Università degli studi di
Roma “LA SAPIENZA”

Advances in Data Analysis and Statistical Modeling

Dr. Mario Fordellone

mario.fordellone@uniroma1.it

What statistical models will we use?

- Disjoint Principal Component Analysis (DPCA)
- K-Means model
- Reduced K-Means model (REDKM)
- Well Structured Partition (WSP)
- Well Structured Perfect Partition (WSPP)

Disjoint Principal Component Analysis (DPCA)

DPCA is a sort of PCA with different constraints on the loading matrix **A**. In particular, the loading matrix obtained by DPCA has the column squared sum equal to 1, each element bigger than 0, and each variable can be related to only one latent factor.
DPCA algorithm optimizes the objective function:

$$\min_{\mathbf{A}} ||\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}'||^2$$

$$subject\ to \quad \sum_{j=1}^{J} a_{j,q}^2 = 1$$

$$\sum_{j=1}^{J} \left(a_{j,q} a_{j,r}\right)^2 = 1$$

$$\sum_{j=1}^{J} a_{j,q} > 0$$

K-Means

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (K) fixed a priori. The main idea is to define K centroids given by the arithmetic mean of each group.

K-Means algorithm optimizes the objective function:

$$\min_{\mathbf{U},\bar{\mathbf{X}}} ||\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}||^2$$

$subject\ to \quad u_{i,k} \in \{0,1\}$

$$\mathbf{U1_K} = \mathbf{1_N}$$

K-Means

1. Place *K* points into the space represented by the objects that are being clustered. These points represent initial centroids;
2. Assign each object to the group that has the closest centroid;
3. When all objects have been assigned, recalculate the positions of the *K* centroids;
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated

Reduced K-Means (REDKM)

PCA:
$\mathbf{X} = \mathbf{XAA}' + \mathbf{E}_1$
subject to:
$\mathbf{A}'\mathbf{A} = \mathbf{I}_P$ (orthogonal)

K-means:
$\mathbf{XA} = \mathbf{U}\bar{\mathbf{X}}\mathbf{A} + \mathbf{E}_2$
subject to:
$\mathbf{U}[u_{ik} \in \{0, 1\}]$ (binary)
$\mathbf{U1}_K = \mathbf{1}_N$ (row stochastic)

$\mathbf{E}_1$ and $\mathbf{E}_2$ are the errors matrices

Reduced K-means (RKM)

$$\mathbf{X} = (\mathbf{U}\bar{\mathbf{X}}\mathbf{A} + \mathbf{E}_2)\mathbf{A}' + \mathbf{E}_1$$

if $\mathbf{E}_{RKM} = \mathbf{E}_2\mathbf{A}' + \mathbf{E}_1$

$$\mathbf{X} = \mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}' + \mathbf{E}_{RKM}$$

by post-multiplying both members of the previous equation by $\mathbf{A}$ and rewriting the error $\mathbf{E}_{FKM} = \mathbf{E}_{RKM}$, we have the Factorial *K*-means (FKM):

$$\mathbf{XA} = \mathbf{U}\bar{\mathbf{X}}\mathbf{A} + \mathbf{E}_{FKM}$$

subject to:
$\mathbf{U}[u_{ik} \in \{0, 1\}]$ (binary)
$\mathbf{U1}_K = \mathbf{1}_N$ (row stochastic)
$\mathbf{A}'\mathbf{A} = \mathbf{I}_P$ (orthogonal)

Reduced K-Means (REDKM)

Objective function PCA K-means

$$\min_{\mathbf{U},\bar{\mathbf{X}},\mathbf{A}} \left\|\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'\right\|^2 \text{ or } \max_{\mathbf{U},\bar{\mathbf{X}},\mathbf{A}} \left\|\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'\right\|^2$$

subject to

$\mathbf{U}[u_{ik} \in \{0,1\}];\ i = 1,\ldots,N;\ k = 1,\ldots,K$

$\sum_{k=1}^{K} u_{ik} = 1$

$\mathbf{A}'\mathbf{A} = \mathbf{I}_P$

Alternative least-squares PCA K-means algorithm

- **Step 0**: Initial random values are chosen for $\hat{\mathbf{A}}$ and $\hat{\mathbf{U}}$.
- **Step 1**: Update $\bar{\mathbf{X}}$ by $\bar{\mathbf{X}} = (\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}'\mathbf{X}$, given $\hat{\mathbf{U}}$.
- **Step 2**: Update $\mathbf{U}$ by $F([u_{ik}]) = \left\|\mathbf{X} - \mathbf{U}\hat{\bar{\mathbf{X}}}\hat{\mathbf{A}}\hat{\mathbf{A}}'\right\|^2$, given the current values of $\hat{\mathbf{A}}$ and $\hat{\bar{\mathbf{X}}}$. The problem is solved by taking $u_{ik} = 1$, if $F([u_{ik}]) = min\{F([u_{iv}]) : v = 1,\ldots,P; (v \neq k)\}$; $u_{ik} = 0$, otherwise.
- **Step 3**: Update $\mathbf{A}$, given $\hat{\mathbf{U}}$ and $\hat{\bar{\mathbf{X}}}$, by minimizing $\left\|\mathbf{X} - \hat{\mathbf{U}}(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}'\mathbf{X}\mathbf{A}\mathbf{A}'\right\|^2$. The problem is solved by taking the first p eigenvectors of $\mathbf{X}'(\hat{\mathbf{U}}(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}')\mathbf{X}$ (e.g., see Ten Berge, 1993).
- **Stopping Rule**: compute the objective function for the current values of $\hat{\mathbf{U}}$, $\hat{\bar{\mathbf{X}}}$ and $\hat{\mathbf{A}}$. If the function decreases more than an arbitrary small constant, it return to **Step 1**; otherwise, stop the process.

Well Structured Partition (WSP)

WSP algorithm optimizes the objective function:

$$\min_{\mathbf{U},\mathbf{D_B},\mathbf{D_W}} ||\mathbf{D} - \mathbf{U}\mathbf{D_B}\mathbf{U}' - \mathbf{U}\mathbf{D_W}\mathbf{U}' + diag(\mathbf{U}\mathbf{D_W}\mathbf{U}')||^2$$

$$subject\ to \quad u_{i,k} \in \{0,1\}$$

$$\mathbf{U}\mathbf{1_K} = \mathbf{1_N}$$

Well Structured Perfect Partition (WSPP)

WSPP algorithm optimizes the objective function:

$$\min_{\mathbf{U},\alpha_1,\alpha_2} ||\mathbf{D} - \alpha_2(\mathbf{1}_n\mathbf{1}_n' - \mathbf{U}\mathbf{U}') - \alpha_1(\mathbf{U}\mathbf{U}' - \mathbf{I}_n)||^2$$

$$subject\ to \quad u_{i,k} \in \{0,1\}$$

$$\mathbf{U1}_\mathrm{K} = \mathbf{1}_\mathrm{N}$$

$$0 < \alpha_1 \leq \alpha_2$$

Statistical applications on European Social Survey Data set

Items	Topics
Core A1-A6	Media use; internet use; social trust.
Core B1-B43	Politics, including: political interest, trust, electoral and other forms of participation, party allegiance, socio-political orientations, immigration.
Core C1-C44	Subjective wellbeing, social exclusion, crime, religion, perceived discrimination, national and ethnic identity, test questions (sect. I), refugees.
Rotating D1-D32	Climate change and energy, including: attitudes, perceptions module and policy preferences.
Rotating E1-E42	Welfare, including attitudes towards welfare provision, size of module claimant groups, attitudes towards service delivery and likely future dependence on welfare, vote intention in EU referendum.
Core F1-F61	Socio-demographic profile, including household composition, sex, age, marital status, type of area, education and occupation, partner, parents, union membership, income and ancestry.
Core Section H	Human values scale.
Core Section I	Test questions.

Statistical applications on European Social Survey Data set

	DIMENSION	VARIABLES
1	POLITICS	from 1 to 19
2	ECONOMICS	from 20 to 21
3	SOCIAL	from 22 to 23
4	CULTURAL	from 24 to 25
5	CRIME	from 26 to 27
6	RELIGION	from 28 to 29
7	STRUCTURAL	from 30 to 40
8	HOUSE HOLD	from 41 to 49
9	EMPLOYMENT	from 50 to 64

Statistical applications on European Customer Satisfaction Index (ECSI)

ECSI is an economic indicator that measures the customer satisfaction. It is an adaptation of the Swedish Customer Satisfaction Barometer (Fornell, 1992; Fornell et al., 1996; Hackl and Westlund, 2000) and is compatible with the American Customer Satisfaction Index (ACSI). A model has been derived specifically for the ECSI. In the complete model, there are seven interrelated latent variables

Statistical applications on European Customer Satisfaction Index (ECSI)

ima1	IMAGE	qua1	PERCEIVED QUALITY	val1	PERCEIVED VALUE
ima2		qua2		val2	
ima3		qua3		sat1	SATISFACTION
ima4		qua4		sat2	
ima5		qua5		sat3	
exp1	EXPECTATIONS	qua6		comp	COMPLAINTS
exp2		qua7		loy1	LOYALTY
exp3				loy2	
				loy3	