**Noname manuscript No.**
(will be inserted by the editor)

# Detecting structural changes in large portfolios

**Peter N. Posch · Daniel Ullmann · Dominik Wied**

Received: date / Accepted: date

**Abstract** Model free tests for constant parameters often fail to detect structural changes in high dimensions. In practice, this corresponds to a portfolio with many assets and a reasonable long time series. We reduce the dimensionality of the problem by looking a compressed panel of time series obtained by cluster analysis and the principal components of the data. Using our methodology we are able to extend a test for a constant correlation matrix from a sub portfolio to whole indices and exemplify the procedure with the NASDAQ-100 index.

**Keywords** correlation · structural change · cluster analysis · portfolio management

**JEL classification** C12 · C55 · C58 · G11

**Acknowledgements** Financial support by the Collaborative Research Center Statistical Modeling of Nonlinear Dynamic Processes (SFB 823) of the German Research Foundation (DFG) is gratefully acknowledged.

## 1 Introduction

Portfolio optimization is most often based on the empirical moments of the portfolio constituents' returns, where the diversification effect is based on some measure of pairwise co-movement between the constituents, e.g. correlation. Whenever the characteristics of either the individual moments or the correlation changes, the portfolios

Peter N. Posch
TU Dortmund University, Faculty of Business, Economics and Social Science, 44227 Dortmund, Germany
E-mail: peter.posch@udo.edu

Daniel Ullmann
TU Dortmund University, Faculty of Business, Economics and Social Science, 44227 Dortmund, Germany
E-mail: daniel.ullmann@udo.edu

Dominik Wied
University of Cologne, Faculty of Business, Economics and Social Science, 50923 Cologne, Germany
E-mail: dwied@uni-koeln.de

optimality is affected. Thus, it is important to test for the occurrence of changes in these parameters and there exist tests for detecting retrograde *structural breaks*, (see Andreou & Ghysels (2009); Aue & Horváth (2013) for an overview). In the last few years, there is a growing interest in the literature for detecting breakpoints in dependence measures, especially for the case that the time of potential breaks need not be known for applying the test. People look, e.g., at the whole copula of different random variables (Bücher et al. 2014), but also at the usual bivariate correlation (Wied et al. 2012). The motivation for such approaches comes from empirical evidence that correlations are in general time-varying (see Longin & Solnik (1995) for a seminal paper on this topic), but it is unclear whether this is true for both conditional and unconditional correlations.

In this paper, we investigate in detail potential changes in correlation using the nonparametric fluctuation test for a constant correlation matrix proposed by Wied (2015). This test, as many others constructed in similar fashion, needs a high number of time observations relative to the number of assets for sufficient size and power properties. In practice, a typical multi asset portfolio has several hundreds of assets under management, but the joint time series for all assets is considerably smaller.

So how can we test for a structural break of the correlation structure of a portfolio where the number of assets is large and possibly larger than the number of observations? Our approach is to reduce the dimensionality and then applying the Wied (2015) test to the reduced problem. We consider two classical approaches to reduce the dimension of the problem, in order to check, if they can be used, or not. In the latter, more advanced methods like the Multi-way principal components analysis (Wold et al. 1987) could be used.

First, we employ cluster analysis, second principal component analysis, cf. e.g. Fodor (2002) for a discussion of their use in dimensionality reduction. Cluster analysis has a wide range of applications such as biology (Eisen et al. 1998), medicine (Haldar et al. 2008) and finance, resp. econophysics (Bonanno et al. 2004; Brida & Risso 2010; Mantegna & Stanley 1999; Mantegna 1999; Tumminello et al. 2010), where it is also applied in portfolio optimization (Onnela et al. 2002, 2003; Tola et al. 2008). Yet, as far as we know, we are the first to combine clustering and tests for structural breaks, which might be an interesting contribution to the existing literature on portfolio optimization. The principal component analysis (PCA) dates back to Pearson (1901) and is one of the most useful and most popular techniques of multivariate analysis (Hallin et al. 2014) with wide applications in finance, cf. Greene (2008) or Alexander (1999). It is suitable for reducing the dimensionality of a problem (Fodor 2002), since its central idea is the transformation and dimension reduction of the data, while keeping as much variance as possible (Jolliffe 2002).

The structure of the paper is as follows: In section 2 we develop the test for structural changes in a large portfolio context. In section 3 we apply the test to a real-life data set, report the resulting clusters and present the result of our analyses, while the final section 4 concludes.

## 2 Methodology

In this section we develop the test approach to detect structural breaks in a large portfolios correlation structure. The basis for our analysis is the correlation matrix:

$$\underline{\rho} = (\rho_{i,j})_{i,j=1,\dots,n} \text{ where } \rho_{i,j} = \frac{E\big((x_i-\mu_i)(x_j-\mu_j)\big)}{\sqrt{\sigma_j^2\sigma_i^2}} \tag{1}$$

with $\mu_i$ as the first moment and $\sigma_i^2$ as the variance of the corresponding time series. For the estimator $\hat{\underline{\rho}}$ we use the empirical average $\hat{\mu}_i$ and the empirical variance $\hat{\sigma}_i^2$. Since the estimated correlation matrix has to be positive definite (Kwan 2010), we need more observations than assets. Looking at $n$ assets with $T$ observations, a sequence of random vectors $X_t = (X_{1,t}, X_{2,t}, \dots, X_{n,t})$, we alculate the correlation matrix from the first $k$ observations according to equation 1 and denoted as $\hat{\underline{\rho}}_k$.

We set the hypotheses as
$H_0 : \underline{\rho}_t = \underline{\rho}_{t+\tau} \forall t = 1,\dots,T \ \tau = 1,\dots,T-t \quad$ vs.
$H_1 : \exists t, \tau : \underline{\rho}_t \neq \underline{\rho}_{t+\tau}$
and define the difference to the correlation matrix from all $T$ observations as

$$\hat{P}_{k,T} = vech\left(\hat{\underline{\rho}}_k - \hat{\underline{\rho}}_T\right), \tag{2}$$

where the operator $vech(A)$ denotes the half-vectorization:

$$vech(A) = (a_{i,j})_{1\le i<j\le dim(A)} \tag{3}$$

The test statistic is given by Wied (2015) as

$$\hat{A}_T := \max_{2\le k\le T} \frac{k}{\sqrt{T}} \left\| \hat{\underline{E}}^{-\frac{1}{2}} \hat{P}_{k,T} \right\|_1, \tag{4}$$

where $\|\cdot\|_1$ denotes the L1 norm. The null is rejected if $\hat{A}_T$ exceeds the threshold given by the 95% quantile of $A$ with the following definition:

$$A := \sup_{0\le s\le 1} \left\| B^{\frac{n(n-1)}{2}}(s) \right\|_1 \tag{5}$$

Here $B^k(s)$ is the vector of $k$ independent standard Brownian Bridges. If the test statistic exceeds this threshold, we define the structural break date $k$ as the following time point:

$$\arg\max_k \frac{k}{\sqrt{T}} \left| \hat{P}_{k,T} \right| \tag{6}$$

For the limiting distribution of the statistic we need a 'scaling' matrix $\hat{\underline{E}}$, which can be obtained by bootstrapping in the following way:

- We define a block length $l$ and devide the data into $T-l-1$ overlapping blocks:

$$\underline{B}_1 = (X_1,\dots,X_l), \underline{B}_2 = (X_2,\dots,X_{l+1}), \dots$$

- In each repetition $b \in [1,B]$ for some large $B$, we sample $\lfloor \frac{T}{l} \rfloor$ times with replacement one of the blocks and merge them all to a time series of dimension $l \cdot \lfloor \frac{T}{l} \rfloor \times n$.
- For each bootstrapped time series we calculate the covariance matrix. We convert the scaled elements above the diagonal to the vector $\hat{v}_b = \sqrt{T}\left(\hat{\rho}^{i,j}_{b,T}\right)_{1\le i<j\le n}$
- We denote the covariance of all these vectors with $\hat{\underline{E}} := Cov(\hat{v}_1, \ldots, \hat{v}_B)$

For the original test to provide a good approximation of the limiting distribution the ratio of observations to assets needs to be much larger than one. A portfolio with a large number of assets and insufficient time length of observation thus cannot be analyzed with the present test for structural changes in its correlation matrix.

To quantify this problem, we simulate for 4, 6 and 10 assets time series with length 200, 500, 1000 and 2000 observations. To assess the power and size of the test, we simulate in one case with a constant correlation matrix and in another with a break in the middle of the time series. For the latter, we choose randomly with replacement $n/2$ correlations ($\rho^*_{i,j}$) in the given correlation matrix randomly to a new correlation in [-1,1]. We transform the resulting matrix into a positive definite matrix, and consequently the resulting matrix differs from the original one in all cases, but most prominently in the randomly selected $\rho^*_{i,j}$.

The result of the simulation study is shown in table 1. Thereby we use the critical values 4.47, resp. 9.13, resp. 23.21 for the number of assets being equal to 4, resp. 6, resp. 10. Our simulations runs 2500 times.

< Insert table 1 about here >

We find comparable results as in Wied (2015) for the size distortions. It seems that the level converges to the 5% level. Concerning the power, we find an increase in the number of observations and a decrease with the number of assets. A very extreme decrease can be found for 500 observations and the shift from 6 to 10 assets. Whereas in the former case we find a rejection rate of about 72% it drops in the latter case to only about 20%. Such drops seem to exist for all cases, but are not that prominent. The lower rejection rate in the case of 4 assets can be explained in the random choice of pairwise correlations. Since we draw with replacement, the chance of a given pair is drawn twice considerably higher than for 6 and 10 dimensions and thus the size of the break in the correlation structure is smaller when compared to the number of assets.

In the following, we analyze how the use of dimension reduction techniques before applying the test changes these findings. We discuss two techniques, clustering and principal component analysis, to reduce the dimensionality of the problem and apply the test afterwards. A first approach is to use exogenous clusters such as e.g. industry sectors which, however, imposes an exogenous structure possibly not present in the data. Clustering endogenously based on the present correlation structure instead preserves this information cf. Mantegna (1999), an approach which is applied widely, e.g. in financial markets by Brida & Risso (2010), in medicine by Eisen et al. (1998).

The first step is to transform the correlation into a distance metric d fulfilling the following four requirements (7)–(10) where in the application of a clustering

algorithm (10) is replaced by the stronger (11), since an ultra-metric is used.

$$d(x,y) \geq 0 \tag{7}$$
$$d(x,y) = 0 \Leftrightarrow x = y \tag{8}$$
$$d(x,y) = d(y,x) \tag{9}$$
$$d(x,z) \leq d(x,y) + d(y,z) \tag{10}$$
$$d(x,z) = \max\{d(x,y), d(y,z)\} \tag{11}$$

Following Anderberg (2014) and Mantegna (1999) we use

$$d(x_i,x_j) = \sqrt{2(1-\rho_{i,j})}, \tag{12}$$

which is the Euclidean distance between the standardized data points $x_i$ and $x_j$. The metric is bounded in the interval $d \in [0,2]$ and smaller values correspond to a smaller distance and thus to more similarity. The clustering algorithm itself runs as follows:

1. Find the pair $i, j$ which satisfies: $d(x_i,x_j) = \min_{m,n} d(x_n,x_m)$
2. Merge the pair $i$ and $j$ into a single cluster
3. Calculate the distance to the other clusters
4. Repeat steps 1 and 2 as often as desired

To calculate the distance in step 3, there exist several algorithms, which all result in different cluster constituents. In order to choose the 'best' one, we use the concept of equally sized clusters: If we seek to form $m$ clusers out of $n$ assets and one of them contains $n-m+1$ assets, we end up with the most unequal cluster size possible. In this extreme case the likelihood for a randomly chosen asset to be in the one large cluster is highest and its cluster weight lowest. In contrast clusters are uniformly distributed in size the sensitivity of the cluster formation is higher. To gain a homogenously sized clustering we use the (Herfindahl) Hirschman index (Hirschman (1964, 1980)).

In our application, the ward algorithm, where the criterion for selecting two clusters to merge is such that the variance within them becomes minimal (Anderberg (2014); Murtagh & Legendre (2014); Ward Jr (1963)) is the one which satisfies best our specifications. In general, it leads to the most homogeneous clusters, due to the minimal variance criterion, while other linkage algorithms like *complete linkage*, *single linkage* or *average linkage* are based on the correlation. Complete linkage is reacting most conservative to correlation, the single linkage most aggressively and the average linkage providing a middle way between these two. All algorithms result in a hierarchical, which does not make any statement about the clusters itself. Instead they are formed by cutting the hierarchical tree horizontally at a height such that the desired amount of clusters is formed.

In a final step we transform the each cluster into a cluster-portfolio, which is a sub-portfolio of the initial portfolio of all assets. In general the weights needed are free parameters and a choice for their determination is needed. As examples one can use the market capitalization, like in the NASDAQ-100 index, weights according to the position in the portfolio observed or simple equal weights. Since the first two

methods impose an external structure, which prefers some assets to others, we use the third option of equally weighting all assets within each cluster-portfolio.

This concludes the clustering approach. We now turn to the PCA as an alternative technique to reduce the dimensionality. Here the given data $\underline{X}$ in the high dimensional Cartesian coordinate system are projected onto another orthonormal coordinate system, which is based on the eigenvalue decomposition of the data's covariance matrix.

Let $\underline{\Sigma} = \mathrm{Var}(\underline{X}) \in \mathrm{R}^{n\times n}$ denote the empirical covariance matrix. Than there exists a transformation (Bronstein 2012) as:

$$\underline{\Sigma} = \underline{P}\,\underline{\Lambda}\,\underline{P}', \quad \underline{\Lambda} \in \mathrm{R}^{n\times n}, \quad \underline{P} \in \mathrm{R}^{n\times n} \tag{13}$$

where $\underline{P}$ is the matrix formed by the eigenvectors of $\underline{\Sigma}$ and $\underline{\Lambda}$ as a diagonal matrix with the eigenvalues $\lambda_i$ on its diagonal. The matrix of principle components is calculated as $\underline{Z} = \underline{X}\underline{P}$. This is the representation of the given data in the eigenvector coordinate system. For the variance of the rotated data we get

$$\begin{aligned}\mathrm{Var}(\underline{Z}) &= \mathrm{Cov}\left((\underline{X}\,\underline{P})', \underline{X}\,\underline{P}\right) = \mathrm{Cov}\left(\underline{P}'\underline{X}, \underline{X}\,\underline{P}\right) \\ &= \underline{P}'\mathrm{Cov}\left(\underline{X}', \underline{X}\right)\underline{P} = \underline{P}'\underline{\Sigma}\,\underline{P} = \underline{\Lambda}\end{aligned} \tag{14}$$

which is a diagonal matrix and therefore uncorrelated data in the rotated system.

By using only the k largest eigenvalues, we can now reduce the dimensions in the orthonormal base. The obtained data is an approximation of the original data but with only $k$ dimensions in the orthonormal base and $n$ dimensions in the Cartesian space, cf. Hair et al. (2006). But the following concern comes with the application of the PCA: As the dimensions increases, we have to neglect relatively more dimensions, since only a rather small number of time series in the basis of the eigenvectors. We are then left only with the high volatile dimensions and a structural break detected therein can be regarded as a rather big and thus most obvious break. In this light we like to address the error made in neglecting small volatile dimensions. In general, the variation in the small eigenvalues corresponds to noise in the estimation of the true correlation matrix, cf. Laloux et al. (1999, 2000); Plerou et al. (2002). As a result one can obtain a more accurate estimate when intentionally not using these dimensions. Thus is not true that a reduction in the percentage of variation represented by the first k eigenvalues likewise reduces the likelihood of detecting a structural break. Improving the accuracy of the estimated correlation may even occur, since noise has been neglected. Concluding, we only have to worry about the information associated with eigenvalues, which are smaller than the fourth largest eigenvalue and larger then this threshold eigenvalue.

## 3 Limiting distribution and finite sample evidence

The two methods of clustering and PCA are two approaches to reduce the dimension of the problem at hand resulting in a linear transformation of the given data and forming actually indices or sub-portfolios, which are analyzed in the Euclidean or another "Eigen" space. Both transformations are linear. Thus the limiting distribution

should not change. An economic interpretation is the following: Suppose one chooses stocks from a given universe and creates portfolios thereof. This investor then could use a test to analyze the correlation structure, resp. breaks in the structure between her portfolios. No matter how the weights are constructed, e.g. all weights are unequal to zero (PCA case) or some of them are zero (hierarchical tree case), a valid financial time series is created.

To test this, we simulate 2500 realizations of 1000 assets with 1500 time points and a fixed random correlation matrix. We randomly choose the eigenvalues in the interval $[1,10]$ and use the columns of a random orthogonal matrix as the corresponding eigenvectors. Figure 1 shows the histogram of the test statistic on the left hand side and the corresponding empirical p values on the right hand side.

< Insert figure 1 about here >

We do not find deviations from the original limiting distribution in Wied (2015) for both methods. The p values seem to be uniformly distributed, so we conclude, that the limiting distribution does not change. Since these transformations are all linear, this result is what one can expect and the test seems to have the same validity as the original. Furthermore, we replace the generic correlation matrix from above with an empirical estimate from a sample of 50 stocks from the NASDAQ-100 index. Applying the same methodology as above, we find a very similar shape of the test statistic and of the distribution of the p-values.

The finite sample properties of the proposed test using a Monte-Carlo simulation clarifies what breaks are detected, resp. what it means, if the test statistic exceeds the critical value. We illustrate these for different number of observations per time series $t \in \{500, 1000, 2000, 4000\}$, and simulate such from a 100-dimensional normal distribution. In one case the correlation matrix is kept fix, in the other case we include a break in the middle of the time series. To do so, we change randomly $n_{\rho^*_{i,j}} \in \{100, 1000\}$ entries in the correlation matrix and transform it in a positive definite matrix. It is then used for the second half of the time series. We use 3, 4 and 5 clusters (with corresponding critical values of 2.27, 4.47 and 6.60) and check if the maximum of the test statistic is larger than the corresponding critical value. We repeat this process 2500 times. The result is shown in table 2. In all cases we find an increasing power in the number of reduced dimensions, but as a trade off the size distortion increases as well.

< Insert table 2 about here >

The upper part of the table shows the empirical rejection rate given H0 holds. For the PCA case, the size distortion seems to converge monotonously to the level of 5%. For the hierarchical tree, we find comparable sizes of size distortions but the convergence process seems not to be monotonous. The middle part corresponds to the case, where the break in the correlation structure of size $n_{\rho^*_{i,j}} = 100$ occurs. Keeping in mind that in a portfolio of 100 assets, a change in 100 pairwise correlations corresponds to roughly one asset changing its correlation to all other assets. So this is a rather small break in our setting. We find for nearly all cases a higher rejection rate for the hierarchical tree approach than for the PCA. The lower part corresponds to

the break size of $n_{\rho^*_{i,j}} = 1000$. Compared to the $100 \cdot 99/2 = 4950$ pairwise correlations overall, this is a rather big break. We find higher rejection rates compared to the small break before. This is what one expects, since the L1 norm increases and thus the probability of an exceedance of the critical value. In contrast to the case before, the rejection rates are higher for the PCA than for the hierarchical tree approach. This suggest that no approach dominates the other and when using both approaches one can create two information sets, with a non-empty overlap. Compared with the original test, we find comparable rejection rates and size distortions for a break in 1000 pairwise correlations. This suggests that when dealing with big breaks in a large portfolio the upstream clustering creates time series which can be treated as in the original test, such that the test has comparable finite sample properties.

## 4 Correlations within and between clusters

In the following we answer the question which correlation breaks we are able to detect. Using the notation of figure 2, we check for the correlation between clusters ($\rho$) and not for correlations within cluster ($\rho_{i,j}$) in the first place.

< Insert figure 2 about here >

An implicit assumption is that the correlation changes within a cluster are transmitted to the clustered time series and thus are implicitly detectable. Consider the situation in figure 2. We have 4 assets (symbolized as circles) and form 2 clusters (symbolized as ellipses). Suppose the time series of the single assets are denoted as $x_i$ for $i \in \{1,2,3,4\}$ with zero mean and finite variance. Using the constant cluster weights $w_i$, the clustered time series $y_i$ are calculated as follows:

$$y_1 = w_1 x_1 + w_2 x_2 \tag{15}$$
$$y_2 = w_3 x_3 + w_4 x_4 \tag{16}$$

For the correlation between the clusters, it holds:

$$\rho := \mathrm{Cor}(y_1, y_2) = w_1 w_3 \rho_{13} + w_1 w_4 \rho_{14} \tag{17}$$
$$+ w_2 w_3 \rho_{23} + w_2 w_4 \rho_{24} \tag{18}$$
$$\tag{19}$$

Since the weights are constant over time, a change in this correlation is then a consequently a weighted sum of the single pairwise correlation changes:

$$\mathrm{d}\rho = \mathrm{dCor}(y_1, y_2) = w_1 w_3 \mathrm{d}\rho_{13} + w_1 w_4 \mathrm{d}\rho_{14} \tag{20}$$
$$+ w_2 w_3 \mathrm{d}\rho_{23} + w_2 w_4 \mathrm{d}\rho_{24} \tag{21}$$

It follows that the change of correlations within a cluster does not change the correlation between clusters per se, since $\mathrm{d}\rho$ is independent of e.g. $\mathrm{d}\rho_{1,2}$. Besides

a very low probability of a change of one asset with solely one another within the cluster, there is also the condition of a positive definite correlation matrix of the whole $x_i$. So a change of e.g. $\rho_{1,2}$ cannot be independent of all other things, so that the other correlations are affected as well.

To check if the cluster formation is a stable process over time, we use a sample of 50 stocks over a timespan of 1500 observations and calculate the cluster members for each time point $t \in [100, 1500]$. We choose a lag of 100 data points to get a stable first estimate. We can conclude that cluster changes occur frequently especially at the beginning of the time interval which corresponds to the stability of the correlation matrix estimation. Also there exist many breaks in the correlation structure, which are so big, that assets would leave the cluster. In the following we simulate such breaks. We use 4 normal distributed variables correlated with a given correlation matrix with lengths of 500,1000,1500 and 2000 data points. To simulate a larger cluster, in which only one asset changes the within correlation, we set the weights randomly in the interval (i) $[0.1, 0.5]$, (ii) $[0.01, 0.05]$ and (iii) $[0.01, 0.5]$. So we can compare the behavior for small (i) and large (ii) clusters. Case (iii) serves as an overall assessment.

As a starting point we chose the matrix

$$\underline{\rho}_0 = \begin{pmatrix} 1 & & & \\ 0.7717 & 1 & & \\ 0.2597 & 0.1328 & 1 & \\ -0.0589 & 0.1665 & 0.8523 & 1 \end{pmatrix}$$

which is positive semi definite and has a high correlation between assets 1 and 2 and assets 3 and 4. This matrix results in the desired clusters, so we create the situation shown in figure 2. This is our starting position in each simulation and we now distinguish the following cases:

(a) We change only $\rho_{12}$: This corresponds to fluctuations within a cluster that are not so large that the asset would leave the cluster. To do this, we set this value to a random number in the interval $[0.7, 1.0]$.
(b) We decrease $\rho_{12}$ and increase $\rho_{23}, \rho_{24}$: This corresponds to a change of the cluster members in a way, that assets 2, 3 and 4 build a cluster. In order to simulate this, we set $\rho_{12} \in [-0.2, 0.2], \rho_{32} \in [0.6, 1.0], \rho_{42} \in [0.6, 1.0]$.
(c) We change the correlations of assets 2 and 3: This corresponds to a situation where a new clustering would result in a cluster of asset 1 and 3 and a cluster of asset 2 and 4. Thus we set $\rho_{12} \in [-0.2, 0.2], \rho_{13} \in [0.6, 1.0], \rho_{24} \in [0.6, 1.0], \rho_{34} \in [-0.2, 0.2]$.

In all cases we draw from the given interval until a positive semi definite matrix with the desired entries is obtained which is then used to change the time series correlation at the middle of the simulated time span. Table 3 reports the detection of test, where 'BC' refers to the situation where we have just two equally weighted assets in each of the two clusters.

< Insert table 3 about here >

The table shows the cases (a) to (c) with the corresponding definitions from above. In general the size of the change in the correlation structure increases from (a) to (c). So we find a general tendency for higher rejection rates when moving from the left to the right hand side of the table.

The table is split up into four horizontal parts. The upper one, labeled with 'BC' is our base case, where we simulate only 2 assets in the two clusters. we find rejection rates of about $4\% - 5\%$ for the 95% significance level, which is a very poor power. When to size of break increases, the power converges to 1. also we find a general convergence with an increasing number of observations. For the second part, labeled with (i) we simulate a larger portfolio of 2-10 assets in the corresponding cluster. We find lower rejection rates as in the base case in general, which is expected, since the break size, relative to the number of other assets decreases. When dealing with a big break, we find for a sufficient number of observations high detection rates. The third part, labeled with (ii) simulates a portfolio of 20-100 clustered assets. Thus the relative size of the break decreases even more and the rejection rates drop further. Only in the case (c), where the break is such, that a re-clustering would lead to a switch of assets, rejection rates converge to 1 for a realistic length of the time series. The last part, labeled with (iii) we simulate a portfolio with 2-100 in the corresponding cluster. Thus it is located between (i) and (ii), with rejection rates in between (i) and (ii) for the corresponding observations and scenarios.

Concluding we find that the structural break test is not only able to detect breaks between clusters, but it may also be able to detect larger fluctuations within (unchanged) clusters. But this does not mean that finding no breaks between clusters means that there are no breaks within clusters.

## 5 Application to a real-life problem

We now turn to the application of the before mentioned methods. We use the stock prices of the constituent of the NASDAQ100 index as of November 11th, 2015. Our observations span from January 2000 to December 2016. Thus we proxy a rather large portfolio, although it does not represent the NASDAQ100 index at all times, since we do not adjust for fluctuations in the index constituents. We calculate log returns for all series and cluster them into 3, 4 and 5 clusters. We use the algorithm described in Galeano & Wied (2014) to detect multiple breaks in the correlation matrix.

Figure 3 to 5 show the clustered time series for reduced dimension of 3, 4 and 5. The left hand side corresponds to the hierarchical clustering and the right hand side to the PCA. It can be seen that, while the volatility in the time series decreases in the PCA case with decreasing eigenvalues, this is not necessarily the case for the hierarchical tree. In the former case, we can preserve 40.8%, 44.1% and 46.8% percent of the total variation in the original data.

< Insert figure 3 about here >

< Insert figure 4 about here >

< Insert figure 5 about here >

At the beginning of our time interval from 2000 onwards to roughly 2003, we find very volatile time series. This could be connected to the Dot-Com Bubble which affected most the technology companies. Another obvious peak in volatility is located in the year 2008 and corresponds roughly to the default of Lehman Brothers, which is probably the best known starting point of the global financial crisis in public. As the NASDAQ index does not contain financial companies, figure 3 to figure 5 illustrate the distress in the whole market, not only in the financial sector.

The breaks in the correlation are indicated by the vertical lines in the figures and are listed in table 4.

< Insert table 4 about here >

In most cases, we find comparable results for the two clustering algorithms, for a given dimension. In two of the three cases, we find the same number of breaks, but we find deviations in the time location in general. A very short time difference is a break identified in 3 dimensions in 2013. Here the time difference is roughly one month when comparing the PCA with the hierarchical tree. As a more general example, for dimensions 4 and 5 the break date differs in about a year time difference. In the case of 3 dimensions, the first break is located in 2002 for PCA, whereas it is located in 2008 for the hierarchical tree. The latter case corresponds to 12 days before the bankruptcy of Lehman Brothers.

Keeping the clustering algorithm and looking at different dimensions, we find very similar break dates. For the hierarchical tree, a break in the second half of 2008 is indicated for all dimensions, as well as a break at the beginning, reps. middle of August 2015. For the PCA case, a break in the 3rd quarter of 2007, one in September 2014 and in September 2015 is located. Thus we are confident that these are actual breaks. Additionally there are four other breaks in the range from end 2010 to end 2012. The break in 2010 is only indicated once, as well as the one in 2011. The two breaks in 2012 for the PCA case in 4 and 5 reduced dimensions have a five month time difference.

On the other hand, we also find some outliers. As mentioned before, the first break in 2002 in the PCA case for 3 dimensions is not indicated by any of the other 5 tests. Another example is the break in 2016 detected by the hierarchical tree with 4 dimensions.

## 6 Conclusion

Monitoring the correlation matrix of a portfolio is a daily task in financial portfolio management and modern portfolio optimization heavily relies on the matrix in calculating the portfolio weights. Likewise triggers the occurrence of changes in this correlation structure, a structural break, a portfolio adjustment. While test for structural breaks are readily available and discussed in the literature, they can only deal with a low number of assets, e.g. Wied (2015) with only four assets. In practice, however, we most commonly have a situation in which the number of assets by far exceeds the available time series. Thus so far no test on structural changes in the correlation structure of typical portfolios was possible.

In order to tackle this dimensionality problem, we propose the use of a hierarchical tree as foundation for cluster formation. This technique has already applied to cluster financial markets, especially in the econophysics literature. As an alternative, the standard principal component analysis is used to form sub-portfolios. In a second stage a test on structural changes is then applied to the reduced problem.

Using a Monte Carlo simulation, we showed that we are able to extend the original test to a larger portfolio in the sense that we can reproduce the power with the corresponding size distortion as the original test for a sufficient big break in the correlation matrix for a large portfolio. Although we check for breaks between clusters we showed that, due to the fact of a positive semi definite correlation matrix, a change in the correlation structure within a cluster translates to some extend to the correlation between the clusters. We quantified this effect through a simulation as well.

Finally we applied both techniques to the case of constituents of the NASDAQ-100 index over a time period of 17 years. We do not check the index itself (since there are fluctuating constituents), but simulate a rather large portfolio compared to the 4 assets in the example in Wied (2015). We found some breaks in the correlation structure, which are independent of the reduced dimensionality and one, which is independent of the clustering approach at the first stage.

## 7 Figures and Tables

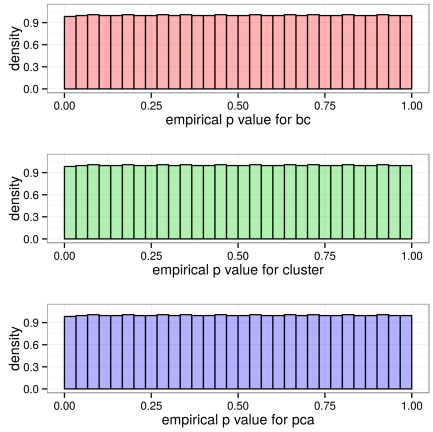


**Fig. 1 Empirical p values for the cluster and PCA approach.** The figure shows the p values of the test statistic for both the cluster and the PCA approach. The original test statistic of Wied (2015) with 4 assets is denoted with bc.

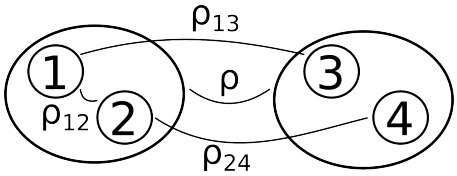


**Fig. 2 Notation for correlations within and between cluster.** Shown are 4 assets (circles), where 2 assets are included in a cluster (ellipse). The correlation between clusters is denoted as $\rho$ and the correlation between assets $i, j$ as $\rho_{i,j}$

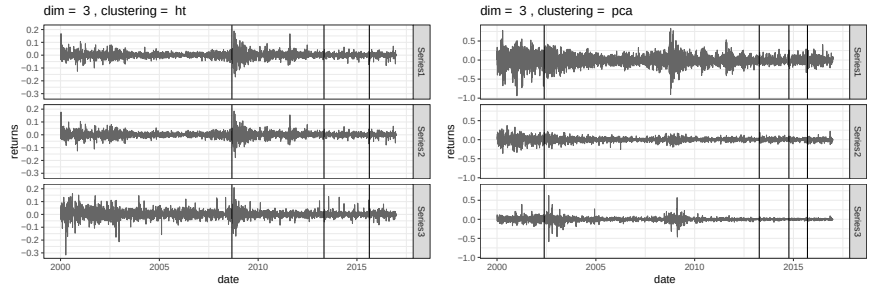


**Fig. 3 Break detection dates for 3 reduced dimensions.** The figure shows the three clustered time series together with the associated break dates. The left hand side corresponds to the hierarchical clustering and the right hand side to the PCA.

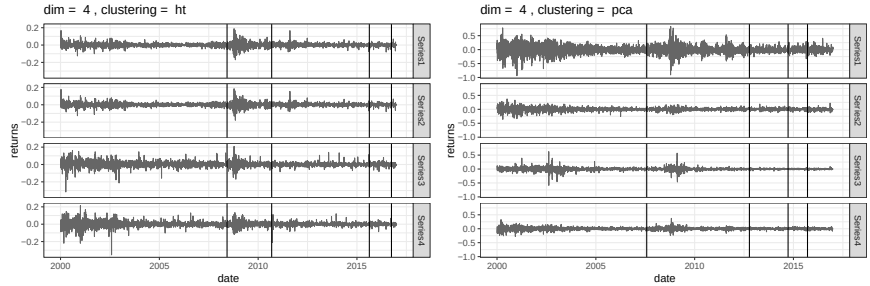


**Fig. 4 Break detection dates for 4 reduced dimensions.** The figure shows the three clustered time series together with the associated break dates. The left hand side corresponds to the hierachical clusteing and the right hand side to the PCA.

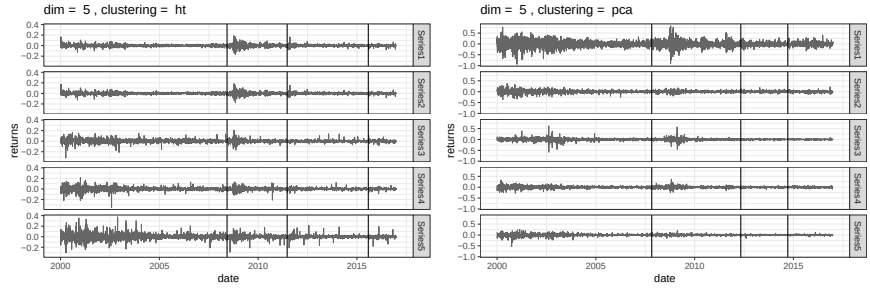


**Fig. 5 Break detection dates for 5 reduced dimensions.** The figure shows the three clustered time series together with the associated break dates. The left hand side corresponds to the hierachical clusteing and the right hand side to the PCA.

**Table 1** Empirical rejection rate (in percent) of H0 of the test in Wied (2015) based on 2500 simulations.

| | number of assets | | |
|---|---|---|---|
| | 4 | 6 | 10 |
| number of observations | $\Delta\rho^*_{i,j} = 0$ | | |
| 200 | 4.60 | 2.16 | 0.80 |
| 500 | 4.33 | 3.29 | 0.75 |
| 1000 | 4.07 | 4.01 | 1.30 |
| 2000 | 5.44 | 4.11 | 4.20 |
| number of observations | $\Delta\rho^*_{i,j} \neq 0$ | | |
| 200 | 39.48 | 31.77 | 4.73 |
| 500 | 70.31 | 71.98 | 19.73 |
| 1000 | 72.84 | 84.22 | 75.93 |
| 2000 | 79.64 | 96.49 | 95.13 |

**Table 2** Empirical rejection rate (in percent) of H0 given a portfolio of 100 assets and different number of clusters, given the 95% quantile.

| | reduced dimensions PCA | | | reduced dimensions HT | | |
|---|---|---|---|---|---|---|
| number of observations | 3 | 4 | 5 | 3 | 4 | 5 |
| | $\Delta\rho^*_{i,j} = 0$ | | | | | |
| 500 | 3.28 | 2.88 | 2.18 | 5.12 | 3.78 | 3.87 |
| 1000 | 4.58 | 3.81 | 2.66 | 4.54 | 4.67 | 3.13 |
| 2000 | 4.74 | 3.97 | 4.00 | 4.10 | 4.42 | 5.51 |
| 4000 | 4.80 | 4.67 | 4.14 | 4.32 | 4.93 | 3.97 |
| | $\Delta\rho^*_{i,j} \neq 0,$ | | | $n_{\rho^*_{i,j}} = 100$ | | |
| 500 | 6.06 | 10.31 | 16.56 | 11.76 | 12.75 | 19.43 |
| 1000 | 5.30 | 10.81 | 26.00 | 12.88 | 27.18 | 23.18 |
| 2000 | 7.75 | 17.00 | 35.31 | 16.38 | 14.56 | 43.56 |
| 4000 | 13.69 | 23.56 | 50.88 | 22.50 | 42.50 | 50.69 |
| | $\Delta\rho^*_{i,j} \neq 0,$ | | | $n_{\rho^*_{i,j}} = 1000$ | | |
| 500 | 75,14 | 97.43 | 99.86 | 46.29 | 60.85 | 77.57 |
| 1000 | 74.86 | 98.57 | 100 | 68.14 | 92.86 | 96.43 |
| 2000 | 84.14 | 99.43 | 100 | 87.00 | 98.29 | 99.43 |
| 4000 | 92.43 | 99.86 | 100 | 88.43 | 100 | 100 |

**Table 3** Empirical rejection rate (in percent) of H0 given two clusters of different assets. Base case has four assets, (i) 10-20 assets, (ii) 20-100 assets and (iii) 10-100 assets. The correlation structure was changed in the middle of the time series, depending on the scenario (a) with fluctuation within, (b) with an asset outflow or (c) with a switch of assets.

| | | (a) | | | | (b) | | | | (c) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | observations | | | | observations | | | | observations | | | |
| | $\alpha/\%$ | 200 | 500 | 1000 | 2000 | 200 | 500 | 1000 | 2000 | 200 | 500 | 1000 | 2000 |
| BC | 90 | 8.28 | 8.94 | 9.62 | 9.84 | 86.62 | 99.84 | 100 | 100 | 87.76 | 99.98 | 100 | 100 |
| | 95 | 4.26 | 4.54 | 4.66 | 4.52 | 77.06 | 99.62 | 100 | 100 | 75.38 | 99.86 | 100 | 100 |
| | 99 | 0.84 | 0.70 | 0.78 | 0.88 | 47.54 | 97.28 | 100 | 100 | 37.16 | 98.02 | 99.98 | 100 |
| (i) | 90 | 8.72 | 8.12 | 9.08 | 9.46 | 39.60 | 61.02 | 71.86 | 80.68 | 77.60 | 98.48 | 99.98 | 100 |
| | 95 | 3.82 | 3.92 | 4.52 | 4.98 | 29.76 | 53.52 | 66.96 | 77.30 | 62.64 | 96.72 | 99.86 | 100 |
| | 99 | 0.72 | 0.62 | 0.74 | 0.90 | 12.94 | 39.62 | 57.92 | 70.28 | 29.02 | 89.42 | 99.38 | 100 |
| (ii) | 90 | 8.18 | 8.32 | 8.76 | 8.68 | 8.68 | 8.84 | 9.73 | 12.44 | 61.24 | 92.96 | 99.56 | 100 |
| | 95 | 3.64 | 3.76 | 4.26 | 4.18 | 4.06 | 4.42 | 5.01 | 6.63 | 46.57 | 88.12 | 99.04 | 100 |
| | 99 | 0.56 | 0.56 | 0.56 | 0.98 | 1.04 | 0.70 | 1.04 | 1.87 | 20.11 | 71.95 | 96.39 | 99.96 |
| (iii) | 90 | 8.89 | 8.83 | 8.89 | 9.41 | 39.79 | 61.20 | 72.27 | 80.12 | 76.32 | 98.35 | 99.91 | 100 |
| | 95 | 4.31 | 4.29 | 4.25 | 4.87 | 29.69 | 54.12 | 67.56 | 76.24 | 61.51 | 96.92 | 99.85 | 100 |
| | 99 | 0.73 | 1.01 | 0.93 | 0.93 | 12.51 | 40.36 | 58.29 | 69.19 | 27.99 | 89.03 | 99.31 | 100 |

**Table 4** Detected structural breaks in the correaltion matrix of the NASDAQ 100 index according to different clustering algorithms and number of reduced dimensions.

| reduced dimensions | cluster | break dates | | | |
|---|---|---|---|---|---|
| 3 | ht | 3.9.2008 | 2.5.2013 | 18.8.2015 | |
| | pca | 22.5.2002 | 9.4.2013 | 9.10.2014 | 17.9.2015 |
| 4 | ht | 5.6.2008 | 8.9.2010 | 18.8.2015 | 30.9.2016 |
| | pca | 30.7.2007 | 9.10.2012 | 24.9.2014 | 18.9.2015 |
| 5 | ht | 5.6.2008 | 21.6.2011 | 30.7.2015 | |
| | pca | 31.10.2007 | 3.5.2012 | 19.9.2014 | |

## References


Alexander, C. (1999). *Optimal hedging using cointegration*, volume 357. The Royal Society.

Anderberg, M. R. (2014). *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*. Academic press.

Andreou, E. & Ghysels, E. (2009). Structural breaks in financial time series. In *Handbook of Financial Time Series* (pp. 839–870). Springer.

Aue, A. & Horváth, L. (2013). Structural breaks in time series. *Journal of Time Series Analysis*, 34(1), 1–16.

Bonanno, G., Caldarelli, G., Lillo, F., Miccichè, S., Vandewalle, N., & Mantegna, R. N. (2004). Networks of equities in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2), 363–371.

Brida, J. G. & Risso, W. A. (2010). Hierarchical structure of the German stock market. *Expert Systems with Applications*, 37(5), 3846–3852.

Bronstein, I. N. (2012). *Taschenbuch Der Mathematik*. Deutsch Harri GmbH; Auflage: 4., überarb. u. erw. A.

Bücher, A., Kojadinovic, I., Rohmer, T., & Segers, J. (2014). Detecting changes in cross-sectional dependence in multivariate time series. *Journal of Multivariate Analysis*, 132, 111–128.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25), 14863–14868.

Fodor, I. K. (2002). A survey of dimension reduction techniques.

Galeano, P. & Wied, D. (2014). Dating multiple change points in the correlation matrix. *TEST*, (pp. 1—-22).

Greene, W. H. (2008). *Econometric analysis*. Granite Hill Publishers.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ.

Haldar, P., Pavord, I. D., Shaw, D. E., Berry, M. A., Thomas, M., Brightling, C. E., Wardlaw, A. J., & Green, R. H. (2008). Cluster analysis and clinical asthma phenotypes. *American Journal of Respiratory and Critical Care Medicine*, 178(3), 218–224.

Hallin, M., Paindaveine, D., & Verdebout, T. (2014). Efficient R-estimation of principal and common principal components. *Journal of the American Statistical Association*, 109(507), 1071–1083.

Hirschman, A. O. (1964). The paternity of an index. *The American Economic Review*, (pp. 761–762).

Hirschman, A. O. (1980). *National power and the structure of foreign trade*, volume 1. Univ of California Press.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Kwan, C. C. Y. (2010). The Requirement of a Positive Definite Covariance Matrix of Security Returns for Mean-Variance Portfolio Analysis: A Pedagogic Illustration. *Spreadsheets in Education (eJSiE)*, 4(1), 4.

Laloux, L., Cizeau, P., Bouchaud, J.-P., & Potters, M. (1999). Noise dressing of financial correlation matrices. *Physical Review Letters*, 83(7), 1467.

Laloux, L., Cizeau, P., Potters, M., & Bouchaud, J.-P. (2000). Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(03), 391–397.

Longin, F. & Solnik, B. (1995). Is the correlation in international equity returns constant: 1960-1990? *Journal of International Money and Finance*, 14(1), 3–26.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1), 193–197.

Mantegna, R. N. & Stanley, H. E. (1999). *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press.

Murtagh, F. & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3), 274–295.

Onnela, J.-P., Chakraborti, A., Kaski, K., Kertesz, J., & Kanto, A. (2003). Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E*, 68(5), 56110.

Onnela, J.-P., Chakraborti, A., Kaski, K., & Kertiész, J. (2002). Dynamic asset trees and portfolio analysis. *The European Physical Journal B-Condensed Matter and Complex Systems*, 30(3), 285–288.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., Guhr, T., & Stanley, H. E. (2002). Random matrix approach to cross correlations in financial data. *Physical Review E*, 65(6), 66126.

Tola, V., Lillo, F., Gallegati, M., & Mantegna, R. N. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1), 235–258.

Tumminello, M., Lillo, F., & Mantegna, R. N. (2010). Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior amd Organization*, 75(1), 40–58.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.

Wied, D. (2015). A nonparametric test for a constant correlation matrix. *Econometric Reviews*, (pp. 1–21).

Wied, D., Arnold, M., Bissantz, N., & Ziggel, D. (2012). A new fluctuation test for constant variances with applications to finance. *Metrika*, 75(8), 1111–1127.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1), 37–52.