Clustering time series – application to inventory investments

Attila Chikán, Erzsébet Kovács, Zsolt Matyusz, Magdolna Sass, Péter Vakhal

Corvinus University of Budapest

Problem

- Multiple (multivariate) time series
- Time domain adds a further dimension to data
  - Variable
  - Item (case)
  - + Time
- Compared to cross-sectional datasets multiple time series are usually not iid.
- Time series are almost always autocorrelated, many times cross-correlated
- Traditional clustering algorithms (k-means, NN, etc) can not be used

Why traditional methods fail?

- Traditional clustering methods are based on metric distances
- Segmentation is based on proximity matrix
- In case of cross-sectional data - distances make sense
- In case of multiple time series - distances do NOT make sense

- For example:
  - Distance between two variables at time T and time T+13…
  - Distances at the SAME time domain make sense

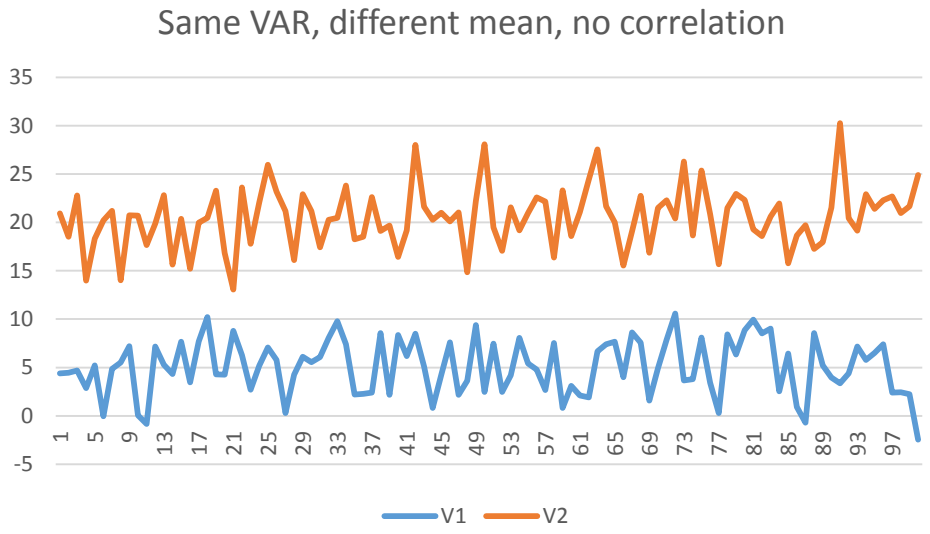
Possible alternatives

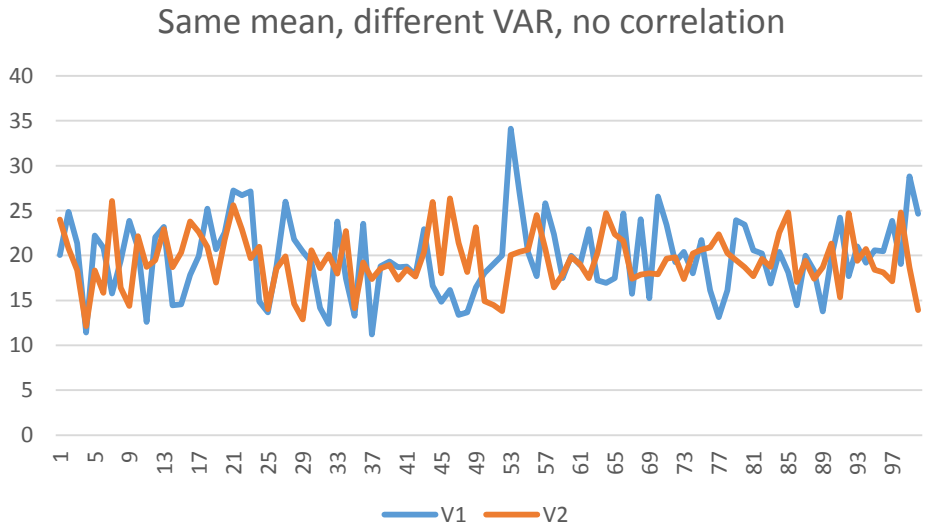
- Correlation (covariance)
  - Pro: models mutual dependence well, its measure can be well interpreted
  - Contra: in case of non-stationary time series (that is, E(x)=VAR(x)=infinite) can not be applied, only bivariate, scale-invariant (that is, amplitude is neglected)
- Correlation of linear trends
  - Pro: easy to calculate and interpret, amplitude is taken into account
  - Contra: sensitive to shocks, non-stationary time series can be problematic
- PCA – based on the assumption that similar time series will be assigned to the same component
  - Pro: based on multiple correlation – thus based on quasi similarities
  - Contra: information loss, non-stationary time series are problematic, goal of PCA is dimension reduction by creating non-correlating components – variables in the same component are similar but this does not mean that variables in different components are not similar

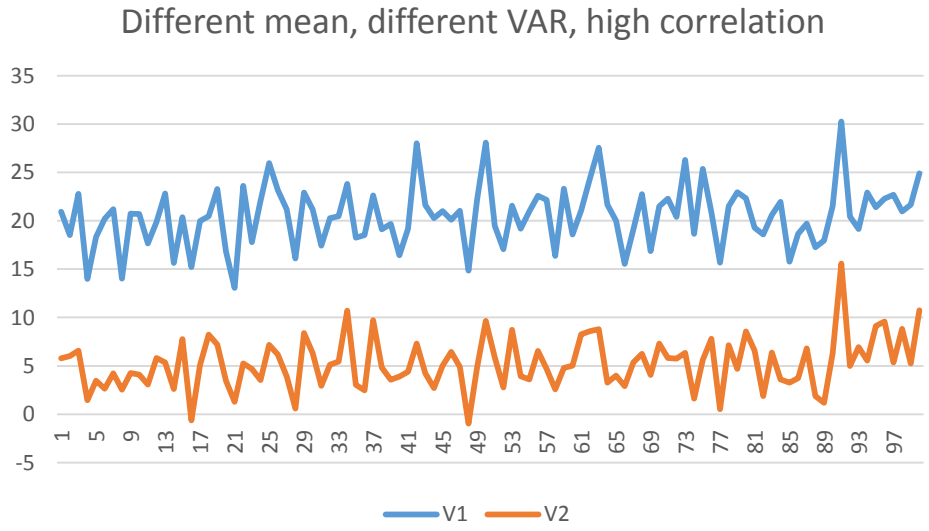
The problem of amplitudes and non-stationarity

- Non-stationarity has an affect on clustering (not elaborated here)
- ADF or other econometric tests
- By differencing stationarity can be usually achieved
- BUT! By differenciation the amplitude is lost
  - Why? After differentiation the mean is zero.
- Dilemma: two time series with the same variation but different mean are in the same cluster? Sometimes not…

Clustering definition in case time series







If we were deciding upon distances, case 3 will be a cluster, while case 1 and case 2 most probably not.

Proposed framework for clustering time series

- The clustering criteria should be based on correlation and distance at the same time.
- Instead of bivariate correlations multiple correlation should be applied.
  - Multiple correlation: $R^2 = c^T R_{xx}^{-1} c$, where c is the pairwise correlation vector between var x_1 and vars x_n, R_{xx}^{-1} is the inverse of variance-covariance matrix of variables.
- Time series must be detrended without loosing their shift.
- Time series should be stationary.

Proposed methodology I.

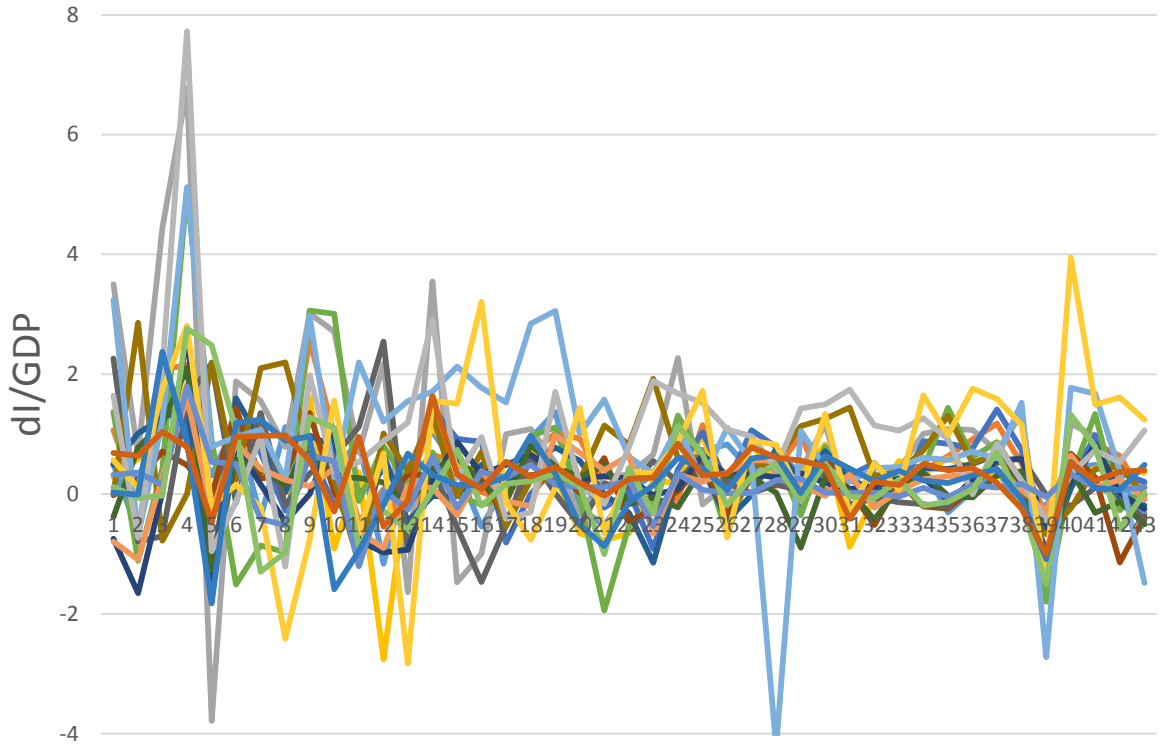
- Principal Component Analysis (PCA)
  - Reminder: PCA is a dimension reduction method based on the multiple correlation of the variables. It applies orthogonal transformation of correlated variables.
- PCA is done by the eigenvalue decomposition of the covariance matrix.
- The first component explains the largest proportion of the total variance.
- Component scores of the first component can be served as cluster centre.

Proposed methodology II.

- Component loadings represent the pairwise correlation of the variables with the component.
- Therefore component loadings can be used as distance measure from the cluster centre.
- Moreover as correlation is measured in metric space it is also possible to compare within group distances.
- Loadings therefore can be drawn like a dendrogram (agglomerative clustering).
- In case of one component hierarchical clustering can be done.
- Extraction of more components are not suggested as orthogonal transformation and rotations artificially segment the database.

Dataset

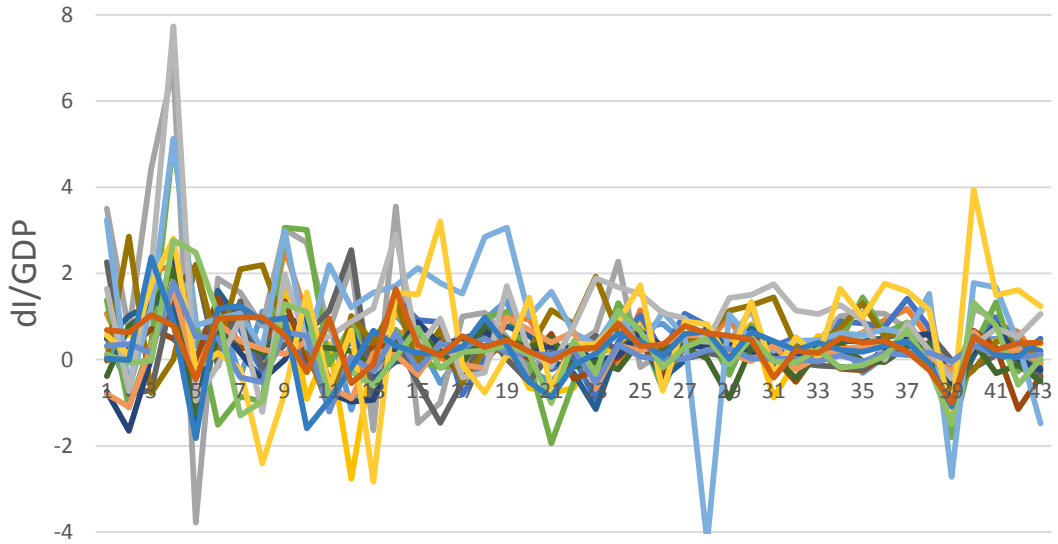
- International inventory investment database of market 20 market economies between 1970 and 2013.

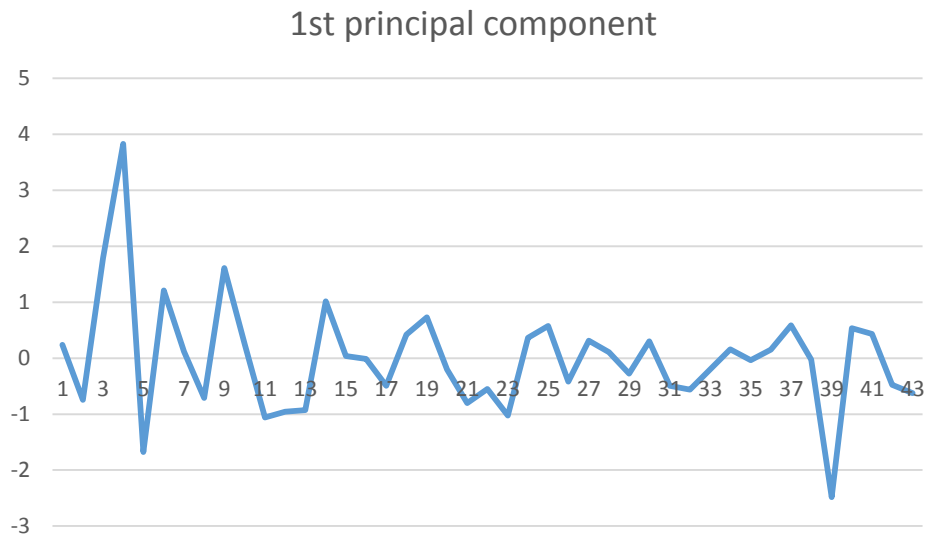


Steps

1. Stationarity tests (ADF)
2. Detrend dataset
   - $y = a + y_{t-1} + \epsilon$ done by OLS
   - $\ddot{y} = \epsilon + a$, where $\ddot{y}$ is the detrended variable with shift
3. Create one component by PCA.
4. Visualize component scores in a dendrogram.

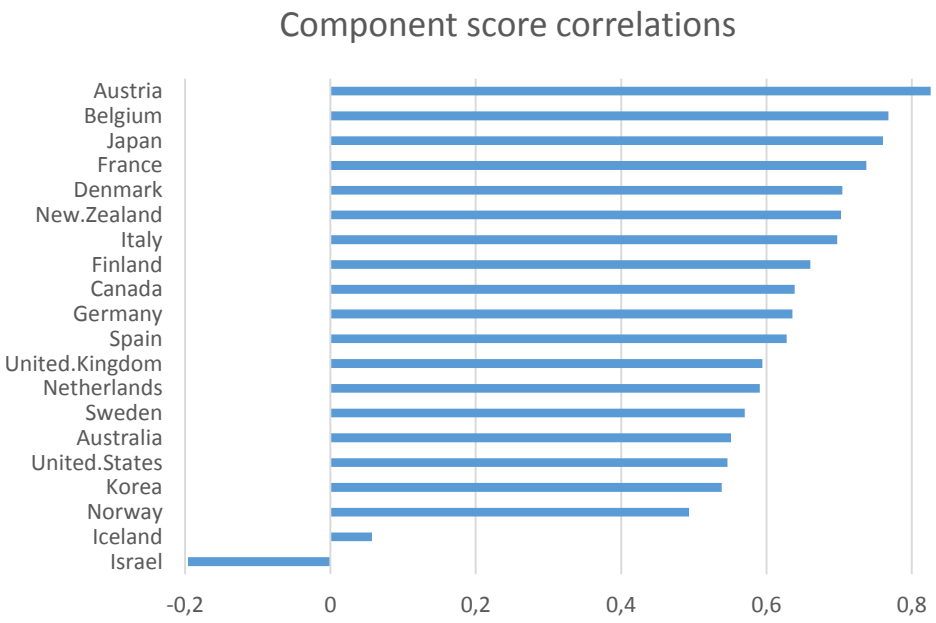
Principal component of the dataset

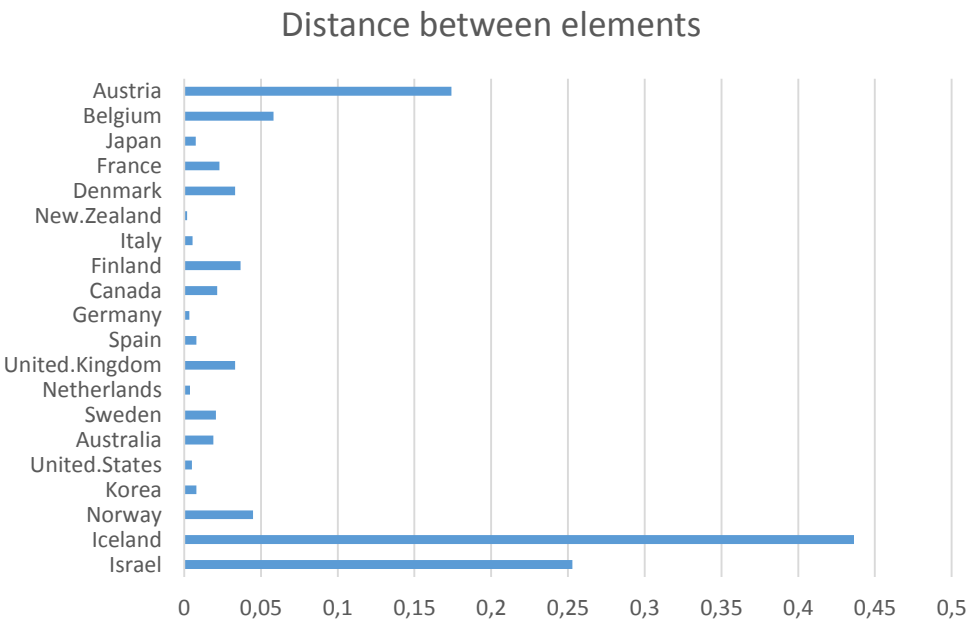




KMO test: 0,741 ($p<0.001$)
Total variance explained: 38%

Results



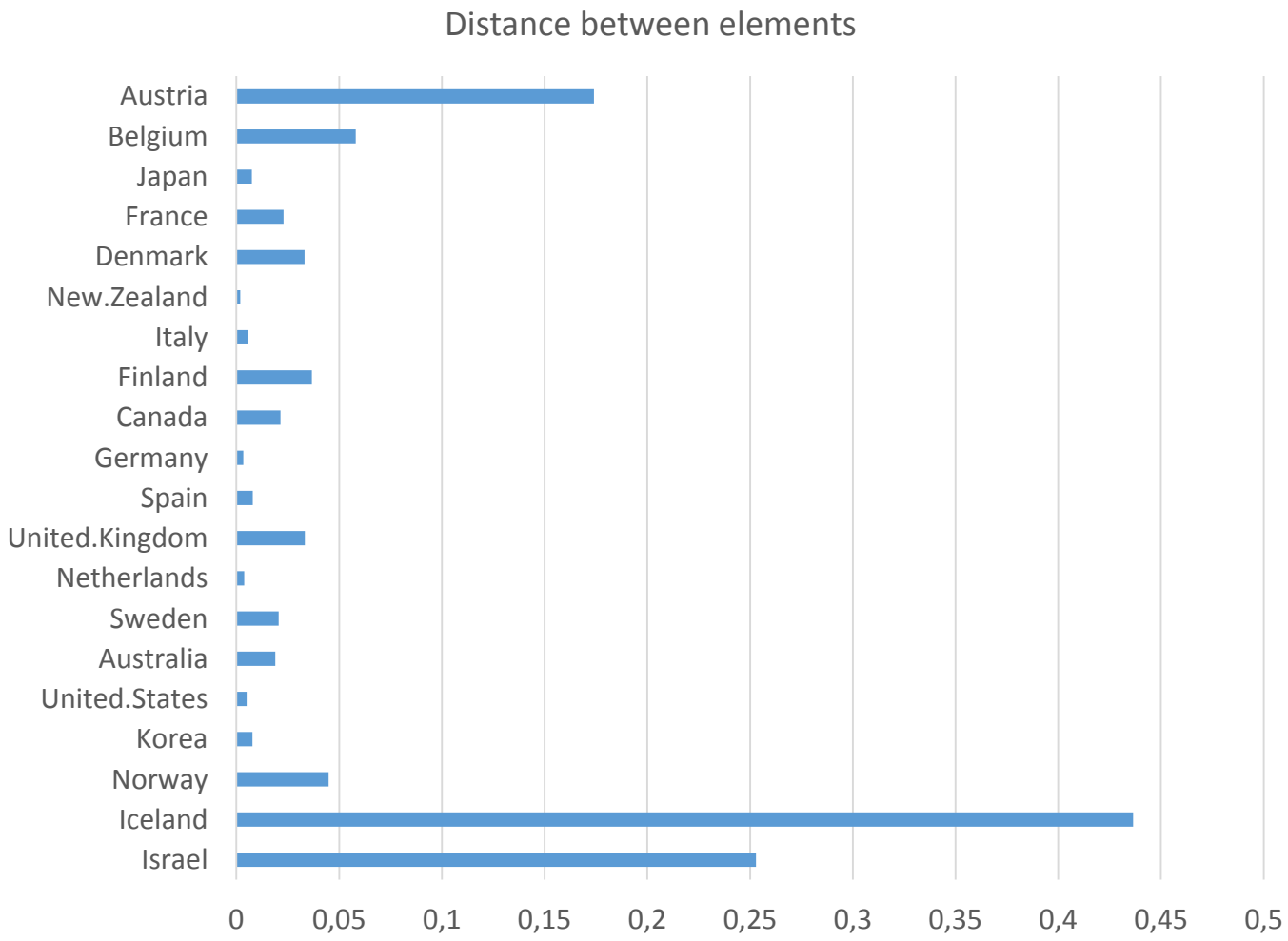


Component score (PCA) correlation is r_n=Cor(PCA, X_n)

Distance is $d_{1,2}=r_2-r_1$

Distance is the difference between two correlations, that is Cor(PCA, X_1)-Cor(PCA, X_2). The larger the difference the less stronger is the correlation X_2 and PCA compared to X_1 and PCA.

Clustering



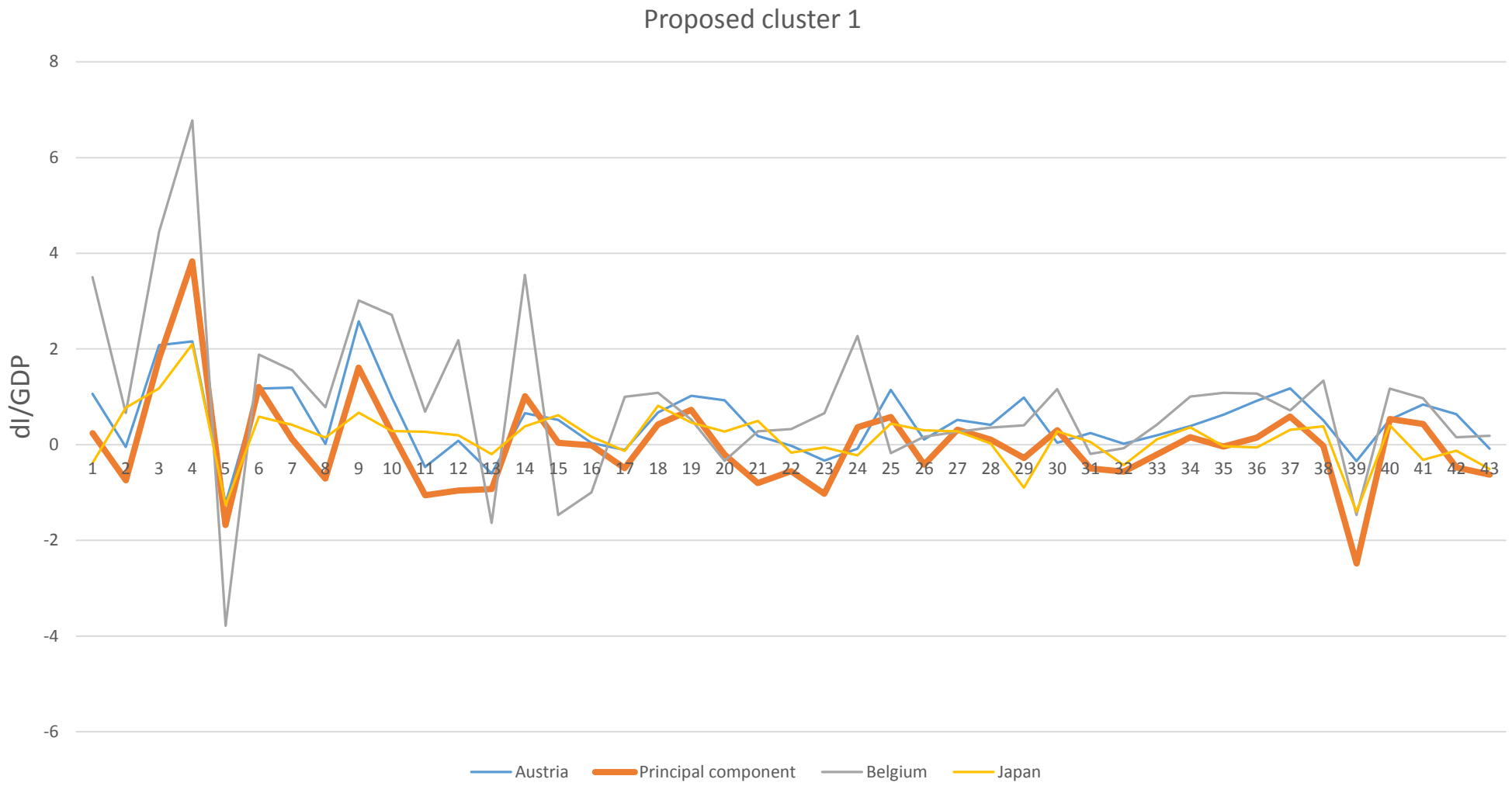
Clustering can be done visionally (for the sake of simplicity) or by predifined algorithm based on threshold values.

Regardless the clustering method larger distances should sign the existance of a new cluster.

In our case Austria, Belgium and Japan constitutes the first cluster (visionally).

Second cluster can be: FR, DK, NZ, IT.

Results


Proposed cluster 1
dI/GDP
8
6
4
2
0
-2
-4
-6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
Austria
Principal component
Belgium
Japan

Conclusion

- Traditional cross-sectional methods can not be applied for time series clustering
- Time series should be stationary but non-stationary series must not be differentiated but detrended.
- Clustering criteria should be based on correlation and distance at the same time
- Proposed clustering method is based on a one component PCA where the component is the cluster centre
- Clustering should be done visionally or by threshold values as in case of agglomerative methods.

Thank you for your attention!

Questions?