Machine Learning for Social Sciences
Part 2: Factorisation

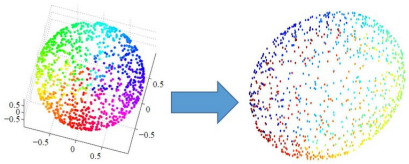
Felix Hagemeister

TUM School of Social Sciences and Technology
April 2022

Factorisation

Dimensionality Reduction: find a lower-dimensional summary of a high-dimensional **x**.

- **Supervised:** Response variable y dictates direction of dimensionality reduction (e.g. regression finds $\hat{y}$).
- **Unsupervised:** There is no response or outcome, reduce **x** for its own sake.



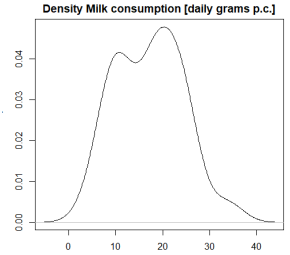
Factorisation

Clustering: Represent data as outcome of a *mixture distribution*.

- Assume that $\mathbf{x}$ is drawn from K different probability distributions $p_k(\mathbf{x})$ for $k = 1...K$
- Properties of these distributions define the clusters
- Mixtures of distributions can yield all sorts of complicated distributions

$$p(\mathbf{x}) = \pi_i p_i(\mathbf{x}) + ...\pi_K p_K(\mathbf{x})$$

where π_K is the probability for component k in the population



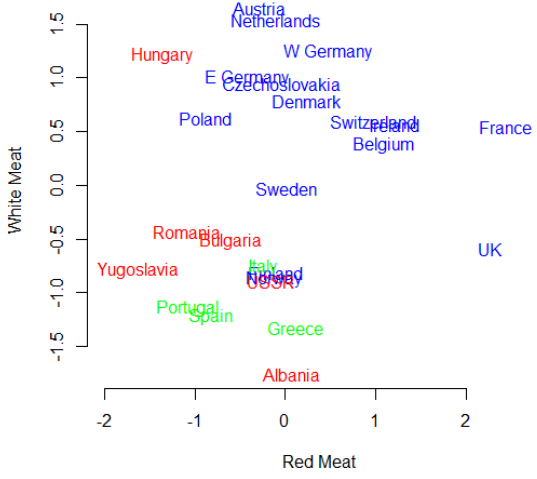
Factorisation

K-Means: Estimate cluster membership across K components.

- Membership of observation $\mathbf{x}_i$ is in one single cluster.
- K-Means is indeterminate (running again might produce slightly different results).
- Could run K-Means mulitple times using different random starts and use solution with smallest deviance.
- Choose K subjectively, might choose K with smallest BIC.

Factorisation

K-means with K=3 for protein consumption data in EU countries



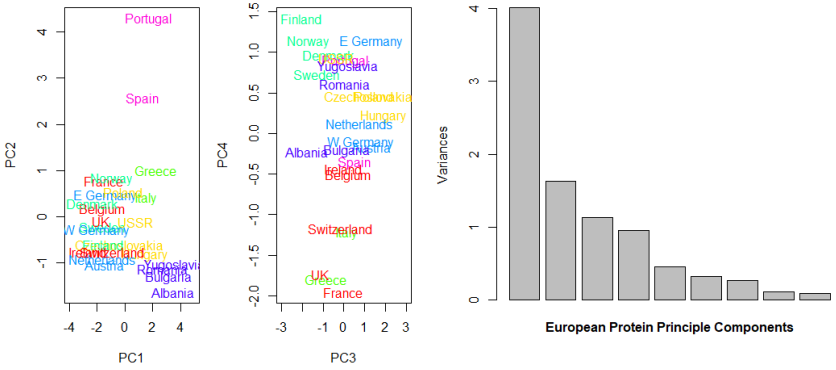
Factorisation

Factor Models and PCA: Allow for *mixed membership* in clusters.

$E[x_{ij}] = \phi_{i1}\nu_{i1} + ... + \phi_{iK}\nu_{iK}, j = 1...p$

- where ϕ_{ik} coefficients are called *loadings* or *rotations*,
- ν_{ik} are K factors (the lower-dimensional summary of $\mathbf{x}_i$).
- Interpretation of K can be bottom-up (using big individual ϕ_{ki} rotations) or top-down (using fitted ν_{ik} *anddomainknowledge*)
- Use proportion of variance explained by each PC (*screeplot*) to choose K

Factorisation



Factorisation

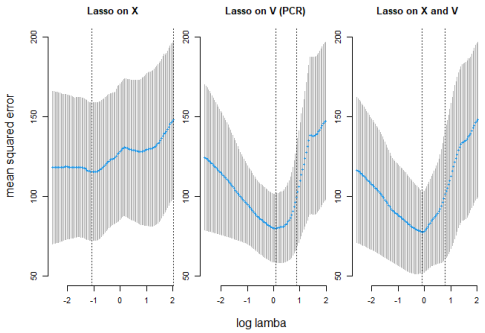
Let's implement k-means and principal component analysis ourselves using data on protein consumption from EU countries.

Here is the coding example.

Factorisation

Principal Component Regression (PCR): Use factors from PCA as inputs in regression

- Could use PCA on unlabeled data (unsupervised) and then use results in (supervised) regression
- One tactic is to use both ν and $\mathbf{x}$ – PCs and raw inputs – in a lasso regression



Factorisation

Let's implement principal component regression (PCR) on television survey data.

Here is the coding example.