

CONSENSUS CLUSTERING VIA PIVOTAL METHODS



Leonardo Egidi · Roberta Pappadà

Francesco Pauli · Nicola Torelli

**Department of Economics, Business,
Mathematics and Statistics 'B. de Finetti'
University of Trieste**



**StaTalk2019
Trieste, 11.22.19**

Summary

The Cluster Ensemble Framework

Pivotal methods in K -means clustering

Concluding remarks

Appendix

The Cluster Ensemble Framework

K-Means Clustering Problem

K-means clustering is a simple and widely used approach for partitioning a data set into k distinct, non-overlapping clusters.

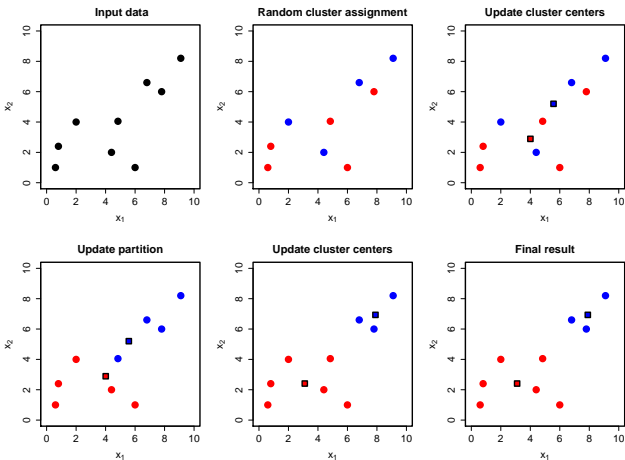
Given a dataset $\mathbf{Y} = (y_1, \dots, y_n)$, with $y_i \in \mathcal{Y} \subset \mathbb{R}^d$, K-means seeks to find a partition of the data into K clusters by minimizing the sum of squared distances between every data point and its nearest cluster center, that is

$$\underset{C_1, \dots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^K \sum_{y_i \in C_k} \|y_i - \mu_k\|^2 \right\},$$

where μ_k is the k -th cluster center.

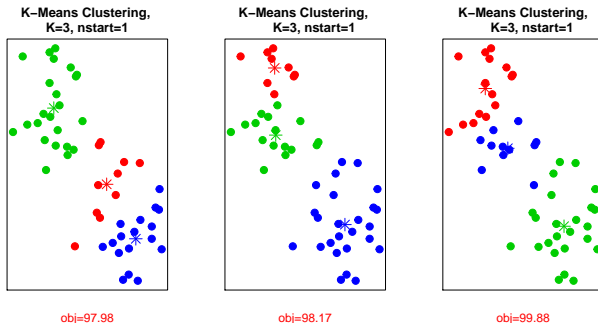
K-Means Algorithm

The k -means algorithm finds a local solution by iteratively assigning each data point to the cluster with the closest center, using random cluster initialization.



K-Means Clustering: different initializations

Different initial centers may lead to different solutions:



One might have to run the algorithm multiple times with random starting centers (*seeds*) for each run: the final solution is the one for which the objective is minimized

The K-Means algorithm: weaknesses

Despite its popularity, well known limitations of k -means algorithm are:

- convergence to a local minimum may produce ‘wrong’ results;
- it is sensitive to the choices of initial cluster centres
- the k -means algorithm works poorly in the case of unbalanced cluster sizes and non-spherical shapes.

Extensions of k -means and initialization strategies have been developed [Jain, 2010; Arthur & Vassilvitskii, 2007].

Cluster ensemble methods have emerged from the need to combine data partitions and generate a better clustering result [Strehl & Ghosh (2002)]

Clustering ensembles can be generated by

- applying different clustering algorithms
- exploring different proximity measures
- using the same clustering algorithm with different parameters or initializations

Evidence accumulation clustering is based on the assumption that each clustering result as an independent evidence of data structure [Fred & Jain (2005)]

Pairwise-similarity based approach

Given a set of n observations (y_1, \dots, y_n) and a set of H partitions $\mathcal{P} = \{P^1, P^2, \dots, P^H\}$ into K disjoint clusters, a co-association matrix can be used to represent the ensemble information:

$$c_{i,j} = \frac{n_{i,j}}{H}$$

where $n_{i,j}$ is the number of times the pair (y_i, y_j) is assigned to the same cluster among the H partitions of the ensemble. The $n \times n$ matrix C

- is used to obtain the consensus partition
- fulfils the conditions of a similarity matrix
- is expected to contain a non-negligible number of zeros

Pivotal methods in K -means clustering

The proposed approach can be summarized as follows:

1. cluster ensembles are created using repeated runs of a single clustering algorithm (e.g. the K -means technique with a random initialization of cluster centres)
2. specific units called **pivots** can be identified via the co-association matrix, such that they are representative of the group they belong to (because they never or very rarely co-occur with members of other groups)
3. the pivots are used to define a new initialization step in the K -means algorithm, in order to reduce the effect of random seeding

The starting point for the pivotal methods we propose is a partition G_1, \dots, G_K of y_1, \dots, y_n into K groups (*reference partition*) obtained, for instance, by applying agglomerative hierarchical clustering:

- We assume that K distinct pivots do exist and propose some criteria to identify them
- each pivot can be chosen so that it is as far as possible from units that might belong to other groups and/or as close as possible to units that belong to the same group

Again $C = (c_{ij})$ denotes the co-association matrix produced by multiple clusterings of the same data set.

Pivot identification criteria

The pivot y_{i_k} for group G_k , $k = 1, \dots, K$, is chosen so that

- it **maximizes** the 'global' within similarity

$$(a) \quad \max_{i_k} \sum_{j \in G_k} c_{i_k, j}$$

- it **minimizes** the 'global' similarity between one group and all the others

$$(b) \quad \min_{i_k} \sum_{j \notin G_k} c_{i_k, j}$$

- or **maximizes** the discrepancy between global within and between similarities

$$(c) \quad \max_{i_k} \left(\sum_{j \in G_k} c_{i_k, j} - \sum_{j \notin G_k} c_{i_k, j} \right)$$

When the number of groups is small, an alternative strategy to detect pivotal units is the algorithm of Maxima Units Search (MUS):

- firstly introduced in the framework of *label switching problem* in Bayesian estimation of finite mixture models
- Given a large and sparse 0-1 matrix, the MUS algorithm seeks those elements, among a specified number of candidate pivots, whose corresponding rows contain more zeros compared to all other units
- ideally, the submatrix of C with only the rows (columns) of the selected pivotal units is identical

[Egidi, Pappadà, Pauli, Torelli (2018a,b)]

1. Initialization

[a] Generate the clustering ensemble \mathcal{P} of H partitions, where each clustering is the result of a K -means run with randomly selected cluster centers μ_1, \dots, μ_K

1. Initialization

[a] Generate the **clustering ensemble** \mathcal{P} of H partitions, where each clustering is the result of a K -means run with randomly selected cluster centers μ_1, \dots, μ_K

[b] Build the **co-association matrix** $C = (c_{i,j})$, $i, j = 1, \dots, n$, from \mathcal{P}

piv_KMeans: Algorithm

1. Initialization

[a] Generate the **clustering ensemble** \mathcal{P} of H partitions, where each clustering is the result of a K -means run with randomly selected cluster centers μ_1, \dots, μ_K

[b] Build the **co-association matrix** $C = (c_{i,j})$, $i, j = 1, \dots, n$, from \mathcal{P}

[c] Given a partition of the data set in K groups, G_1, \dots, G_K , apply a **pivotal method** to C , choosing among MUS, (a)–(c), to find the pivots $y_{i_1}, \dots, y_{i_K} \rightarrow$ set $\mu_k = y_{i_k}$

piv_KMeans: Algorithm

1. Initialization

[a] Generate the **clustering ensemble** \mathcal{P} of H partitions, where each clustering is the result of a K -means run with randomly selected cluster centers μ_1, \dots, μ_K

[b] Build the **co-association matrix** $C = (c_{i,j})$, $i, j = 1, \dots, n$, from \mathcal{P}

[c] Given a partition of the data set in K groups, G_1, \dots, G_K , apply a **pivotal method** to C , choosing among MUS, (a)–(c), to find the pivots $y_{i_1}, \dots, y_{i_K} \rightarrow$ set $\mu_k = y_{i_k}$

2. Obtain the consensus partition

Run the K -means algorithm using the pivots as initial cluster centers

The R package 'pivmet'

The `piv_KMeans` algorithm is implemented in the R package **pivmet** via the function `piv_KMeans()`:

- `x`: the data, in either matrix or data frame format
- `centers`: the number of clusters K

```
piv_KMeans(x, centers, alg.type = "KMeans",  
  piv.criterion = "MUS", H = 1000, ...)
```

By default `piv_KMeans` obtains a partition of `x` into the number of groups specified by `centers` via the K -means algorithm.

Then, finds K pivots from the co-association matrix derived from $H = 1000$ runs of K -means with random seeds, using the selected pivotal method. If `centers` < 5, default is MUS.

The `piv_KMeans` function

Argument	Description
<code>x</code>	The data object with N observations (a matrix, vector or data frame)
<code>centers</code>	The number of clusters in the solution
<code>alg.type</code>	The clustering algorithm for the initial partition: "KMeans" (default) or "hclust"
<code>method</code>	If <code>alg.type="hclust"</code> , the agglomeration method: "single", "complete", "average", "ward.D", "ward.D2", "mcquitty", "median", "centroid"
<code>piv.criterion</code>	The pivotal criterion: "MUS", "maxsumint", "min-sumnoint", "maxsumdiff"
<code>H</code>	The number of distinct k -means runs used for building the $N \times N$ co-association matrix. Default is 1000
<code>prec_par</code>	If <code>piv.criterion="MUS"</code> , the user-specified value of candidate pivots for each group.

Table 1: Overview of arguments for `piv_KMeans()`.

Illustration: 'Two-sticks' data

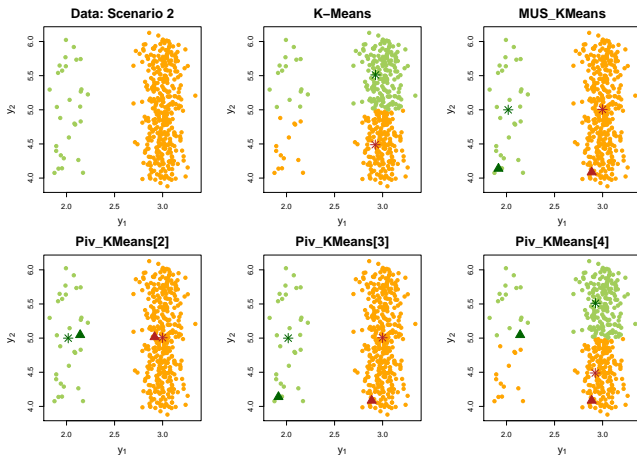


Figure 1: 'Two-sticks' data (two groups of 30 and 370 observations, respectively). Group centers (asterisks) and pivots (triangles) are shown on the plots.

Simulation settings

- We simulate 1000 bivariate datasets from a mixture of three Gaussian distributions
- The components have sample size 20, 100 and 500, respectively
- The K -means algorithm with random seeds is used to generate a cluster ensemble of dimension $H = 1000$
- The solution with $K = 3$ from hierarchical clustering (Average-Linkage) is used as reference partition

Comparing the clustering solutions

ARI: K-Means vs Piv_KMeans (1000 iterations)

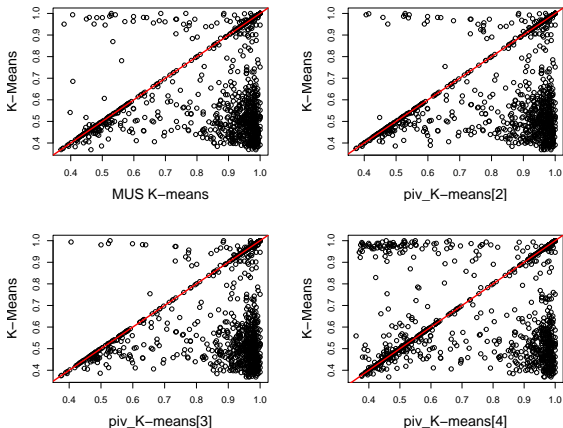


Figure 2: ARI over 1000 runs of piv_KMeans and K-means on 3-component 2D Gaussian data.

Comparison with other consensus methods

Ensemble methods based on the pairwise similarity amongst data points uses the co-association matrix $C = (c_{i,j})$ that summarizes the entire ensemble in order to define the dissimilarities

$$d_{i,j} = 1 - c_{i,j}$$

The final clustering result is generated by applying any similarity-based clustering algorithm (as a consensus function) to this matrix (e.g. agglomerative hierarchical clustering (agnes), PAM for partitioning 'around medoids')

Comparison with other consensus methods

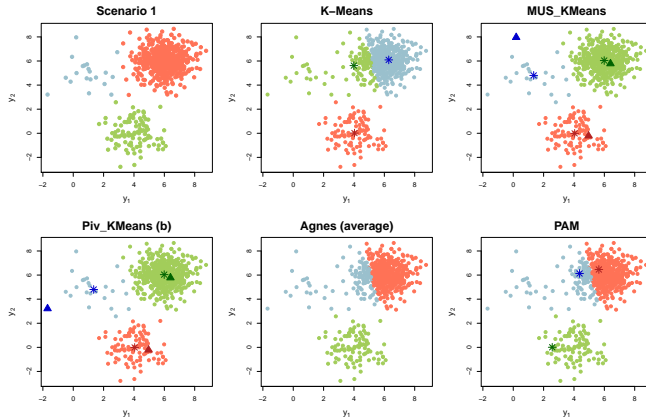


Figure 3: Mixture of three Gaussian distributions (sample size $n=620$) with mean vectors $\mu_1 = (1, 5)$, $\mu_2 = (4, 0)$, $\mu_3 = (6, 6)$, and covariance matrix the 2×2 identity matrix.

Preliminary results

The table shows the mean Adjusted Rand Index (1000 simulations) between the consensus partition and the true data partition for the 2D Gaussian data.







Pivotal methods	MUS 0.857	(a) 0.865	(b) 0.883	(c) 0.779
Other consensus methods	agnes (<i>AL</i>) 0.512	agnes (<i>SL</i>) 0.535	agnes (<i>CL</i>) 0.514	PAM 0.506

Concluding remarks

We propose a modified version of the standard K -means algorithm, by considering a pivot-based initialization step:

- The co-association derived from a cluster ensemble is used to identify pivotal units, different criteria can be used
- Simulation results reveal that a careful seeding based on pivotal units may improve the final clustering results
- Pivotal methods for relabelling and data clustering are implemented in the R package **pivmet**, available from the Comprehensive R Archive Network
- Ongoing work explores the advantage of using pivotal methods for selecting the appropriate number of groups

References

-  Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms, pp. 1027-1035 (2007)
-  Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: Relabelling in Bayesian mixture models by pivotal units. Stat. Comput. **28**(4), 957–969 (2018a)
-  Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: Maxima Units Search (MUS) algorithm: methodology and applications. In: Perna, C. , Pratesi, M., Ruiz-Gazen A. (eds.) Studies in Theoretical and Applied Statistics, pp. 71–81 (2018b)
-  Fred, A. L., Jain, A. K.: Combining multiple clusterings using evidence accumulation. IEEE Trans Pattern Anal Mach Intell **27**(6), 835–850 (2005)
-  Jain, A.: Data clustering: 50 years beyond K-means. Patt. Recog. Lett. **31**(8), 651–666 (2010)
-  Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2002)

Thank you for your attention!

Appendix

MUS Algorithm

Consider a large and sparse $n \times n$ symmetric matrix where each row (column) is a statistical unit and the units belongs to \mathcal{K} groups. Here $\mathcal{K} = 3$ and groups are $C_1 = \{1, 2, 3\}$, $C_2 = \{4, 5, 6\}$, $C_3 = \{7, 8, 9\}$.

$$X = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{matrix} & \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \end{matrix}$$

MUS Algorithm: Step (i)

For $k \in \{1, 2, 3\}$ first select \bar{m} units in the k -th group with a higher number of zeros with respect to the other groups

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

$$\mathcal{K} = 3$$

$$\bar{m} = 2 : \text{Candidates} = \{2, 3, 4, 6, 8, 9\}$$

MUS Algorithm - Step (ii)

- For each group k and each candidate i^* , consider the set of units in a **different group** with a zero in the corresponding cell (denote it by \mathcal{P}^k)
- count the identity matrices of rank \mathcal{K} that can be constructed with i^* and units in \mathcal{P}^k .

Example: Group C_1 (red), unit $i_1^* = 2$

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

MUS Algorithm - Step (ii)

- For each group k and each candidate i^* , consider the set of units in a **different group** with a zero in the corresponding cell (denote it by \mathcal{P}^k)
- count the identity matrices of rank \mathcal{K} that can be constructed with i^* and units in \mathcal{P}^k .

Example: Group C_1 (red), unit $i_1^* = 2$

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \xrightarrow{\mathcal{P}^2} \{4, 6, 8, 9\}$$

$\underbrace{\{2, (p, q) \in \mathcal{P}^2\}}_{M_2} \begin{matrix} & \overset{2}{p} & \overset{q}{q} \\ \overset{2}{p} & \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} & = I_3 \end{matrix}$

Group C_1 (red), unit $i_1^* = 3$

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

MUS Algorithm

Group C_1 (red), unit $i_1^* = 3$

$$X = \begin{pmatrix}
 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\
 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\
 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\
 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1
 \end{pmatrix} \xrightarrow{\mathcal{P}^3} \{4, 6, 8, 9\}$$

$\{3, (p, q) \in \mathcal{P}^3\}: \begin{matrix} \text{red } 3 & \text{green } p & \text{blue } q \\ \text{green } p & & \\ \text{blue } q & & \end{matrix} \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} = I_3$

M_3

MUS Algorithm

Group C_1 (red), unit $i_1^* = 3$

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \xrightarrow{\mathcal{P}^3} \boxed{\{4, 6, 8, 9\}}$$

$\underbrace{\{3, (p, q) \in \mathcal{P}^3\}: \begin{matrix} 3 & p & q \\ \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} \end{matrix}}_{M_3} = I_3$

Step (iii) For group $C_1 = \{1, 2, 3\}$, the **pivot** is $i_1^* = 2$ if $M_2 > M_3$, $i_1^* = 3$, otherwise.