FEATURE GROUPING BASED ON THE OPTIMAL SEQUENCE OF CORRELATION MATRIX MINORS

DVOENKO SERGEI D.[1] , PSHENICHNY DENIS O.[1]

[1]TULA STATE UNIVERSITY, TULA, RUSSIAN FEDERATION

SPEAKER: DVOENKO S.D.

BARCELONA 10/2016

20 slides

1. DATA ANALYSIS PROBLEM

- Data analysis problem usually arises in early stages of investigations, when a model of a phenomenon in researching has not been developed yet. Hence, it is too early to introduce a problem of a model identification. It needs to collect and study a lot of miscellaneous information about most significant characteristics of a phenomenon under investigation in this case.
- Such a situation forces us to use inconsistent approach, since we do not know what characteristics are important, and what knowledge needs to be collected.
- Therefore, data analysis methods must resolve the contradiction and focus on the correct description of the phenomenon.
- Specifically, the problem of informal interpretation of factors and groups arises in the grouping problem. Factors are synthetic features and difficulties can arise in informal interpretation of them. Therefore, after groups and corresponding factors have been built the representative usually is defined for each group as a feature, the most correlated with the group factor. As a result, it is possible to denote groups informally as such initial features.

2. FEATURE GROUPING

- Let $X(N,n)$ be a data matrix with N measurements of n features. With lines $\mathbf{x}_i = (x_{i1}, \dots x_{in})$ and columns $X_j = (x_{1j}, \dots x_{Nj})^T$ it can be represented as a table of lines-objects $X(N,n) = (\mathbf{x}_1, \dots \mathbf{x}_N)^T$ and columns-features $X(N,n) = (X_1, \dots X_n)$.

- It is supposed all objects are concentrated in K clusters, and all features are depended on L hidden factors $F_i, i = 1, \dots L$.

- It is supposed factors are statistically independent each from other. Factors form a system of orthogonal axes. According to hidden factors, all features are divided in groups $G_i, i = 1, \dots L$.

- It is the problem to get orthogonal common factors with so-called "simple structure" by orthogonal rotation. On the contrary, the oblique rotations allow to get correlated factors.

- If features are divided in groups first, then each group factor is defined as the most correlated with features in group. As a result, the system of generally correlated factors is naturally defined. This result can be improved to get less correlated factors by re-grouping features again relative to factors, etc.

2. FEATURE GROUPING

- The special algorithms of Extreme Grouping were developed previously for principal (Square) and centroid (Module) factors. First of them is known as LPCA (local principal component analysis) or LPF (local principal factors). Second can be denote as LCF (local centroid factors).

- Unfortunately, factors are synthetic essences or artificial features. It is the well-known problem to get interpretation of them. Usually, the feature, most correlated with group factor, can be well interpreted as the group representative. Group representatives are usually used instead of factors itself. As a result, we can get a system of less correlated initial features as a system of well interpreted features relevant to system of hidden factors.

- According to this approach, based on factor model we produce some intermediate transformations to reduce the initial set of features to the set of less correlated representatives.

- Can we group features or reduce dimensionality based on some other idea without calculating principal or centroid factors themselves as intermediate steps?

3. METRIC CONFIGURATION AND VIOLATIONS

- Let data be directly presented by paired comparisons between elements (objects or features) of the limited set in the form of a square matrix of similarities or dissimilarities.
- This is the usual situation in modern intelligent data processing (data mining, expert's evaluations, decision problem, qualitative data, etc.).
- Usually, we would like similarities to be scalar products or correlations in the positive quadrant of a metric space, and dissimilarities to be distances. In this case set elements can be immersed in a metric space as a correct configuration. But usually this is not so. We denote it as "metric violations".
- It seems this is not a problem for a data matrix $X(N,n)$, given by correct measurements. Distances $D(N,N)$, weighed scalar products $R(n,n)$, similarities $S(n,n)=R^2(n,n)$ usually can be calculated correctly, if no errors in calculations.

3. METRIC CONFIGURATION AND VIOLATIONS

- Let the matrix of weighed (normalized) scalar products of features $R(n,n)$ be given (correlations). In a case of a correct feature configuration, $R(n,n)$ appears to be the positively definite one with a sequence of eigenvalues $\lambda_1 > ... > \lambda_n > 0$.
- In a case of metric violations $R(n,n)$ appears to be the non-positively definite one with $q < n$ last negative eigenvalues in the sequence of them $\lambda_1 > ... > \lambda_{m=n-q} > 0 > \lambda_{n-q+1} > ... > \lambda_n$. build
- Usually it is not the problem in PCA based on the Karhunen-Loeve transform to built a projection of $X(N,n)$ in the space of first $m = n - q$ eigenvectors with corresponded positive eigenvalues $\lambda_1 > ... > \lambda_m > 0$ to get so called calculated (and correct) data $Y(N,m)$.
- In this case we get the correlation matrix $R_{q-}(n,n)$ of residuals. But it is non-correct one with $r_{ii} > 1, i = 1,...n$ and $\sum_{i=1}^{n} r_{ii} > n$. As a result, we have got additional to n dispersion in data “from nowhere”. Of course, we can correct it immediately by normalizing, but later (or without data matrix X) we can’t control this situation (we don’t know, what normalizations have been used before).

3. METRIC CONFIGURATION AND VIOLATIONS

- In other case metric violations can arise in transformations according to factor models. Here we need to define communalities for the correlation matrix to reduce it according to principal or centroid factor models. The correlation matrix $\overline{R}(n,n)$ of residuals appears to be non-normalized with $r_{ii} < 1, i = 1,\dots n$ to explain dispersion $\sum_{i=1}^{n} r_{ii} < n$ of common factors.

- Communalities are usually defined based on some empirical recommendations, since it is the complicated theoretical problem. But in the most of real situation we take the risk to get the non-positively definite residual matrix $\overline{R}(n,n)$, that is the metric violation.

4. OPTIMAL SEQUENCE OF FEATURES

- We have developed before another idea of corrections to get positively definite correlation matrix and not to change data dispersion. We don't eliminate "layers" of eigenvectors corresponded to negative eigenvalues, but correct individually only some correlations in the matrix.
- According to Silvester's criterion, the symmetric matrix $S(n,n)$ of quadratic form is positively definite, if all its principal (upper left) minors are positive $S_k = S(k,k), k = 1,...n$, $\det S_k > 0$, where $S_1 = S(1,1) = s_{11} = 1$ for normalized $S(n,n)$ with the diagonal of units.
- According to Silvester's law of inertia, the number q of negative eigenvalues is equal to sign changes of principal minors decreasing from $\det S_1 = 1$ in the sequence $S_0 = 1, S_1, S_2, ... S_n = S(n,n)$.

4. OPTIMAL SEQUENCE OF FEATURES

- There are arbitrary places of sign changes in such the sequence of minors. Let us permute elements in the set to concentrate sign changes at the end of the sequence of minors $S_k, k=1,\dots n$. Therefore, in ideal case the principal minor S_{n-q+1} appears to be the first time negative one with $\det S_{n-q+1} < 0$, and signs of other $q-1$ minors alternate. Hence, not more than q elements violate metrics. We denoted such idea as a localization of negative eigenvalues in the non-positively definite similarity matrix.
- Let u be the number of additional steps without the sign change for all q negative eigenvalues. Hence, we correct not more than $q+u$ last elements (corresponded pair correlations) in the optimal sequence of set elements. As a result, violating elements, collected at the end of the sequence, decrease additional violations from other elements and total deviation of corrected matrix from the initial similarity matrix.
- If no violations, we get the suboptimal sequence of principal minors $S_k, k=1,\dots n$ with the most slowly decreasing positive values of determinants $\det S_k > 0$.

4. OPTIMAL SEQUENCE OF FEATURES

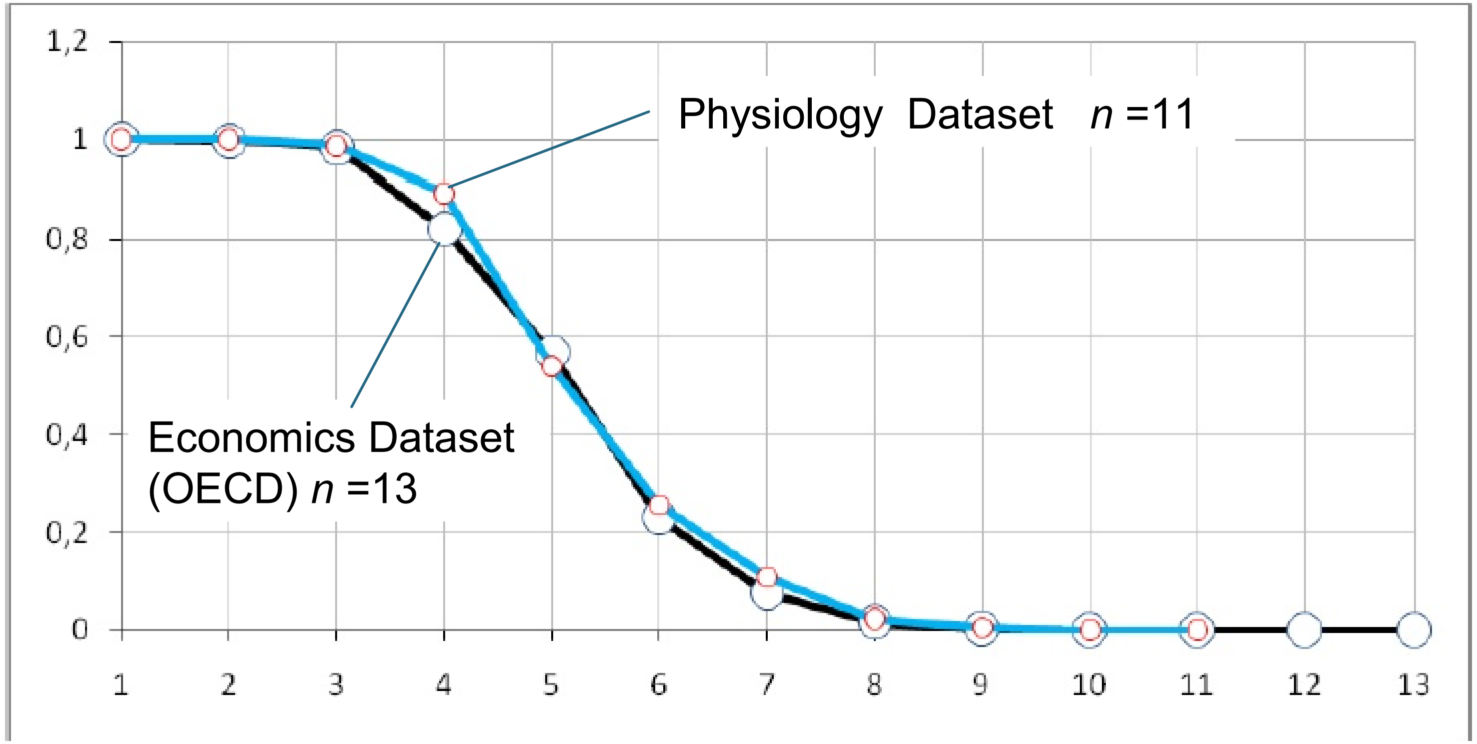
- The $\det R$ of the correlation matrix $R(n,n)$ depends on the mutual orthogonality of the feature set. The more orthogonality – the more closed to 1 determinant, and otherwise. It is clear for $n=2$, since $\det R = 1 - r^2$. It can be showed for $n=3$, since $\det R = 1 + 2r_{12}r_{13}r_{23} - r_{12}^2 - r_{13}^2 - r_{23}^2$ for different cases without metric violations.
- Hence, the correct correlation matrix $R(n,n)$ defines the suboptimal sequence of principal minors and, as a result, defines the suboptimal sequence of nested sets of most orthogonal features. In other words, it is the sequence of the most orthogonal features at the beginning, and the least orthogonal ones at the end of the sequence.
- Therefore, we can define first m optimal representatives of m groups without calculating principal (for squared correlations) or centroid (for modules of correlations) factors.
- The optimal sequence can be used for initial partition in algorithms of Extreme Grouping (Square and Module), or independently as a result of grouping.

5. EXPERIMENTS

- Let L be a number of groups $G_i = 1,\dots L$ with $|G_i| = n_i, \sum_{i=1}^{L} n_i = n$, and $r(X_j, F_i)$ be correlation of a factor $F_i = (f_{1i},\dots f_{Ni})^T$ with a feature $X_j = (x_{1j},\dots x_{Nj})^T$.
- Let us use criteria $I_Q = \sum_{i=1}^{L} \sum_{j \in G_i} r^2(X_j, F_i)$ and $I_M = \sum_{i=1}^{L} \sum_{j \in G_i} | r(X_j, F_i) |$ for Square and Module to evaluate the quality of the grouping.
- Different types of initial partitions are investigated for Square and Module algorithms based on criteria I_Q and I_M :
  - First L features according to Optimal Sequence
  - L features with minimal mutual correlations
  - First L features (as is) – bad quality
  - Random L features – bad quality

5. EXPERIMENTS

- It is used a heuristic idea (like the well-known one in cluster-analysis) to evaluate the number L: for Optimal Sequence of positive principal minors $S_k, k=1,\dots n$ (after correction, if it was necessary) the optimal L corresponds to interval of sharp decreasing of determinants $\det S_k$.



The optimal number of groups is:
L=3÷5 for Economics Dataset
L=4÷5 for Physiology Dataset

6. ECONOMICS DATASET

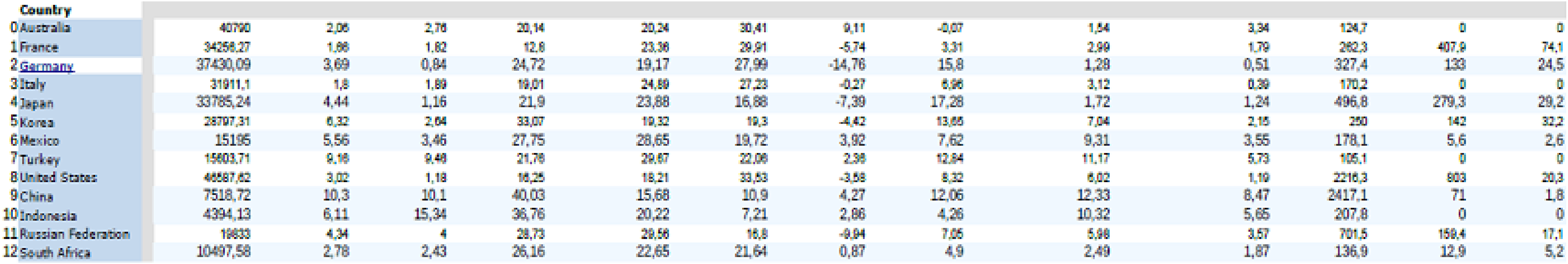
- The Dataset of Organization for Economic Cooperation and Development (OECD) is extracted from the Fastbook Country Statistical Profiles – 2013 Edition for 13 economic characteristics of 13 countries: Australia, France, Germany, Italy, Japan, Korea, Mexico, Turkey, US, China, Indonesia, RF, and South Africa (http://stats.oecd.org/).

- Economics activities are showed in following subjects as features:

1. Gross Domestic Product per capita ($)
2. Real GDP growth (%)
3. Value added in agriculture, hunting, forestry, fishing (%)
4. Value added in industry, including energy (%)
5. Value added in wholesale and retail trade, repairs, hotels and restaurants, transport (%)
6. Value added in financial intermediation, real estate, renting and business activities (%)
7. Real value added in agriculture, hunting, forestry, fishing (%)
8. Real value added in industry, including energy (%)
9. Real value in wholesale and distributive trade, repairs, transport, food service, communication (%)
10.Real value in financial and insurance activities, real estate, professional support activities (%)
11.Total primary energy supply (TW/h)
12.Nuclear electricity generation (TW/h)
13.Nuclear electricity generation as percentage of total electricity (%)

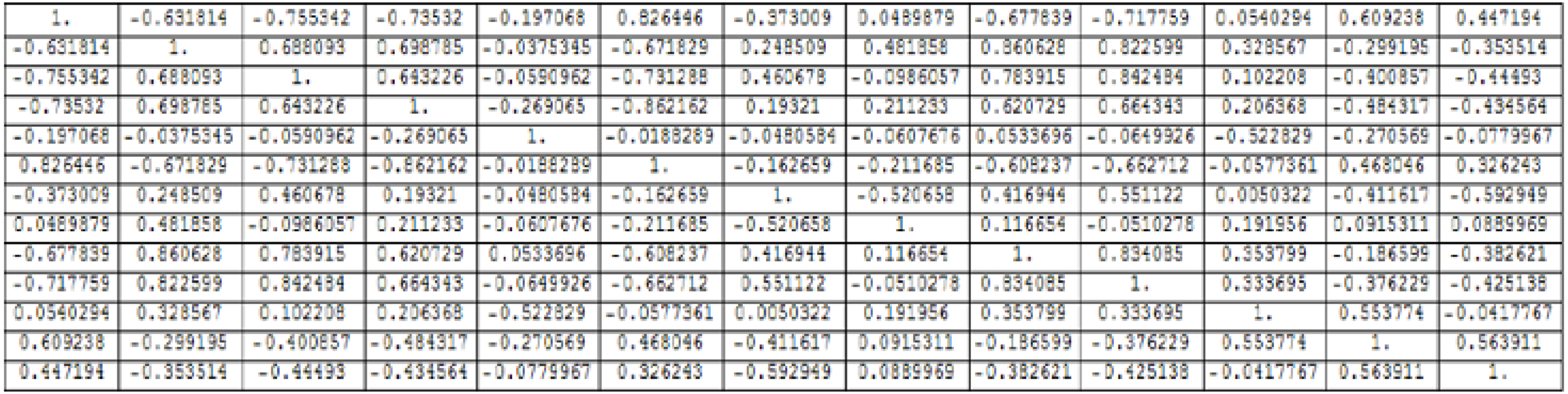
6. ECONOMICS DATASET

- Correlations of 13 features are presented in the matrix



	Country													
0	Australia	40790	2,06	2,76	20,14	20,24	30,41	9,11	-0,07	1,54	3,34	124,7	0	0
1	France	34258,27	1,86	1,62	12,8	23,38	29,91	-5,74	3,31	2,99	1,79	262,3	407,9	74,1
2	Germany	37430,09	3,69	0,84	24,72	19,17	27,99	-14,76	15,8	1,28	0,51	327,4	133	24,5
3	Italy	31911,1	1,8	1,89	19,01	24,89	27,23	-0,27	6,96	3,12	0,39	170,2	0	0
4	Japan	33785,24	4,44	1,16	21,9	23,88	16,88	-7,39	17,28	1,72	1,24	496,8	279,3	29,2
5	Korea	28797,31	6,32	2,04	33,07	19,32	19,3	-4,42	13,65	7,04	2,15	250	142	32,2
6	Mexico	15195	5,56	3,46	27,75	28,65	19,72	3,92	7,62	9,31	3,55	178,1	5,6	2,6
7	Turkey	15603,71	9,16	9,48	21,76	29,67	22,08	2,38	12,64	11,17	5,73	105,1	0	0
8	United States	46587,62	3,02	1,18	16,25	18,21	33,53	-3,58	8,32	6,02	1,19	2216,3	803	20,3
9	China	7518,72	10,3	10,1	40,03	15,68	10,9	4,27	12,06	12,33	8,47	2417,1	71	1,8
10	Indonesia	4394,13	6,11	15,34	36,76	20,22	7,21	2,86	4,26	10,32	5,65	207,8	0	0
11	Russian Federation	19833	4,34	4	28,73	28,56	16,8	-9,94	7,05	5,98	3,57	701,5	159,4	17,1
12	South Africa	10497,58	2,78	2,43	26,16	22,65	21,64	0,87	4,9	2,49	1,87	136,9	12,9	5,2

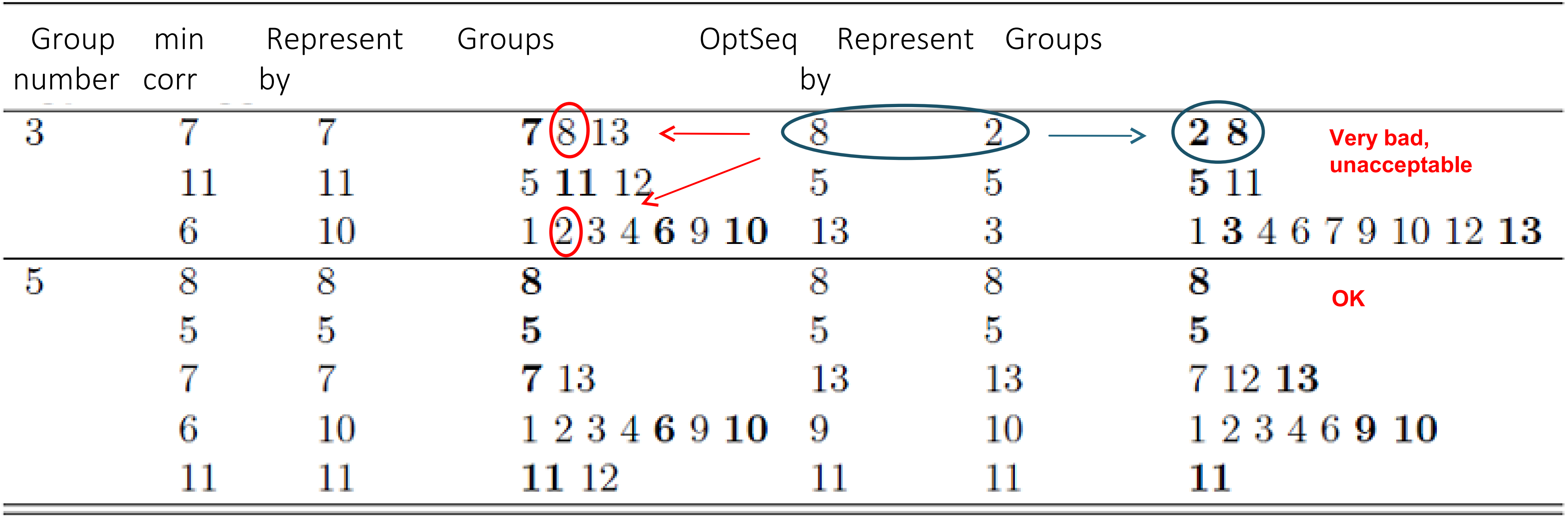
Out[33]=



1.	-0.631814	-0.755342	-0.73532	-0.197068	0.826446	-0.373009	0.0489879	-0.677839	-0.717759	0.0540294	0.609238	0.447194
-0.631814	1.	0.688093	0.698785	-0.0375345	-0.671829	0.248509	0.481858	0.860628	0.822599	0.328567	-0.299195	-0.353514
-0.755342	0.688093	1.	0.643226	-0.0590962	-0.731288	0.460678	-0.0986057	0.783915	0.842484	0.102208	-0.400857	-0.44493
-0.73532	0.698785	0.643226	1.	-0.269065	-0.862162	0.19321	0.211233	0.620729	0.664343	0.206368	-0.484317	-0.434564
-0.197068	-0.0375345	-0.0590962	-0.269065	1.	-0.0188289	-0.0480584	-0.0607676	0.0533696	-0.0649926	-0.522829	-0.270569	-0.0779967
0.826446	-0.671829	-0.731288	-0.862162	-0.0188289	1.	-0.162659	-0.211685	-0.608237	-0.662712	-0.0577361	0.468046	0.326243
-0.373009	0.248509	0.460678	0.19321	-0.0480584	-0.162659	1.	-0.520658	0.416944	0.551122	0.0050322	-0.411617	-0.592949
0.0489879	0.481858	-0.0986057	0.211233	-0.0607676	-0.211685	-0.520658	1.	0.116654	-0.0510278	0.191956	0.0915311	0.0889969
-0.677839	0.860628	0.783915	0.620729	0.0533696	-0.608237	0.416944	0.116654	1.	0.834085	0.353799	-0.186599	-0.382621
-0.717759	0.822599	0.842484	0.664343	-0.0649926	-0.662712	0.551122	-0.0510278	0.834085	1.	0.333695	-0.376229	-0.425138
0.0540294	0.328567	0.102208	0.206368	-0.522829	-0.0577361	0.0050322	0.191956	0.353799	0.333695	1.	0.553774	-0.0417767
0.609238	-0.299195	-0.400857	-0.484317	-0.270569	0.468046	-0.411617	0.0915311	-0.186599	-0.376229	0.553774	1.	0.563911
0.447194	-0.353514	-0.44493	-0.434564	-0.0779967	0.326243	-0.592949	0.0889969	-0.382621	-0.425138	-0.0417767	0.563911	1.

6. ECONOMICS DATASET

- Optimal Sequence for correlations squared is $[8,5,13,9,11,4,7,12,3,1,10,6,2]$.
- Quality of groups based on I_Q



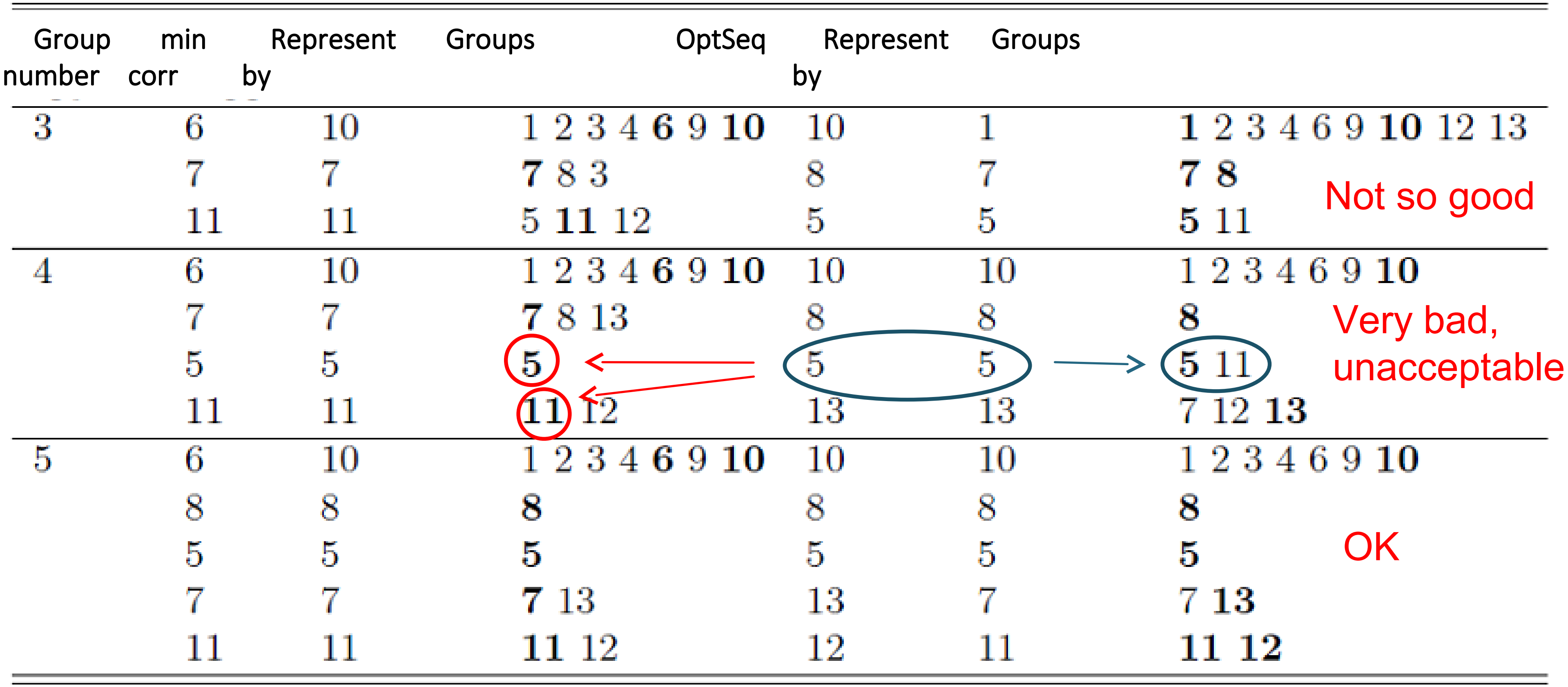
Group number	min corr	Represent by	Groups	OptSeq	Represent by	Groups
3	7	7	**7** 8 13	8	2	**2** **8**
	11	11	5 **11** 12	5	5	**5** 11
	6	10	1 2 3 4 **6** 9 **10**	13	3	1 **3** 4 6 7 9 10 12 **13**
5	8	8	**8**	8	8	**8**
	5	5	**5**	5	5	**5**
	7	7	**7** 13	13	13	7 12 **13**
	6	10	1 2 3 4 **6** 9 **10**	9	10	1 2 3 4 6 **9** **10**
	11	11	**11** 12	11	11	**11**

Very bad, unacceptable

OK

6. ECONOMICS DATASET

- Optimal Sequence for modules of correlations is $[10, 8, 5, 13, 12, 4, 11, 9, 2, 7, 3, 6, 1]$.
- Quality of groups based on I_M



Group number	min corr	Represent by	Groups	OptSeq	Represent by	Groups
3	6	10	1 2 3 4 **6** 9 **10**	10	1	**1** 2 3 4 6 9 **10** 12 13
	7	7	**7** 8 3	8	7	**7** **8**
	11	11	5 **11** 12	5	5	**5** 11
4	6	10	1 2 3 4 **6** 9 **10**	10	10	1 2 3 4 6 9 **10**
	7	7	**7** 8 13	8	8	**8**
	5	5	**5**	5	5	**5** 11
	11	11	**11** 12	13	13	7 12 **13**
5	6	10	1 2 3 4 **6** 9 **10**	10	10	1 2 3 4 6 9 **10**
	8	8	**8**	8	8	**8**
	5	5	**5**	5	5	**5**
	7	7	**7** 13	13	7	7 **13**
	11	11	**11** 12	12	11	**11** **12**

Not so good

Very bad, unacceptable

OK

6. ECONOMICS DATASET

- **Results for both criteria are the same for *L*=5**

- **Group interpretation is:**

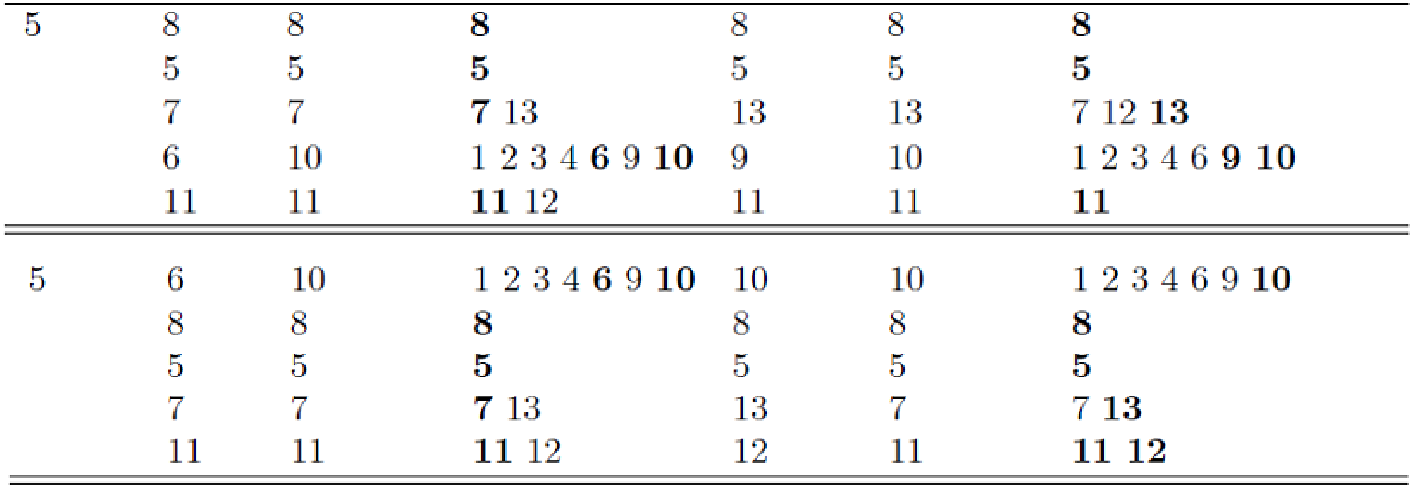
1. Real value added in industry, including energy (8)

2. Value added in trade service and transport (5)

3. Value added in natural production with power inputs (7 13)

4. GDP and value added in all activities (1 2 3 4 6 9 10)

5. Total primary energy supply (11)

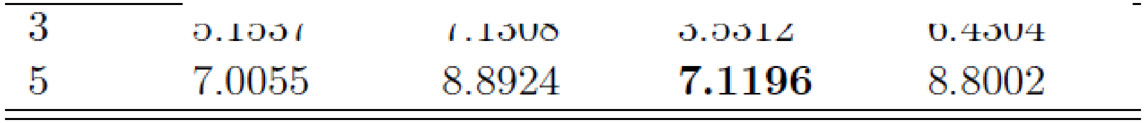


Group number	min corr	Represent by	Groups	OptSeq	Represent by	Groups
5	8	8	**8**	8	8	**8**
	5	5	**5**	5	5	**5**
	7	7	**7** 13	13	13	7 12 **13**
	6	10	1 2 3 4 **6** 9 **10**	9	10	1 2 3 4 6 **9** **10**
	11	11	**11** 12	11	11	**11**
5	6	10	1 2 3 4 **6** 9 **10**	10	10	1 2 3 4 6 9 **10**
	8	8	**8**	8	8	**8**
	5	5	**5**	5	5	**5**
	7	7	**7** 13	13	7	7 **13**
	11	11	**11** 12	12	11	**11** **12**

6. ECONOMICS DATASET

Quality I_Q

by

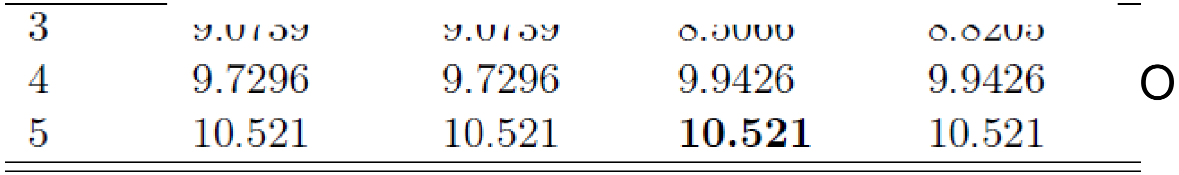


Group number	MinCorr		OptSeq	
	Initial	Final	Initial	Final
3	5.1557	7.1508	5.5512	6.4504
5	7.0055	8.8924	**7.1196**	8.8002

OptSeq is improved by Square

Quality I_M

by



Group number	MinCorr		OptSeq	
	Initial	Final	Initial	Final
3	9.0759	9.0759	8.5000	8.8205
4	9.7296	9.7296	9.9426	9.9426
5	10.521	10.521	**10.521**	10.521

OptSeq and Module are the same

CONCLUSION

- The correct correlation matrix defines the suboptimal sequence of principal minors and the suboptimal sequence of nested sets of most orthogonal features.
- It is the sequence of the most orthogonal features at the beginning, and the least orthogonal ones at the end of the sequence.
- We can define first m optimal representatives of m groups without calculating principal (for squared correlations) or centroid (for modules of correlations) factors.
- The optimal sequence can be used for initial partition in algorithms of Extreme Grouping (Square and Module), or independently as a result of grouping.

THANK YOU FOR ATTENTION!