Non-parametric Clustering of Multivariate Populations with Arbitrary Sizes

Yves I Ngounou Bakam, Denys Pommeret

▶ **To cite this version:**

Yves I Ngounou Bakam, Denys Pommeret. Non-parametric Clustering of Multivariate Populations with Arbitrary Sizes. 2022. hal-03849492

HAL Id: hal-03849492
https://hal.science/hal-03849492

Preprint submitted on 11 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Non-parametric Clustering of Multivariate Populations with Arbitrary Sizes

BY

Yves I. Ngounou Bakam and Denys Pommeret

Abstract

We propose a clustering procedure to group K populations into subgroups with the same dependence structure. The method is adapted to paired population and can be used with panel data. It relies on the differences between orthogonal projection coefficients of the K density copulas estimated from the K populations. Each cluster is then constituted by populations having significantly similar dependence structures.A recent test statistic from Ngounou-Bakam and Pommeret (2022) is used to construct automatically such clusters. The procedure is data driven and depends on the asymptotic level of the test. We illustrate our clustering algorithm via numerical studies and through two real datasets: a panel of financial datasets and insurance dataset of losses and allocated loss adjustment expense.

KEYWORDS

Copula coefficients, data-driven, Legendre polynomials, nonparametric clustering, smooth test.

1 INTRODUCTION AND MOTIVATIONS

The knowledge of the companies that dominate the capitalization of international stock markets and their classification can allow portfolio managers a much more active strategy and a better diversification of risks. In particular, the knowledge of their dependence structure makes it possible to group together various portfolios with similar risks.

In this paper, we propose a data-driven strategy to regroup portfolios or risks having the same dependence structure. Their similarities are measured through their copulas and our procedure is based on simultaneous multiple comparisons. The implementation of our clustering procedure therefore requires a multiple comparison method that has been introduced in Ngounou-Bakam and Pommeret (2021). This K-sample test is a data-driven procedure with a chi-square limit distribution making the method very fast and very easy to implement. The algorithm is based on this test procedure and is also data-driven, depending only on the asymptotic level of the test. The basic idea of this algorithm is to use the test statistics to measure the proximity between populations. If the statistics are close, it is proposed to form a cluster with their associated populations and the test procedure accepts or rejects the validity of the cluster.

Our method applies to $K(\geq 2)$ iid sample observed on K populations, eventually paired. This is the case in the considered problem of portfolios. Our approach differs from recent based copulas clustering algorithms as for instance: the clustering methods which rely on hierarchical Kendall copula with Archimedean clusters (see Su et al. (2019), Joe and Sang (2016), among others); a clustering algorithm based on the likelihood of the copula, called CoClust, which has been introduced in Di Lascio and Giannerini (2012) and further developed and implemented in Di Lascio and Giannerini (2017); Di Lascio (2018); Di Lascio and Giannerini (2019); the clustering algorithms where an iid sample from a finite mixture model is usually considered (see for instance Kosmidis and Karlis (2016); Zhang and Baek (2019) and reference therein); the approach which relies on time-varying copula-based estimators via minimization of the value-at-risk (see De Luca and Zuccolotto (2017)) and the copula-based fuzzy clustering algorithm for spatial time series, called COFUST (see Disegna et al. (2017)). All these previous works concern parametric copulas and classify each individual and not populations.

A numerical study first shows the good behaviour of the proposed clustering algorithm. We then apply the procedure on financial assets [a detailler un peu le ou les jeux de donnees ici].

The paper is organized as follows: in Section 2 we set up notation and we recall the main result of the the test statistic presented in Ngounou-Bakam and Pommeret (2022), making the paper self-contained. Section 3 presents the clustering algorithm. Section 4 is devoted to the numerical study and Section 5 contains two real-life illustrations.

2 NOTATION AND TEST PROCEDURE

Let briefly recall here the test procedure proposed in Ngounou-Bakam and Pommeret (2022). Let $\mathbf{X} = (X_1, \cdots, X_p)$ be a p-dimensional continuous random variable with joint probability distribution function (pdf) $F_{\mathbf{X}}$ that can be expressed in terms of copula as

$$F_{\mathbf{X}}(x_1, \cdots, x_p) \quad = \quad C(F_1(x_1), \cdots, F_p(x_p)), \tag{1}$$

where F_j denotes the marginal pdf of X_j, and C denotes the copula associated to $\mathbf{X}$. Writing

$$U_j = F_j(X_j), \qquad \text{for } j = 1, \cdots, p,$$

we have for all $u_j \in (0, 1)$

$$C(u_1, \cdots, u_p) \quad = \quad F_{\mathbf{U}}(u_1, \cdots, u_p),$$

with $\mathbf{U} = (U_1, \cdots, U_p)$, and deriving this expression p times with respect to $u_1, \cdots, u_p$, we get an expression of the density copula

$$c(u_1, \cdots, u_p) \quad = \quad f_{\mathbf{U}}(u_1, \cdots, u_p), \tag{2}$$

where $f_{\mathbf{U}}$ denotes the joint density of the vector $\mathbf{U}$. Write $\mathcal{L} = \{L_n; n \in \mathbb{N}\}$ the set of orthogonal Legendre polynomials (see Appendix **??** for more detail). Write $\mathbf{j} = (j_1, \cdots, j_p) \in \mathbb{N}^p$ and define

$$\rho_{j_1, \cdots, j_p} \quad = \quad \mathbb{E}(L_{j_1}(U_1) \cdots L_{j_p}(U_p)), \tag{3}$$

the $\mathbf{j}$-th *copula coefficient* associated to $\mathbf{U}$. Note that $\rho_{\mathbf{0}} = 1$ where $\mathbf{0} = (0, \cdots, 0)$, and $\rho_{\mathbf{j}} = 0$ if only one element of $\mathbf{j}$ is non null.

The sequence $(\rho_{\mathbf{j}})_{\mathbf{j} \in \mathbb{N}_*^p}$ permits to summarize the copula and we propose a clustering procedure based on the distances between these coefficients.

In this way assume that we observe K iid samples, possibly paired, with associated copulas denoted by $C_1, \cdots, C_K$.

Our aim is to regroup populations having the same copula coefficients, that is, satisfying the following equality

$$H_0 : \rho_{\mathbf{j}}^{(i_1)} = \quad \cdots \quad = \rho_{\mathbf{j}}^{(i_k)}, \quad \forall \mathbf{j} \in \mathbb{N}_*^p, \tag{4}$$

where $i_1, \cdots, i_k$ are the label of the tested populations and $\rho^{(i_k)}$ stands for the copula coefficients associated to C_{i_k}. Clearly if $C_1 = \cdots = C_K$ then H_0 is immediately satisfied. In order to implement our clustering algorithm we propose to use the

statistic based on the estimation of these quantities proposed in Ngounou-Bakam and Pommeret (2022).

We denote by

$$\mathbf{X}^{(1)} = (X_1^{(1)}, \cdots, X_p^{(1)}), \cdots, \mathbf{X}^{(K)} = (X_1^{(K)}, \cdots, X_p^{(K)}),$$

the K continuous random variables associated to the K populations, with joint cumulative distribution function (cdf) $\mathbf{F}^{(1)}, \cdots, \mathbf{F}^{(K)}$, and with associated copulas $C_1, \cdots, C_K$, respectively. Assume that we observe K iid samples from $\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}$, possibly paired, denoted by

$$(X_{i,1}^{(1)}, \cdots, X_{i,p}^{(1)})_{i=1,\cdots,n_1}, \cdots, (X_{i,1}^{(K)}, \cdots, X_{i,p}^{(K)})_{i=1,\cdots,n_K}.$$

We assume that

$$\text{for all } 1 \le k < \ell \le K, \qquad n_k/(n_k + n_\ell) \to a_{k\ell}, \text{ with } 0 < a_{k\ell} < \infty. \tag{5}$$

We will denote by $F_j^{(k)}$ the marginal cdf of the jth component of $\mathbf{X}^{(k)}$ and we write

$$U_{i,j}^{(k)} = F_j^{(k)}(X_{i,j}^{(k)}).$$

For testing (4) we first estimate the copula coefficients by

$$\widehat{\rho}_{j_1\cdots j_p}^{(k)} \;=\; \frac{1}{n_k}\sum_{i=1}^{n_k} L_{j_1}(\widehat{U}_{i,1}^{(k)}) \cdots L_{j_p}(\widehat{U}_{i,p}^{(k)})), \tag{6}$$

where

$$\widehat{U}_{i,j}^{(k)} = \widehat{F}_j^{(k)}(X_{i,j}^{(k)}),$$

and where $\widehat{F}$ denotes the empirical distribution functions associated to F.

Considering the null hypothesis H_0 as expressed in (4), the test procedure is based on the sequences of differences

$$r_{\mathbf{j}}^{(\ell,m)} := \widehat{\rho}_{\mathbf{j}}^{(\ell)} - \widehat{\rho}_{\mathbf{j}}^{(m)}, \qquad \text{for } 1 \le \ell \le m \le K, \text{ and } \mathbf{j} \in \mathbb{N}_*^p,$$

with the convention that $r_{\mathbf{j}}^{(\ell,m)} = 0$ when only one element of $\mathbf{j}$ is different of zero.

In order to select automatically the number of copula coefficients, for any vector $\mathbf{j} = (j_1, \cdots, j_p)$ we denote by

$$\|\mathbf{j}\|_1 = |j_1| + \cdots + |j_p|,$$

its L^1 norm and for any integer $d > 1$ we write

$$\mathcal{S}(d) \quad = \quad \{\mathbf{j} \in \mathbb{N}^p; \|\mathbf{j}\|_1 = d \text{ and there exists } k \neq k' \text{ such that } j_k > 0 \text{ and } j_{k'} > 0\}.$$

The set $\mathcal{S}(d)$ contains all non null positive integers $\mathbf{j} = (j_1, \cdots, j_p)$ with norm d and such that $j_k < d$ for all $k = 1, \cdots, p$.

We also introduce the following set of indexes:

$$\mathcal{V}(K) \quad = \quad \{(\ell, m) \in \mathbb{N}^2; 1 \leq \ell < m \leq K\}.$$

Clearly $\mathcal{V}(K)$ contains $v(K) = K(K-1)/2$ elements which represent all the pairs of populations that we want to compare.

We construct an embedded series of statistics as follows

$$V_1 = V_{D(n)}^{(1,2)}, \quad V_2 = V_{D(n)}^{(1,2)} + V_{D(n)}^{(1,3)}, \quad \cdots, \quad V_{v(K)} = V_{D(n)}^{(1,2)} + \cdots + V_{D(n)}^{(K-1,K)},$$

or equivalently,

$$V_k \quad = \quad \sum_{(\ell,m) \in \mathcal{V}(K); rank_{\mathcal{V}}(\ell,m) \leq k} V_{D(n)}^{(\ell,m)},$$

where

$$V_k^{(\ell,m)} = n \sum_{\mathbf{j} \in \mathcal{H}(k)} (r_{\mathbf{j}}^{(\ell,m)})^2 \tag{7}$$

where the set $\mathcal{H}(k)$ contains the k first integers of $\mathbb{N}^p$ with respect to the order of $\mathcal{S}(d)$ and where

$$D(n) := \min \Big\{ \underset{1 \leq k \leq d(n)}{\operatorname{argmax}} (V_k^{(1,2)} - kq_n) \Big\}, \tag{8}$$

where q_n and $d(n)$ tend to $+\infty$ as $n \to +\infty$, kq_n being a penalty term which penalizes the embedded statistics proportionally to the number of copula coefficients used.

Moreover, we have the following relation: for all $k \geq 1$ and $j = 1, \cdots, c(k+1)$

$$V_{c(1)+c(2)+\cdots+c(k)+j}^{(1,2)} = T_{k+1,j}^{(1,2)},$$

were $c(k)$ denotes the cardinal of $\mathcal{S} = (\lceil)$ with the convention $c(1) = 0$.

We have $V_1 < \cdots < V_{v(K)}$. The first statistic V_1 compares the first two populations 1 and 2. The second statistic V_2 compares the populations 1 and 2, and, in addition, the populations 1 and 3. And so on. To choose automatically the appropriate number k we introduce the following penalization procedure, mimicking the Schwarz criteria procedure Schwarz (1978):

$$s(\mathbf{n}) = \min \Big\{ \underset{1 \leq k \leq v(K)}{\operatorname{argmax}} \big(V_k - kp_n \big) \Big\}. \tag{9}$$

We make the following assumption:

(A) $d(n)^{(p+4)} = o(q_n)$,

(A') $d(n)^{(p+4)} = o(p_n)$,

and we recall here main result of Ngounou-Bakam and Pommeret (2022).

Theorem 2.1. *Assume that* **(A)** *and* **(A')** *hold. Then under* H_0, $s(\mathbf{n})$ *converges in probability towards 1 as* $n \to +\infty$. *Moreover,* $V_{s(\mathbf{n})}/\widehat{\sigma}^2(1,2))$ *converges in law towards a* χ^2_1 *distribution, where* $\widehat{\sigma}^2(1,2)$ *is given in Appendix.*

It is important to note that if $p_n = o(n)$ then the test is consistent against alternative where at least one copula coefficient differs between two copulas.

3 CLUSTERING PROCEDURE

3.1 Clustering principle

In the sequel we propose to adapt the previous test procedure to obtain a data-driven method to cluster K populations into N subgroups characterized by a common dependence structure. The number N of clusters is unknown and will be automatically chosen by the previous procedure and validated by our testing method.

More precisely, assume that we observe K iid samples from K populations, possibly paired. The clustering algorithm starts by choosing the two populations that are the most similar in terms of dependence structure, through their copulas. In this way, it chooses the smaller two-sample statistic. If the equality of both associated copulas is accepted these two populations form the first cluster. Then the algorithm proposes the closer population of this cluster, that is the smaller statistic having a common population index. While the test accepts the simultaneous equality of the copulas, the cluster growths. If the last test is rejected then the cluster is closed and the last rejected population forms a new cluster. One can iterate this several times until every sample is associated with a cluster.

3.2 Clustering algorithm

We can summarize the clustering algorithm as follows:

Algorithm: K-sample copulas clustering

Initialization: $c = 1$. Set $S = \{C_1, \cdots, C_K\}$ and $S_0 = \emptyset$;
Select $\{\ell^\star, m^\star\} = \text{argmin}\{V_{D(n)}^{(\ell,m))}; \ell \neq m \in S \setminus \bigcup_{k=1}^{c} S_{k-1}\}$;
Test $\tilde{H}_0$ between all $\rho_{\mathbf{j}}^{(\ell^\star)}$ and $\rho_{\mathbf{j}}^{(m^\star)}$;
if $\tilde{H}_0$ *is not rejected* **then**
    $S_1 = \{C_{\ell^\star}, C_{m^\star}\}$;
else
    STOP. There is no cluster.
end
while $S \setminus \bigcup_{k=1}^{c} S_k \neq \emptyset$ **do**
    Select $\{j^\star\} = \text{argmin}\{T_{D(n)}^{(i,j)}; i \in S_c, j \in S \setminus \bigcup_{k=1}^{c} S_k\}$;
    Test $\tilde{H}_0$ the simultaneous equality of all the $\rho_{\mathbf{j}}^{(i)}$, $i \in S_c$ and $\rho_{\mathbf{j}}^{(j^\star)}$;
    if $\tilde{H}_0$ *not rejected* **then**
        $S_c = S_c \bigcup \{C_{j^\star}\}$;
    else
        $S_{c+1} = \{C_{j^\star}\}$;
        $c = c + 1$;
    end
end

This clustering procedure can solve several complex problems in a very short time and is useful in practice, particularly in risk management and more generally in the world of actuarial science and finance markets by making it possible to detect mutualizable risks and not mutualizable; but also to build a well-diversified portfolio.

3.3 Tuning the algorithm

As evoked in Remark **??** we can choose the penalty $q_n = p_n = \alpha \log(n)$. We fix $\alpha = 1$ in the proofs of this paper for simplicity. But in practice we can empirically improve this tuning factor by using the following data-driven procedure:

- Assume we observe K populations.
- We merge all populations to get only one (larger) population.
- Split randomly this population into $K' > 2$ sub-populations.

- Clearly these K' sub-populations have the same copula and then the null hypothesis $\tilde{H}_0$ is satisfied.
- We then approximate numerically the value of the factor $\alpha > 0$ such that the selection rule retains the first component, that is $s(\mathbf{n}) = 1$. From Theorem **??** this is the asymptotic expected value under the null.
- We can repeat N times such a procedure to get N K'-sample under the null.

Finally we fix

$$\widehat{\alpha} = \min\{\alpha > 0;\ \text{such that } s(\mathbf{n}) = 1 \text{ for the previous } N \text{ selection rules}\}.$$

In our simulation we fixed arbitrarily $K' = 3$, which seems to give a very correct empirical level. Note that this transformation only slightly modified the empirical results.

Concerning the value of $d(n)$, the condition **(A)** is an asymptotic condition and from our experience choosing $d(n) = 3$ or 4 is enough to have a very fast procedure which detects alternatives such that copulas differ by a coefficient with a norm less or equal to $d(n)$.

4 NUMERICAL STUDY OF THE ALGORITHM

4.1 Simulation design

In order to evaluate the performance of the algorithm, we consider the following classical copulas families: the Gaussian copulas, the Student copulas, the Gumbel copulas, the Frank copulas, the Clayton copulas and the Joe copulas which we denote for hereafter *Gaus*, *Stud*, *Gumb*, *Fran*, *Clay* and *Joe* respectively. For the explicit functional forms and properties of these copulas we refer the reader to Nelsen (2007) and **?**. For each copula C, the sample is generated with a given kendall's τ parameter, and we denote this model briefly by $C(\tau)$. When τ is close to zero the variables are close to the independence. Conversely, if τ is close to 1 the dependence becomes linear.

4.2 Clustering simulation

We consider the following designs:

- **A100**: $n = 100$, $p = 3$, $K = 6$ populations with 3 groups $C_1 = Gumb(0.8)$ and $C_2 = C_3 = Gaus(0.2)$ and $C_4 = C_5 = C_6 = Clay(0.9)$

- **A500** = **A100** with $n = 500$
- **B100**: $n = 100$, $p = 5$, $K = 4$ different populations with 4 groups $C_1 = Gumb(0.8)$, $C_2 = Gaus(0.2)$, $C_3 = Clay(0.9)$, $C_4 = Gumb(1)$
- **B500** = B100 with $n = 500$
- **C100**: $n = 100$, $p = 4$, $K = 5$ populations with one group $C^{(1)} = C^{(2)} = C^{(3)} = C^{(4)} = C^{(5)} = Clay(0.9)$
- **C500** = C100 with $n = 500$
- **D100**: $n = 100$, $p = 2$, $K = 10$ populations with two unbalanced groups $C_1 = C_2 = \cdots = C_9 = Clay(0.9)$ and $C_{10} = Gumb(0.9)$
- **D500** = D100 with $n = 500$

We applied the clustering algorithm described in Section 3. The results are summarized below:

- Results for **A100**:
  - In 82.5 % of cases the algorithm found 3 groups. In such cases, 74 % of the time it was the 3 correct groups.
  - In 11.4 % of cases the algorithm found 4 groups
  - In 5 % of cases the algorithm found 2 groups
  - In 0.1 % of cases the algorithm found 5 groups.
  - Note that the first group (with the Gumbel copula) was well identified 99 % of the time.
- Results for **A500**: The three groups were well identified in 92 % of cases. In other cases the algorithm essentially obtained 4 groups (merging populations of the second and the third group).
- Results for **B100**: In 78 % of cases the null hypothesis was rejected and we obtained 4 different groups. In other cases the algorithm merged two groups (Clayton with Normal or Clayton with Gumbel) and then proposed 3 clusters.
- Results for **C100**: In 70 % of cases the algorithm found one group. In other cases it gave two groups.
- Results for **D100**: More than 80% of cases the algorithm found the 2 correct groups. In other cases the algorithm found 3 group obtained by a rejection of one of the 9 similar populations.

	1	2	3	4	5	6
1	100	0	0	0	0	0
2		100	100	0	0	0
3			100	0	0	0
4				100	100	100
5					100	100
6						100

	1	2	3	4	5	6
1	100	0	0	0	0	0
2		100	73	29	30	29
3			100	22	25	21
4				100	78	82
5					100	79
6						100

Table 1: Population associations (in %) under model **A100** (n=100). Left: theoretical; Right: observed. The true associations are $\{1\}; \{2, 3\}; \{4, 5, 6\}$.

	1	2	3	4	5	6
1	100	0	0	0	0	0
2		100	100	0	0	0
3			100	0	0	0
4				100	100	100
5					100	100
6						100

	1	2	3	4	5	6
1	100	0	0	0	0	0
2		100	93	0	0	0
3			100	0	0	0
4				100	99	99
5					100	100
6						100

Table 2: Population associations (in %) under model **A500** (n=500). Left: theoretical; Right: observed. The true associations are $\{1\}; \{2, 3\}; \{4, 5, 6\}$.

	1	2	3	4
1	100	0	0	0
2		100	0	0
3			100	0
4				100

	1	2	3	4
1	100	0	0	0
2		100	12	11
3			100	10
4				100

Table 3: Population associations (in %) under model **B100**. Left: theoretical; Right: observed. The true associations are $\{1\}; \{2\}; \{3\}; \{4\}$.

	1	2	3	4	5
1	100	100	100	100	100
2		100	100	100	100
3			100	100	100
4				100	100
5					100

	1	2	3	4	5
1	100	97.9	98.5	98.9	99.2
2		100	99.6	98.4	99.7
3			100	99.4	99.1
4				100	98.7
5					100

Table 4: Population associations (in %) under model **C100**. Left: theoretical; Right: observed. The true associations are $\{1, 2, 3, 4, 5\}$.

	1	2	3	4	5	6	7	8	9	10
1	100	100	100	100	100	100	100	100	100	0
2		100	100	100	100	100	100	100	100	0
3			100	100	100	100	100	100	100	0
4				100	100	100	100	100	100	0
5					100	100	100	100	100	0
6						100	100	100	100	0
7							100	100	100	0
8								100	100	0
9									100	0
10										100

	1	2	3	4	5	6	7	8	9	10
1	100	90.2	89.8	91	94.2	90.5	92	97.1	89	0
2		100	94.1	92	89.9	88.7	91.3	90.9	92	0
3			100	94.4	92.2	95.6	88	97.4	90	0
4				100	91	95.5	89.1	90	93.3	0
5					100	94	88.5	96	97	0
6						100	89.9	91.2	88.2	0
7							100	87	97.1	0
8								100	96	0
9									100	0
10										100

Table 5: Population associations (in %) under model **D100**. Up: theoretical; Down: observed. The true associations are $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}; \{10\}$.

5 REAL DATASETS

5.1 Financial data

The knowledge of the companies that dominate the capitalization of international stock markets and their classification can allow portfolio managers a much more active strategy and a better diversification of risks. We build 33 portfolios.

From the 500 component stocks of the $S\&P500$ stock market index, which are issued by 500 large capitalization companies traded on American stock exchanges, we choose in each sector following the Global Industry Classification (there exists 11 stock market sectors)

- the stocks index of the 3 most high weighted companies. We denote $Sih, i = 1, \cdots, 11$ hereafter,
- the stocks index of the 3 companies with the middle weight. We denote $Sim, i = 1, \cdots, 11$ hereafter,
- the stocks index of the 3 companies with the lowest weight. We denote $Si\ell, i = 1, \cdots, 11$ hereafter.

Table 6 presents the weight, symbol, company and sector of each selected stock index.

The data employed are weekly closing adjusted prices from January, 26^{th}, 2006 to December, 30^{th}, 2021 for a total of 825 observations. Data are available from Yahoo Finance and we consider the rate returns series by using the standard continuously compounded return formula. We note that each price of stock is expressed in the reference country currency.

The application of non-parametric tests of randomness (Wang (2003); Cho and White (2011); Gibbons and Chakraborti (2014)) to these weekly rates of return for each of the 33 stocks in Table 6 reveals that there is no evidence that these series are not iid.

We begin by considering the populations (denoted *pop* in Table 6) of each group (high (h), middle (m) and lower (ℓ)).

Applying the clustering procedure with nominal level $\alpha = 5\%$, we obtain 6,4 and 8 clusters of group ℓ, group m and group h, respectively. The Figures 1, 2 and 3 displays the dendrogram of groups (Grp.) ℓ, m and h respectively. In the three dendrogram we observe that the sector Material is isolated. Moreover at 1% level (see Figures **??** **??** and **??**), the number of clusters and the elements of each cluster remain unchanged. But it is clear that moving this level is an interesting way to reduce or increase the number of clusters.

By looking at the three groups, we now ask whether if there are populations in different groups of similar dependence structure. To this end, we apply the clustering algorithm to all 33 populations with 5% nominal level. We get 12 clusters of populations and the associated dendrogram is presented in Figure 4. We observe that clusters $C4$,$C5$ and $C9$ contain only the populations of group ℓ and clusters $C8$, $C11$ and $C12$ only the populations of group h.

We thus obtain a way to group stocks with the same dependence structure into homogeneous portfolios, while forcing these portfolios not to have the same behavior. This allows for risk diversification, for example.

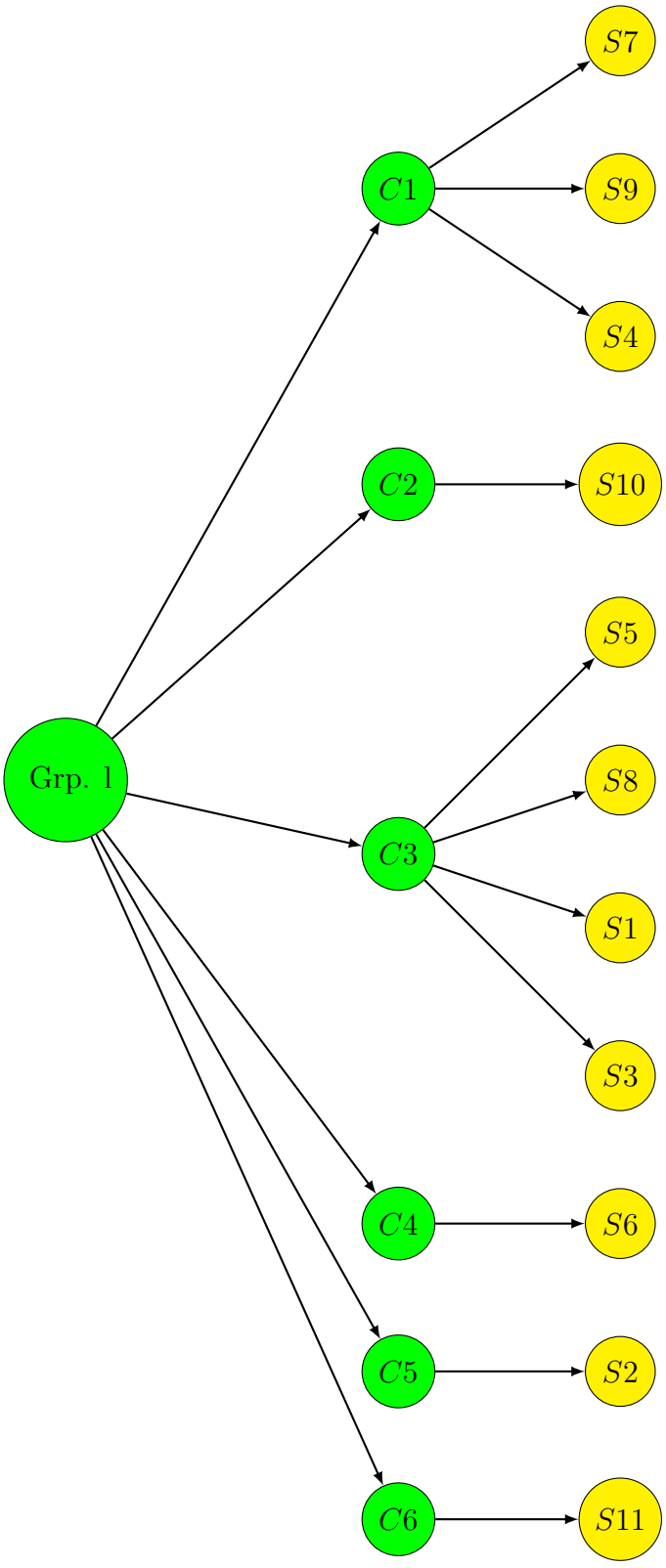

Figure 1: Clustering of group ℓ at 5% level. $c_1, \cdots, c_6$ denote the clusters and $s_1, \cdots, s_{11}$ are defined in Table 6.

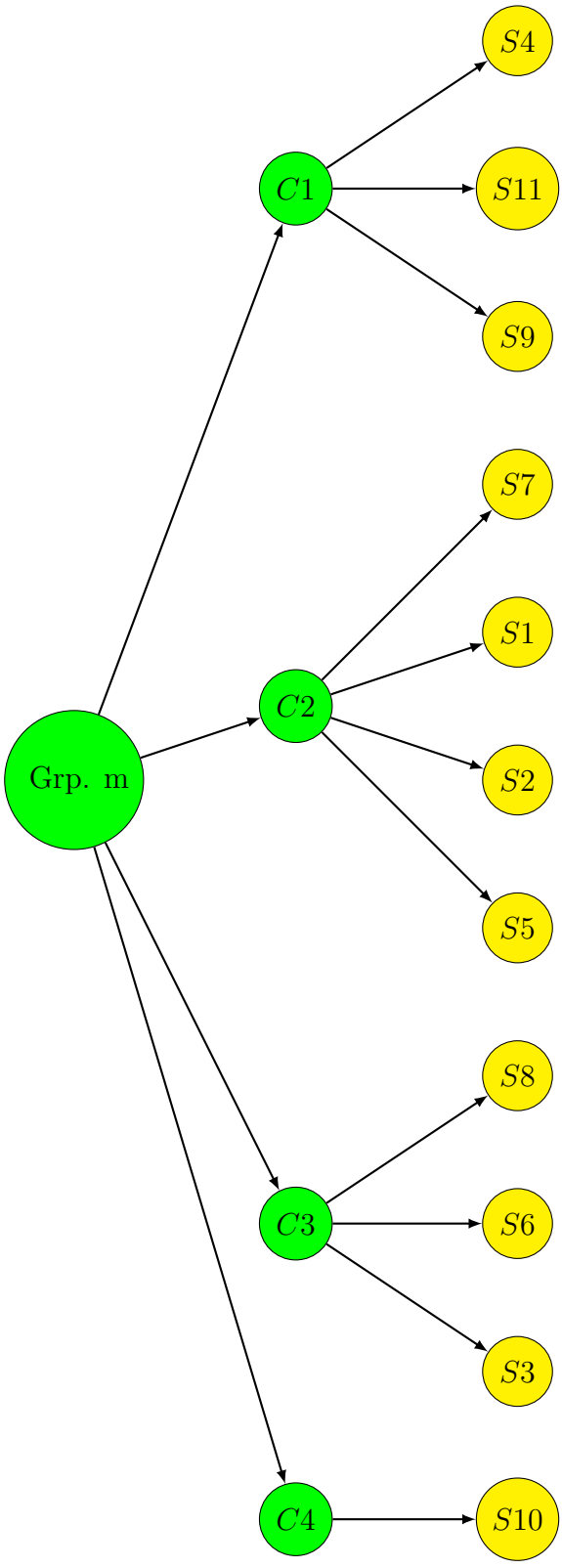

Figure 2: Clustering of group m at 5% level. $c_1 \cdots c_4$ denote the cluster and $s_1 \cdots s_{11}$ are defined in Table 6.

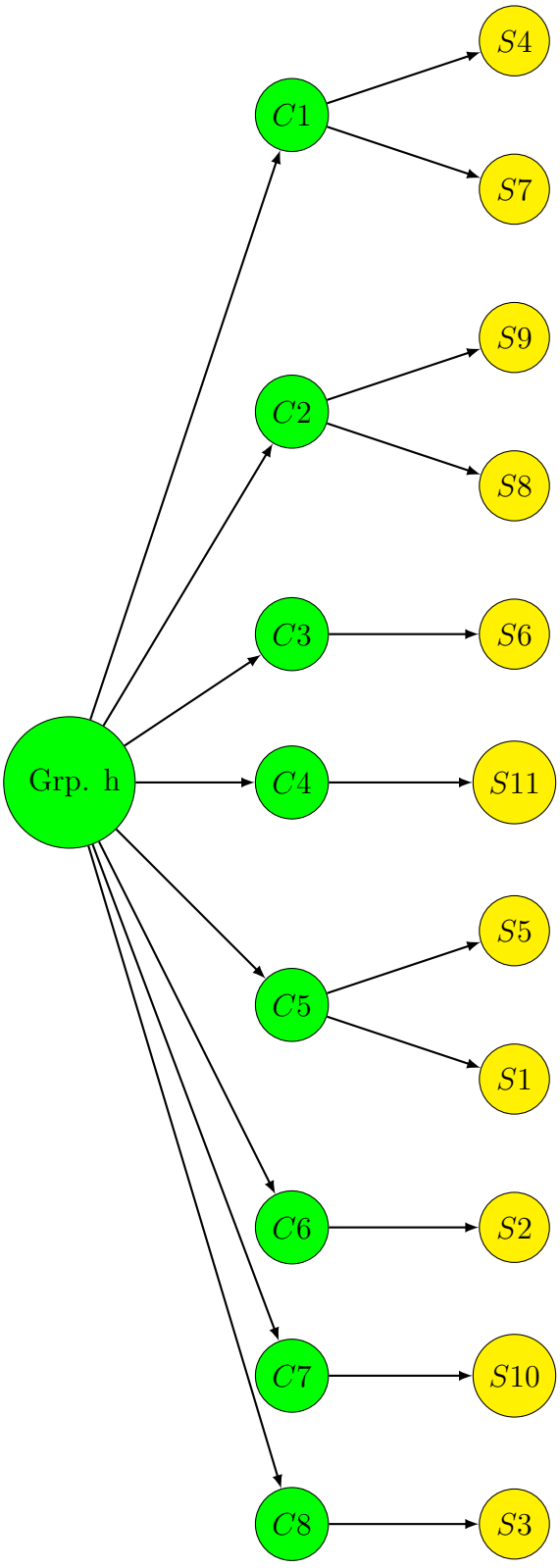

Figure 3: Clustering of group h at 5% level. $c_1 \cdots c_8$ denote the cluster and $s_1 \cdots s_{11}$ are defined in Table 6

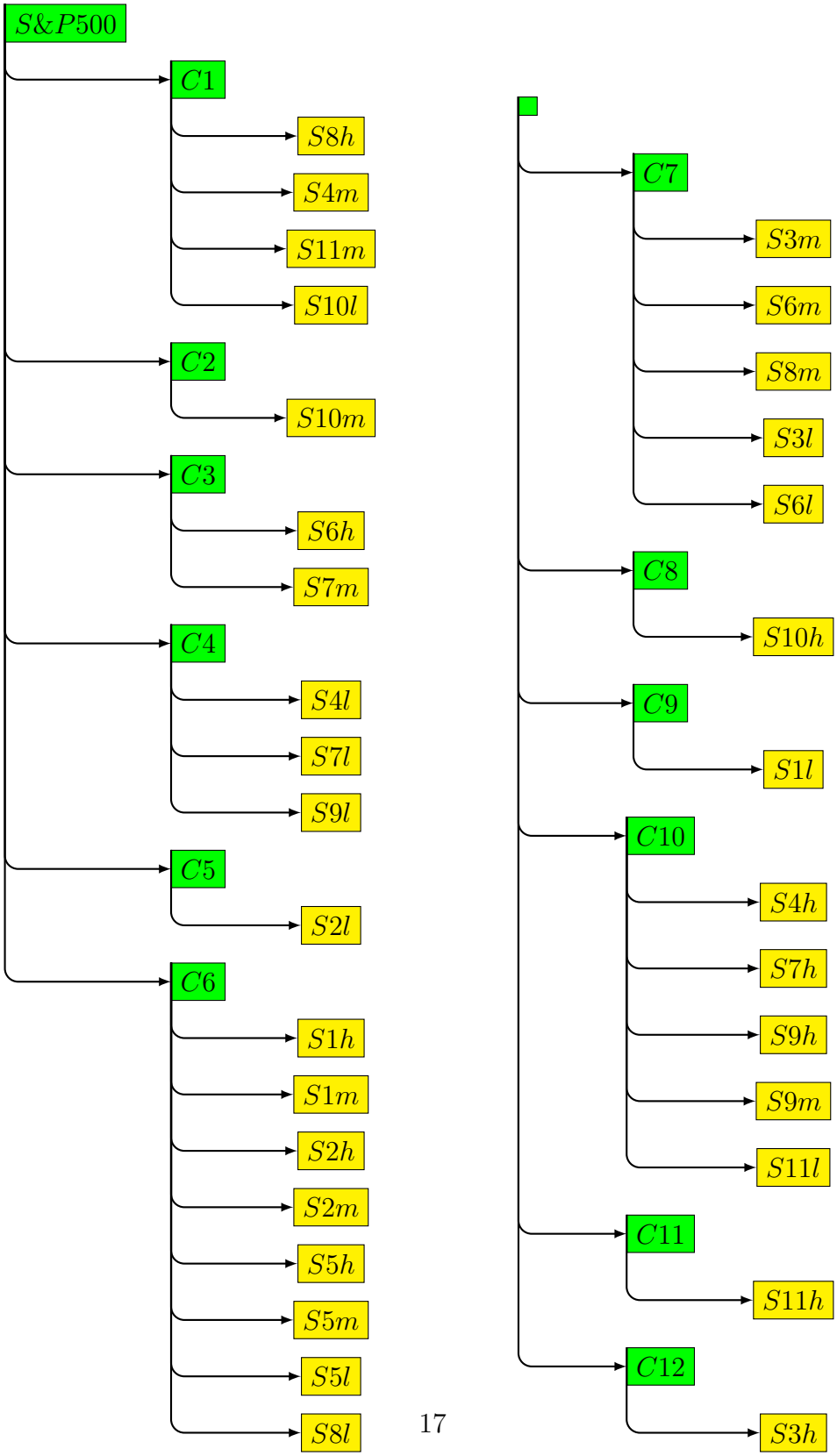

Figure 4: Clustering of $S\&P$ 500 at 5% level. $c_1 \cdots c_{12}$ denote the cluster and the populations are defined in Table 6.

Sectors	Pop	Symbols	Companies	Weights	Sectors	Pop	Symbols	Companies	Weights
Information Technology	S1h	AAPL	Apple Inc.	659.67	Consumer Discretionary	S2h	AMZN	Amazon.com Inc.	286.65
		MSFT	Microsoft Corporation	582.47			HD	Home Depot Inc.	91.66
		NVDA	NVIDIA Corporation	133.75			MCD	McDonald's Corporation	53.55
	S1m	PAYX	Paychex Inc.	11.26		S2m	GPC	Genuine Parts Company	5.65
		CDNS	Cadence Design Systems Inc.	12.26			BBY	Best Buy Co. Inc.	5.17
		MCHP	Microchip Technology Incorporated	11.47			POOL	Pool Corporation	4.62
	S1l	FFIV	F5 Inc.	2.85		S2l	PENN	Penn National Gaming Inc.	1.52
		JNPR	Juniper Networks Inc.	2.88			RL	Ralph Lauren Corporation Class A	1.36
		DXC	DXC Technology Co.	2.50			PVH	PVH Corp.	1.44
Communication Services	S3h	GOOGL	Alphabet Inc. Class A	192.14	Financials	S4h	JPM	JPMorgan Chase & Co.	110.32
		GOOG	Alphabet Inc. Class C	178.22			BAC	Bank of America Corp	74.88
		VZ	Verizon Communications Inc.	61.40			WFC	Wells Fargo & Company	50.74
	S3m	WBD	Warner Bros. Discovery Inc. Series A	11.78		S4m	MTB	M&T Bank Corporation	9.19
		EA	Electronic Arts Inc.	11.14			AMP	Ameriprise Financial Inc.	8.82
		MTCH	Match Group Inc.	6.42			TROW	T. Rowe Price Group	8.47
	S3l	DISH	DISH Network Corp. Class A	1.57		S4l	ZION	Zions Bancorporation N.A.	2.46
		LUMN	Lumen Technologies Inc.	3.27			BEN	Franklin Resources Inc.	2.15
		IPG	Interpublic Group of Companies Inc.	3.60			IVZ	Invesco Ltd.	1.89
Health Care	S5h	UNH	UnitedHealth Group Incorporated	135.88	Consumer Staples	S6h	PG	Procter & Gamble Company	101.38
		JNJ	Johnson & Johnson	135.63			KO	Coca-Cola Company	71.55
		PFE	Pfizer Inc.	86.10			PEP	PepsiCo Inc.	67.62
	S5m	BAX	Baxter International Inc.	10.79		S6m	SYY	Sysco Corporation	12.25
		A	Agilent Technologies Inc.	11.20			STZ	Constellation Brands Inc. Class A	11.49
		IDXX	IDEXX Laboratories Inc.	9.61			KR	Kroger Co.	10.08
	S5l	UHS	Universal Health Services Inc. Class B	2.59		S6l	HRL	Hormel Foods Corporation	3.90
		XRAY	DENTSPLY SIRONA Inc.	2.46			TAP	Molson Coors Beverage Company Class B	2.96
		DVA	DaVita Inc.	1.71			CPB	Campbell Soup Company	2.79
Energy	S7h	XOM	Exxon Mobil Corporation	117.49	Industrials	S8h	UNP	Union Pacific Corporation	40.32
		CVX	Chevron Corporation	97.79			RTX	Raytheon Technologies Corporation	41.10
		COP	ConocoPhillips	42.47			HON	Honeywell International Inc.	38.30
	S7m	OXY	Occidental Petroleum Corporation	17.82		S8m	RSG	Republic Services Inc.	8.15
		VLO	Valero Energy Corporation	15.28			ODFL	Old Dominion Freight Line Inc.	7.02
		WMB	Williams Companies Inc.	12.90			LUV	Southwest Airlines Co.	7.68
	S7l	CTRA	Coterra Energy Inc.	8.18		S8l	AOS	A. O. Smith Corporation	2.31
		MRO	Marathon Oil Corporation	6.94			ROL	Rollins Inc.	2.31
		APA	APA Corp.	4.91			ALK	Alaska Air Group Inc.	1.71
Utilities	S9h	NEE	NextEra Energy Inc.	43.23	Materials	S10h	LIN	Linde plc	48.08
		DUK	Duke Energy Corporation	24.95			SHW	Sherwin-Williams Company	18.91
		SO	Southern Company	22.94			NEM	Newmont Corporation	15.57
	S9m	ES	Eversource Energy	9.09		S10m	PPG	PPG Industries Inc.	8.75
		DTE	DTE Energy Company	7.38			ALB	Albemarle Corporation	8.98
		EIX	Edison International	7.54			BALL	Ball Corporation	6.83
	S9l	AES	AES Corporation	4.24		S10l	AVY	Avery Dennison Corporation	4.08
		NI	NiSource Inc	3.51			EMN	Eastman Chemical Company	4.02
		PNW	Pinnacle West Capital Corporation	2.53			SEE	Sealed Air Corporation	2.71
Real Estate	S11h	AMT	American Tower Corporation						33.82
		PLD	Prologis Inc.						26.80
		CCI	Crown Castle International Corp						23.70
	S11m	EQR	Equity Residential						7.55
		ARE	Alexandria Real Estate Equities Inc.						6.98
		VTR	Ventas Inc.						6.51
	S11l	REG	Regency Centers Corporation						2.99
		FRT	Federal Realty Investment Trust						2.30
		VNO	Vornado Realty Trust						1.59

Table 6: 33 components of S&P500

5.2 Insurance data

Insurance is an area in which the knowledge of the dependence structure between several portfolios can be useful in pricing particularly for risk pooling or price segmentation. As an illustration purposes, we consider the well-known example of pricing insurance contracts involving pairs of dependent variables which consist to compute the premium of a reinsurance treaty on a policy with unlimited liability, some retention level of the losses and a prorata sharing of ALAEs. ALAEs in this context are types of insurance company expenses that are specifically attributable to the settlement of individual claims such as lawyers' fees and claims investigation expenses. The database at issue is the SOA Group Medical Insurance Large Claims Database over the period 1991–92 and is available online at the web page of Society of Actuaries. The database includes more than 171,000 claims of $25,000$ or more, representing over $10 billion in total charges with information collected from 26 insurers. Each row of the database presents a summary of claims for an individual claimant in fields. Fields include diagnosis, type of coverage (HMO, PPO, Indemnity, etc.), claimant status (E-employee or D-dependent), claimant gender (M-male or F-Female) claimant age and charges split into hospital and non-hospital. We refer to Grazier and G'Sell (1997) for a detailed and thorough description of the data. Here, we deal with the 1991 data of females, insured by a Preferred Provider Organization (PPO) plan. We split the variables losses (hospital charges) and ALAEs (other charges) by ten-year age groups shown in Table 7.

age groups	Claimant status	sizes (n)
$[20, 30[$	D	426
	E	568
$[30, 40[$	D	967
	E	1116
$[40, 50[$	D	1079
	E	1177
$[50, 60[$	D	1039
	E	1136
$[60, 70[$	D	595
	E	786
$[70, 80[$	D	102
	E	175

Table 7: Age groups of females in SOA91

Applying our algorithm procedure at 5% level, we obtained four clusters and the dendogram is presented in Figure 5. It appears that the dependence structure of

claim charges change over age where it shows that the status of the policy holder is irrelevant and that premiums charged to both types of individuals should be the same if the size of the observations are substantially identical.

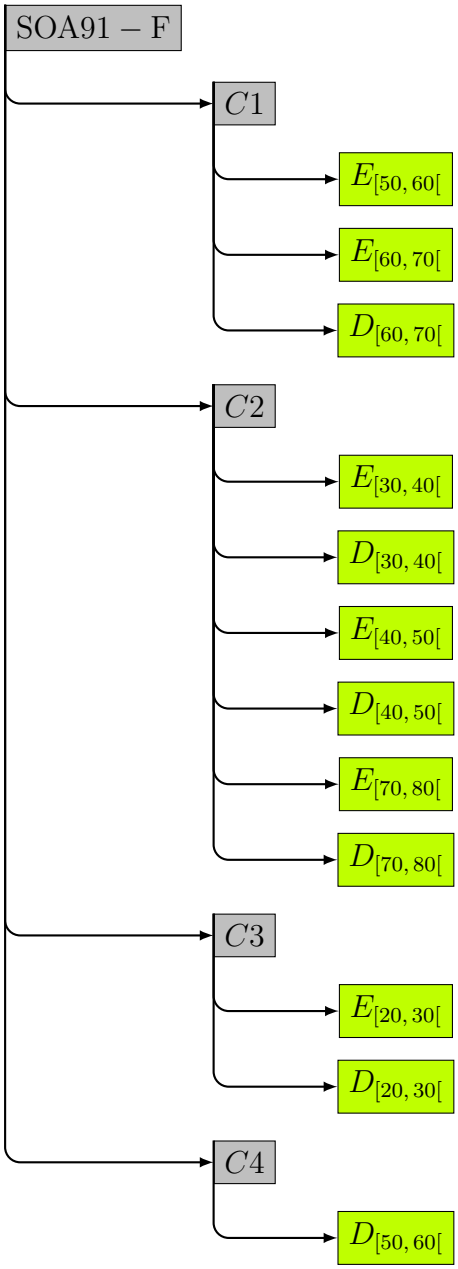

Figure 5: Dendogram of SOA91-Female at 5% level. $C_1 \cdots C_4$ denote the cluster.

6 CONCLUSION

REFERENCES

Jin Seo Cho and Halbert White. Generalized runs tests for the iid hypothesis. *Journal of econometrics*, 162(2):326–344, 2011.

Giovanni De Luca and Paola Zuccolotto. Dynamic tail dependence clustering of financial time series. *Statistical papers*, 58(3):641–657, 2017.

F. Marta L. Di Lascio and Simone Giannerini. A copula-based algorithm for discovering patterns of dependent observations. *Journal of Classification*, 29:50–75, 2012. ISSN 0176-4268. URL http://dx.doi.org/10.1007/s00357-012-9099-y. 10.1007/s00357-012-9099-y.

F. Marta L. Di Lascio and Simone Giannerini. *CoClust: copula based cluster analysis*, 2017. URL https://CRAN.R-project.org/package=CoClust. R package version 0.3-2.

F Marta L Di Lascio and Simone Giannerini. Clustering dependent observations with copula functions. *Statistical Papers*, 60(1):35–51, 2019.

Francesca Marta Lilja Di Lascio. Coclust: An r package for copula-based cluster analysis. In *Recent Applications in Data Clustering*, pages 93–114. IntechOpen, 2018.

Marta Disegna, Pierpaolo D'Urso, and Fabrizio Durante. Copula-based fuzzy clustering of spatial time series. *Spatial Statistics*, 21:209–225, 2017.

Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric statistical inference*. CRC press, 2014.

K.L. Grazier and B. G'Sell. *Group medical insurance large claims database collection and analysis*. Society of Actuaries, 1997.

Harry Joe and Peijun Sang. Multivariate models for dependent clusters of variables with conditional independence given aggregation variables. *Computational Statistics & Data Analysis*, 97:114–132, 2016.

Ioannis Kosmidis and Dimitris Karlis. Model-based clustering using copulas with applications. *Statistics and computing*, 26(5):1079–1099, 2016.

Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

Y. I. Ngounou-Bakam and D. Pommeret. Nonparametric estimation of copulas and copula densities by orthogonal projections. *arXiv:2010.15351*, 2021.

Y. I. Ngounou-Bakam and D. Pommeret. K-sample test for equality of copulas. *arXiv*, 2022.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6: 461–464, 1978.

Chien-Lin Su, Johanna G Nešlehová, and Weijing Wang. Modelling hierarchical clustered censored data with the hierarchical kendall copula. *Canadian Journal of Statistics*, 47(2):182–203, 2019.

Ying Wang. Nonparametric tests for randomness. *ECE*, 461:1–11, 2003.

Lili Zhang and Jangsun Baek. Mixtures of gaussian copula factor analyzers for clustering high dimensional data. *Journal of the Korean Statistical Society*, 48(3): 480–492, 2019.

Denys POMMERET (correponding author)
Aix-Marseille University
CNRS, Centrale Marseille, I2M
Campus de Luminy
13288 Marseille cedex 9
Marseille,France
E-Mail: denys.pommeret@univ-amu.fr

Yves Ismaël NGOUNOU BAKAM
Aix-Marseille University
CNRS, Centrale Marseille, I2M
Campus de Luminy
Marseille,France
E-Mail: yves-ismael.ngounou-bakam@univ-amu.fr