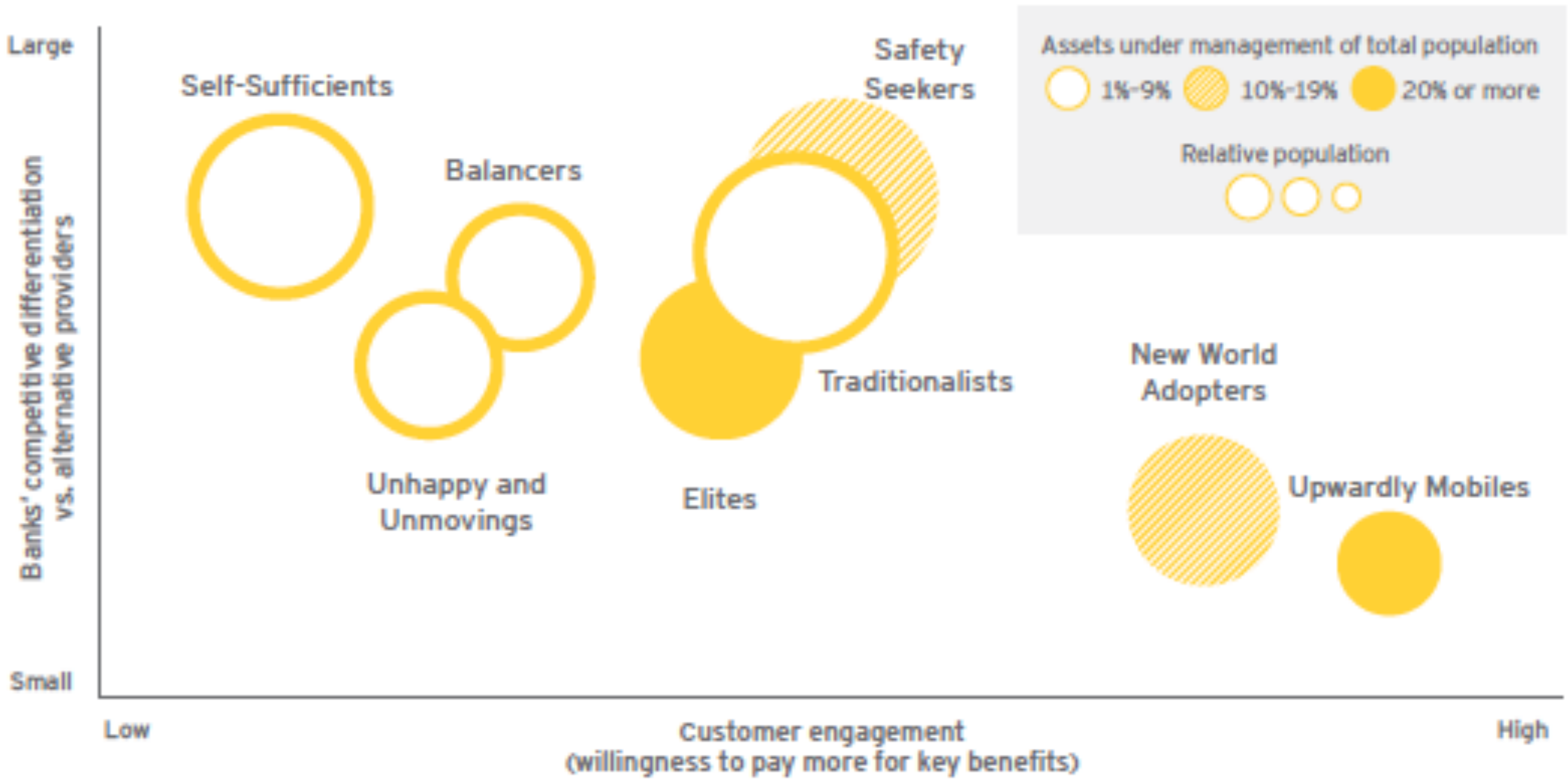
1. Upwardly Mobiles	5. Safety Seekers
2. Elites	6. Traditionalists
3. New World Adopters	7. Self-sufficients
4. Balancers	8. Unhappy and Unmovings

“Global Consumer Banking Survey”. Ernst & Young: www.ey.com

Banking Segment Descriptions

Segment	Age	Education	Household Income	Assets	Description
Upwardly Mobiles	18-34 (43%)	College Graduate (80%)	$48,571	$250,000	Own many products, switchers, value financial advice
Elites	50+ (43%)	College Graduate (70%)	$46,667	$122,393	Financial goals, research online
New World Adopters	18-34 (44%)	College Graduate (75%)	$29,584	$90,750	Like technology, actively switch
Balancers	50+ (38%)	College Graduate (59%)	$41,429	$46,875	Fee transparency, problem resolution
Safety seekers	18-34 (41%)	College Graduate (53%)	$18,667	$31,875	Likely to trust providers
Traditionalists	50+ (39%)	College Graduate (53%)	$16,358	$31,875	Use branches & ATMs
Self-sufficients	18-34 (42%)	College Graduate (51%)	$29,922	$28,684	Low levels of trust, do own research
Unhappy & unmovings	50+ (40%)	College Graduate (47%)	$25,000	$30,984	All providers are the same

Differentiation versus Engagement



"Global Consumer Banking Survey". Ernst & Young: www.ey.com

Matlab

- zscore
- pdist
- linkage
- dendrogram
- cluster
- kmeans
- gmdistribution.fit

Q&A