Nonparametric sign prediction of high-dimensional correlation matrix coefficients

Christian Bongiorno, Damien Challet

Nonparametric sign prediction of high-dimensional correlation matrix coefficients

Christian Bongiorno[(1)*] and Damien Challet[(1)]
[(1)] *Laboratoire de Mathématiques et Informatique pour les Systèmes Complexes, CentraleSupélec, Université Paris Saclay, 3 rue Joliot-Curie, 91192, Gif-sur-Yvette, France*
(Dated: October 16, 2019)

We introduce a method to predict which correlation matrix coefficients are likely to change their signs in the future in the high-dimensional regime, i.e. when the number of features is larger than the number of samples per feature. The stability of correlation signs, two-by-two relationships, is found to depend on three-by-three relationships inspired by Heider social cohesion theory in this regime. We apply our method to US and Hong Kong equities historical data to illustrate how the structure of correlation matrices influences the stability of the sign of its coefficients .

I. INTRODUCTION

Correlation matrices may be pathologically noisy without a proper filtering method. For example building optimal mean-variance portfolios [1] requires so precise estimations of trends, covariances and correlations that portfolio optimization was equated to error maximization by **(author?)** [2]. The main problem is that a precise estimation of a full unfiltered correlation matrix between N features requires $T \gg N$ samples per feature. Regrettably, the non-stationary nature of many real-life dynamical systems, including financial markets, imposes T to be as small as reasonably possible but in any case proportional to N. The impossibility to approximate the $T \to \infty$ limit while keeping N constant is known as the 'curse of dimensionality' as correlation estimators remain noisy even in the N and $T \to \infty$ limit at fixed ratio $q = T/N$.

Ad-hoc filtering techniques include linear shrinkage ([3]), block-diagonal ansatz for the correlation matrix ([4]) and random matrix theory-based eigenvalue clipping ([5, 6]). The latter works reasonably well for $T > N$. More recently, the Rotational Invariant Estimator (RIE), which makes use of eigenvectors as well, was shown to be optimal in the large N and T limit at constant ratio $q = T/N > 1$ ([7]). RMT and RIE assume stationary Gaussian returns and $N < T$ (see [8] for a recent review of the field). Here, we focus on non-stationary correlation structures of possibly non-Gaussian returns when $N > T$ and aim to predict the sign of asset correlations. Statistics calls the $N > T$ case high-dimensional and we will follow this terminology.

Here, we are considering a complex network representation of an asset correlation matrix. Historically, the first use of complex networks in finances is the Minimum Spanning Tree (MST) [9]. In this work, the author showed that a tree network could adequately describe the economic sector taxonomy of a portfolio of N assets, leading to a substantial complexity reduction, i.e., from $N \times N$ correlation coefficients to $N - 1$ links. Later on, in Ref. [10, 11], the authors proposed a relaxation of the topological constraint of the MST, leading to the Maximal Planar Graph, i.e., a filtered graph embedded in a bi-dimensional space. In this work, we use the most straightforward parametric procedure to obtain a network from a correlation matrix, which is the asset graph [12]. The asset graph prescribes to retain only links related to correlations exceeding a threshold value.

Whereas correlations involve two time series (they are dyadic), we find that triadic measures better quantify the global and local stability of the dependence and thus better predict the stability of the sign of correlations when $T < N$. Our approach is related to Heider balance theory ([13, 14]) which aims at explaining the attitude changes of interacting individuals. In the modeling framework of this theory, only two possible interactions between two individuals are possible: the latter can be *friends* or *enemies*. The general observation in social science that 'the enemy of my friend is my enemy' ([15]) becomes particularly relevant when extended to triadic relationships: for example, triads where a is a friend of b and c but c is an enemy of b tend to be unstable. As a consequence, one interaction type is likely to change and lead to a stable triad: a could become an enemy of b, or c could become a friend of b. In a similar way, a triad composed of three individuals that are enemies of each other is considered unstable as two individuals could join their forces against the third one. In summary, this theory identifies four possible triads, two stable ones and two unstable ones, and adds the intuition is that unstable triads tend to evolve into stable ones.

More recently, many authors in the field of network science proposed to extend the mechanism of triad balance to describe the evolution of a signed complex network ([16–19]). In particular, Hedayatifar [20] measured the global social balance with a Hamiltonian whose minimal energy level coincides with the maximal stability and studied the possible paths that drive the system towards minimal energy levels, i.e. to the maximally stable triad states. In a financial context, various properties of network structures have been used to characterize the state of the market ([21, 22]) or of interbank lending networks ([23, 24]).

Here, we link the stability of the dyadic relationships as encoded by correlation matrices and statistically validated networks to triadic relationships. Section II A de-

* christian.bongiorno@centralesupelec.fr

scribes how we processed the data-set; in Section II B, we illustrate the speed at which the correlation matrix structure changes in financial data and the relevance of the high-dimensional regime; Section II C shows how a social balance metric based on triads can describe such evolution; in Section II D, we demonstrate that the principle of social balance can be used to predict the stability of the a correlation matrix coefficient sign; Section III concludes.

II. RESULTS

A. Data Description and Processing

In this work we consider the daily close-to-close returns of equities from US and Hong Kong stock markets, adjusted for dividends, splits and other corporate events. We focus on large capitalization equities for which we have information on their official industrial sector. More precisely:

1. US equities: large-capitalization stocks, from 1992-02-03 to 2018-06-29. The number of stocks with data vary over time: it ranges from 399 in 1992-02-06 to 723 in 2018-06-29, and is roughly constant from 2008.

2. Hong Kong equities: 1277 stocks with the largest capitalization as of 2019-05-01 listed on Hong Kong stock exchange. Our dataset covers the 2002-01-04 to 2017-06-23 period. The number of stocks is similar to the US database: the minimum number of stocks was 320 stocks on 2002-01-01 and the maximum was 1277 on 2017-06-23.

Let $p_{i,t}$ be the matrix of the adjusted close prices, we denote the log-return matrix r whose elements are $r_{i,t} = \log(p_{i,t}) - \log(p_{i,t-1})$.

First, we define partial returns $\tilde{r}_{i,t} = r_{i,t} - m_t$ where m_t is the median of all price returns at time t (a nonparametric definition of the market mode ([25])) and their binarized values $b_{i,t} = \text{sign}(\tilde{r}_{i,t})$.

Then, for a given time window $[t - T + 1, t]$, we only keep the assets without any missing value and evaluate the correlation matrix Φ_t of the binarized returns b ([26]). Given the definition of b, Φ_t is nonparametric. We also will use the notation C_t to denote the correlation matrix of the raw returns $r_{i,t}$ in the time window $[t - T + 1, t]$.

Binarized returns certainly require less bits of storage. Whether they contain less information in practice depends on the situation and on the issue under investigation. For example, the cluster composition of US equities determined by Louvain clustering adapted to usual correlation matrices ([27]) is essentially the same if for Φ and C ([28, 29]). Here, we use binarized returns for two reasons: first to infer statistically validated networks, and second to build a robust nonparametric method.

For the sake of completeness (but not robustness), we repeated all the analysis with Pearson correlation matrices computed of raw returns in Appendix V A, and we achieved qualitatively similar results. In addition, an alternative way to remove the global trend from the Pearson correlation matrix is reported in section V B of the Appendix.

B. Fast correlation structure dynamics

We first illustrate how quickly the structure of correlation matrices change in financial markets, which explains why the prediction of correlation sign changes is important in this context. The idea is to infer statistically significant elements of Φ_t, which then defines a time-dependent adjacency matrix A_t whose evolution reflects some of the structural changes of the financial market in question.

We restrict Φ_t to its significantly positive elements by controlling for multiple hypothesis. Because Φ_t is computed from binary variables, we can use the one-sided Fisher exact test, which equivalent to the hypergeometric test ([30]). The coefficients which pass the test (at a false discovery rate set to $\alpha = 0.1$) form the adjacency matrix A_t whose coefficients are either $A_{ij,t} = 1$ if $\phi_{ij,t}$ is selected or 0 otherwise. This is known as Statistically Validated Networks (SVNs) which may be applied to more than two states ([31]). Section IV A gives for more details about the method. The main advantage of this approach is to obtain a filtered network of stocks even in the high-dimensional regime ($N > T$), without any assumption on the resulting clustering structure such as not-overlapping clusters and without resorting to bootstraps (see e.g. [32, 33])

It is well known that stocks belonging to the same sector are usually strongly correlated with each other: several methods can observe such emergent behavior, for example, the minimum spanning tree ([9]) or principal component analysis coupled with random matrix theory ([5, 6]). Here, we expect that assets that belong in the same sector form clusters in the SVN A_t. Among the arsenal of techniques to screen a complex network, we identified in the assortativity coefficient ([34]) as the most appropriate to measure the similarity with respect to the GICS sector composition. We shall drop the index t when it leads to too heavy notations. The assortativity coefficient is defined as

$$G = \frac{\sum_{ij} \left(A_{ij} - k_i k_j / 2m\right) \delta(c_i, c_j)}{2m - \sum_{ij} (k_i k_j / 2m) \delta(c_i, c_j)}$$

where $\delta(c_i, c_j)$ is 1 if node i and j belong to the same sector and zero otherwise, k_i is the degree of node i i.e. $k_i = \sum_j A_{ij}$ and m is the total number of links.

By definition, assortativity $G \in [-1, 1]$; its expectation equals 0 if the links of the networks are distributed randomly with respect to the sector partition in the case of

a configuration null-model. $G > 0$ indicates a propensity of the nodes to establish links between nodes the same sector and reversely.

This section reports results for US equities; those for Hong Kong equities are to be found in Section V C. We explored the dynamics of G by shifting a time window of $T = 100$ days one day at a time from 2000-01-03 to 2018-06-29. For each time window t, we computed the SVN A_t with $\alpha = 0.1$ and the related assortativity coefficient G_t.

The assortativity coefficient oscillates between 0.2 and 0.85 (Fig. 1(a)). One of the local minimum values is reached during the crisis of 2009 and is close to the random sector linking limit. It is worth noticing that although the link density of SVNs reaches local maxima in the proximity of the minima of the assortativity (Fig. 1(b)), the dynamical evolution of the assortativity should be unbiased by definition with respect to the number of links as indeed the assortativity coefficient is an adjusted metric that considers the configuration model as a null model, i.e., the family of models that preserves the exact degree distribution of the nodes.

In order to support our interpretation of the assortativity coefficient, we show a graphical representation of the networks obtained in 2007-07-20 and 2019-05-19 in Figs 1(c) and Fig. 1(d) respectively. The clusters observed in the network of Fig. 1(c), characterized by an assortativity of 0.8, has a clear association with the macroscopic structure defined by the sectors. However, in the network of Fig. 1(d), such association disappears. Such network, characterized by an assortativity of 0.2, seems to be composed by two large clusters, highly overlapping each other. Results for significantly negative correlations are reported in S.I.; in short, they lead to much sparser networks that are disassortative ($G < 0$) and only non-null in times of crisis.

Thus SVNs of binarized partial returns are able to capture part of the fast evolution of correlation structures in a way that overcomes the usual problems of correlation matrices in the high-dimensional regime $N > T$. That said, because SVNs are built by controling the false positive rate, they do not control the false negative rate, i.e,, the fraction of links that have been wrongly omitted (see [35] for a discussion on this point). The smaller the FDR of the SVN, the larger the risk of a larger false negative rate. Figure 1(b) illustrates this point: only a few links are retained in the SVN for most of the time periods and a high number of nodes are isolated.

At any rate, it is clear that the structure of correlation has a non-trivial fast dynamics, which can only be captured by small calibration windows. In the following, we focus on a specific part of structural changes, i.e. the change of correlation signs.

C. Triads Dynamic

We first further simplify the correlation matrix Φ by taking its sign and by setting its diagonal to 0: we introduce $S = \text{sign}(\Phi) - \mathbb{I}$. In short, one assumes that a link is positive if $\Phi_{ij} \geq 0$ and a negative when $\Phi_{ij} < 0$, and $S_{i,i}$=0. This time, an unknown fraction of false positives may be included in S. However, the global information emerging by considering the whole network structure will compensate for such errors.

The matrix S is nothing else than a signed adjacency matrix and makes it easy to define triads: there are four possible triads, two stable and two unstable (see Fig. 2) ones. In the case of asset returns, the two stable configurations correspond to a triangle of positively correlated assets, and to two assets that are positively correlated but negatively correlated to a third one. The two unstable situations involve two positive links and three negative ones.

Thus, triads with an odd number of negative links are stable and those with an even number of negative links are unstable. **(author?)** [20] introduce a global metric H to characterize the fraction of stable triangles in a system of N nodes, defined as

$$H = -\frac{1}{\binom{N}{3}} \sum_{ijk} S_{ij} S_{ik} S_{jk}. \tag{1}$$

A stable triangle adds $+1$ to the sum and an unstable one -1. The metric is normalized by the maximum number of triangles. Finally the minus sign ensures that a system with only stable triangles has $H = -1$, and a system characterized by only unstable triangles has $H = +1$. Thus, as pointed out in [20], H can be interpreted as the Hamiltonian of a physical system. Only two possible macroscopic states are possible in the lowest energy levels (the most stable ones): the 'paradise', where all the nodes have positive links with other nodes, and the 'bipolar' with two groups with positive interactions within the same group and negative interactions among different groups. Other compositions such as a clustering structure with $K > 2$ clusters can exist in a jammed state or be caused by an external force.

Within this modelling framework, the clustering structure such as the sector composition may therefore to be unstable with respect to perturbations and may evolve towards more stable structures if it was not for stabilizing forces that we do not explicitly account for in the following.

We observed the evolution of H with a time window of $T = 100$ days. As shown in Fig. 3(a), the dynamic of H is strongly correlated with the sector assortativity of the SVNs. In fact, the local minima of the assortativity correspond to the local minima of H, and similarly for the local maxima. This observation confirms that such a dynamic of the composition of the clusters can be well detected by H. It is worth noticing that the evolution of H is strongly anticorrelated with that of the volatility, see

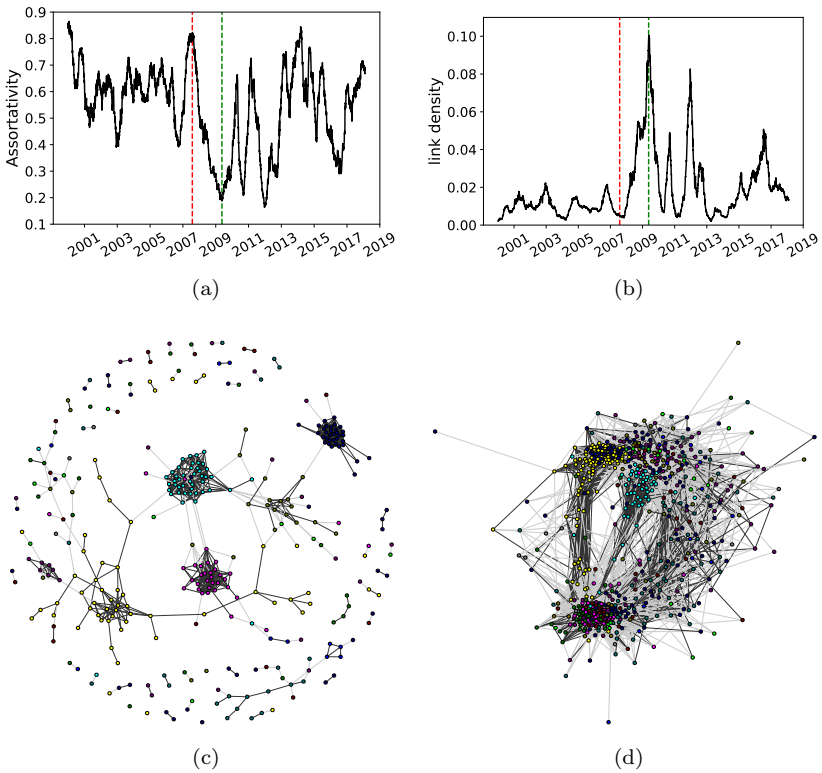


FIG. 1. (a) Assortativity of the network of statistically validated correlation (SVN) with respect to the sector classification, (b) links density of the correlation network, each point refers to the last day of the time window used to compute the SVN; (c) correlation network as of 2007-07-30 (dotted red line in panel (a)); (d) correlation network as of 2009-05-19 (dotted green line in panel (a)); the color code of the nodes of both networks represents different different sectors, the links between different sectors are colored in gray; calibration windows of $T = 100$ days.

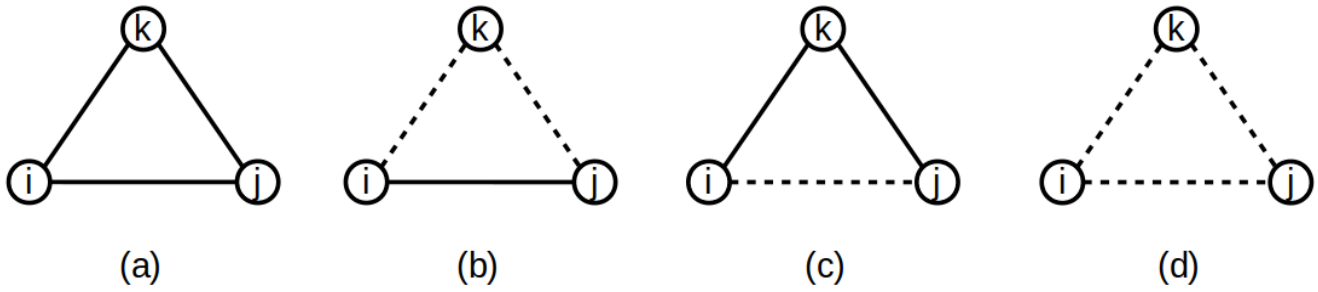


FIG. 2. (a) and (b) stable triads, (c) and (d) unstable triads. The solid line indicates a positive relation, the dotted line indicates a negative relation.

Fig. 3(b). In particular, a large change of the volatility in many cases precedes a similar event for H, as in 2009. This suggests that the volatility could be related with the perturbation that moves the system away from the jammed state characterized by the sector structure

1. *Triads and Spectral Decomposition*

According to the Spectral Decomposition Theorem, a symmetric matrix can be written as a sum of its eigenvectors weighted by their respective eigenvalues

$$\Phi = \sum_{i=1}^{N} \lambda_i \mathbf{v}_i' \, \mathbf{v}_i, \tag{2}$$

where λ_i is the i-th eigenvalue, $\mathbf{v}_i$ its associated eigenvector, and $\mathbf{v}_i'$ the transpose of $\mathbf{v}_i$. In addition, since cor-

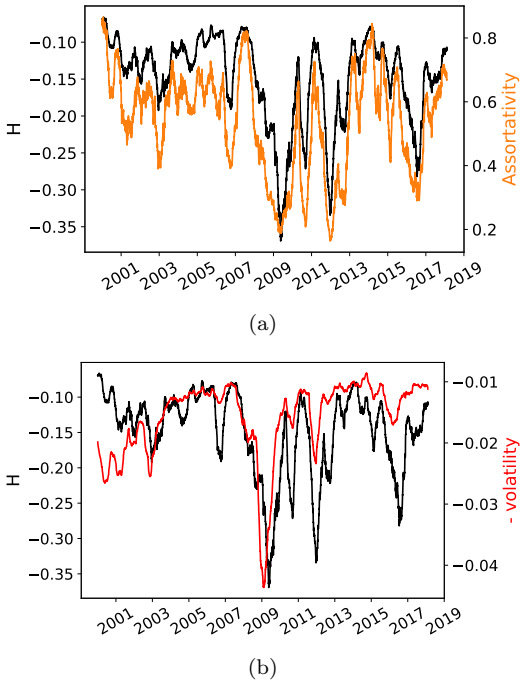


FIG. 3. (a) Evolution of H (black line), assortativity coefficient of the SVNs with respect to the sector composition (orange line); (b) Evolution of H (black line), minus volatility (red line) defined as the average absolute value of the returns in the considered time window.

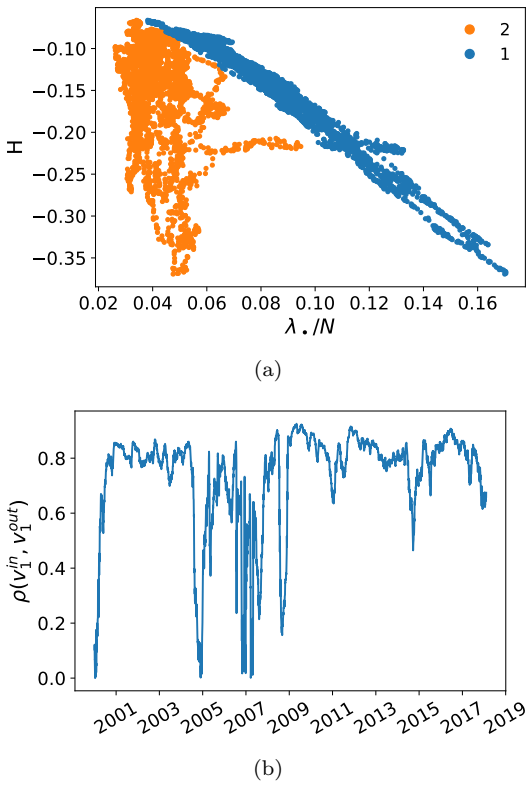


FIG. 4. (a) Scatter plot of H as a function of the fraction of variance explained by the largest and the second largest eigenvalues; (b) Pearson correlation between the eigenvector corresponding to the largest eigenvalue in the in-sample time window $\mathbf{v}_1^{in}$ and in the out-of-sample time window $\mathbf{v}_1^{out}$. $T_{in} = T_{out} = 100$

relation matrices are positive-defined, $\lambda_i \geq 0$ for every i. When $q = T/N < 1$, there are $N - T$ zero eigenvalues and the rank of the matrix is T, hence smaller than N. Finally, $\sum_i \lambda_i = N$. We shall adopt the convention that eigenvalues are sorted, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$.

This decomposition yields interesting insights on triads: indeed, each component of the spectral decomposition $\mathbf{v}_i' \mathbf{v}_i$ is a matrix which only contains stable triangles. This means that the signs of its elements label the groups to which they belong. Indeed, only two possible scenarios can occur:

1. **paradisiac case**: $\mathbf{v}_i$ has only positive (or negative) components, in which case $\mathbf{v}_i' \mathbf{v}_i$ has only positive entries;

2. **bi-polar case**: some components of $\mathbf{v}_i$ are positive and others are negative, in which case the matrix $\mathbf{v}_i' \mathbf{v}_i$ is composed of two groups.

Therefore, if the largest eigenvalue is much larger than the other ones, i.e. $\lambda_1 \gg 1/N$, the $\mathbf{v}_1' \mathbf{v}_1$ matrix dominates in Eq. (2). Reversely, if $\lambda_1 \approx 1/N$, the contribution to the stability of each component $\mathbf{v}_i' \mathbf{v}_i$ may cancel out each other, leading in most of the cases to lower global stability.

The components of the first eigenvector of the correlation matrix C (of real returns r) typically have the same sign because the market mode is still present; however, since we are interested in the macroscopic cluster composition of the stocks, we removed the market mode defined as the median return when computing Φ. This means that the largest eigenvector $\mathbf{v}_1$ of Φ will be composed of positive and negative entries, and it will be strongly correlated with the second largest eigenvector of the correlation matrix C. As was pointed out by [36], the dynamics of the second eigenvalue of C is quite independent from that of the first one, and the direction of the related eigenvector is quite stable over time. In fact, in Fig. 4(a), we show that the fraction of variance explained by the largest eigenvalue of Φ, λ_1/N, is strongly anti-correlated with H ($R^2 = 0.94$), whereas λ_2/N is only weakly anti-correlated with H. Furthermore, we show in Fig. 4(b) that the direction of $\mathbf{v}_1$ is quite stable across the time periods.

The main implication of this observation is that, if the largest eigenvalue λ_1 (Φ) increases and if the direction of $\mathbf{v}_1$ does not change substantially, then the stability of some the triads increases. This is when one can predict the sign of some correlations according to Heider's balance theory.

D. Prediction of correlation signs

As observed in the previous section, triads suggest a mechanism to predict the future states of S, i.e., of the signs of correlation coefficients. For this purpose we define Δ_{ij} as the contribution of asset pair (i, j) to H, which amounts to (in matrix notation)

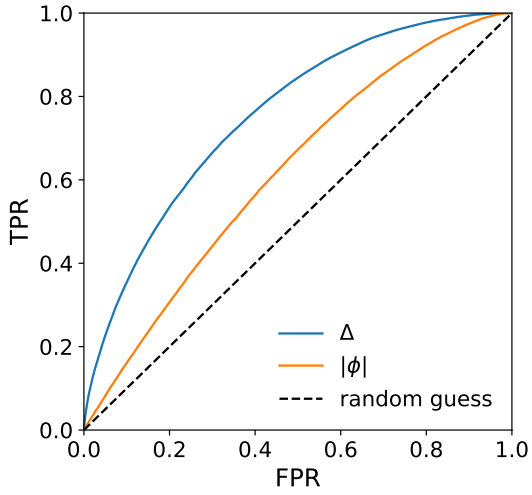
$$\Delta = \frac{S \circ S^2}{N-2}, \tag{3}$$

where S is the signed adjacency matrix defined above and $\circ$ is the Hadamard (element-wise) product. Note that $\Delta_{ij} \in [-1, +1]$: $\Delta_{ij} = +1$ if the link (i, j) forms stable pairs with all the other nodes, and $\Delta_{ij} = -1$ if the link (i, j) forms unstable pairs with all the other nodes.

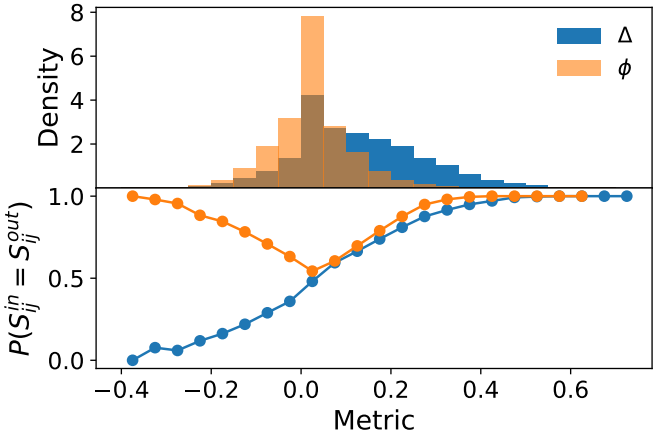
Our hypothesis is that the lower Δ_{ij} in the in-sample window, the higher the probability for the link (i, j) to switch its sign in the future, and reversely: high values of Δ_{ij} should be related to high out-of-sample stability interactions.

In order to test this hypothesis, we build a binary classifier that uses Δ_{ij} as discrimination variable. Specifically, we evaluate Δ_{ij} in an in-sample time window of T_{in} days, and we try to predict the sign switch of Φ_{ij} in the out-of-sample of T_{out}. In our experiments the in-sample and out-of-sample time windows are not overlapping and in general $T_{in} \neq T_{out}$. In order to assess the ability of Δ_{ij} to predict the sign stability, we used the Receiver Operating Characteristic (ROC) curve ([37]), a graphical representation of the True Positive Rate as a function of the False Positive rate as the discrimination threshold varies. As a summary of the performance of a discrimination variable, we use the Area Under the Curve (AUC). We therefore compare the ROC curves obtained with Δ_{ij} as discrimination variable, and the other associated with the value of the correlation Φ_{ij}.

Intuitively, larger correlations (in absolute value) should be more stable than smaller ones, if only because of estimation noise. In the high-dimensional case however, stability depends more on triadic relationships that on the intensity of correlations. In Fig. 5(a), we show an example of the ROC evaluated on 2011-18-04 with $T_{in} = T_{out} = 155$. The variable Δ_{ij} outperforms $|\Phi_{ij}|$ for the prediction of the correlation sign: their respective AUCs are 0.75 and 0.61. The origin of this difference clearly appears in Fig. 5(b) which plots the probability that both in- and out-of-sample signs are the equal as a function of the discrimination parameter. The first obvious observation is that $|\Phi|$ is not able to predict the sign changes as $P(S_{ij}^{in} = S_{ij}^{out})$ is never smaller than 0.5. Furthermore, by looking at the marginal distributions, it is clear that most of the correlations lie close to $|\Phi| = 0$; this explains why $|\Phi|$ is only slightly more informative than a coin toss. On the other hand, the marginal distribution of Δ has better coverage, resulting in a better correction sign prediction performance. This result also holds for the correlation coefficients of raw returns C.



(a)

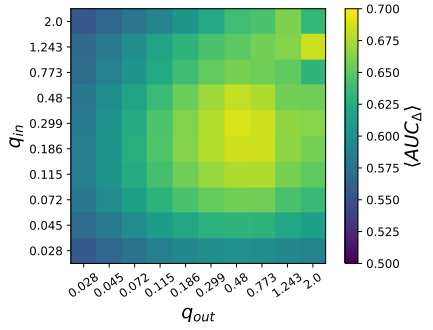


(b)

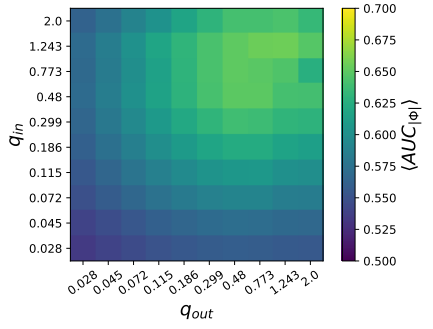
FIG. 5. (*a*) ROC curve for 2011-18-04 with $T_{in} = T_{out} = 155$; (*b*) the lower subplot is the probability to preserve the in-sample sign in the out-of-sample on 2011-18-04 for different values of the discrimination parameter binned in steps of 0.05, the upper subplot is the related marginal distribution.

We computed the performance of both predictors for a wide range of calibration and test window lengths chosen in order to include partial-rank and full-rank correlation matrices, i.e., 10 values between 20 and 2000 with a geometric progression. For each pair (T_{in}, T_{out}), we estimate the AUC of each method in rolling windows with a step of 1 day. Figure 6(c) shows the difference $\langle \mathrm{AUC}_\Delta \rangle - \langle \mathrm{AUC}_{|\Phi|} \rangle$, where $\langle \mathrm{AUC}_X \rangle$ is the average value of the AUC in the considered time-period for predictor $X \in \{\Delta, |\Phi|\}$.

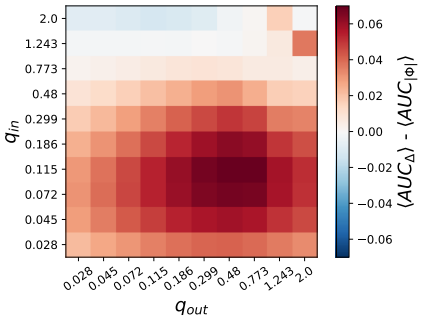
The dependence of the AUC as a function of q_{in} and q_{out} is worth discussing: first, $\mathrm{AUC}_{|\Phi|}$ increases monotonically as a function of both q_{in} and q_{out}; second, AUC_Δ has a local maximum at about $(q_{in}, q_{out}) \simeq (0.1, 0.5)$, i.e., deep in the high-dimensional regime. The difference between the two is clear: triads are better as long as $q > 1$



(a)



(b)

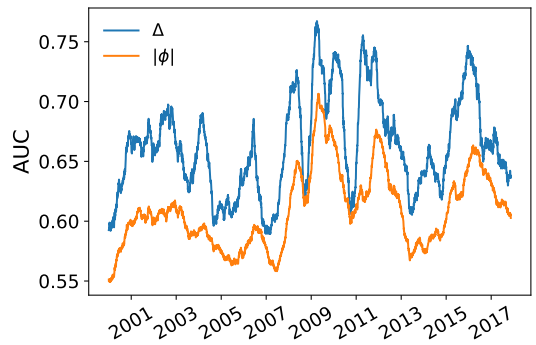


(c)

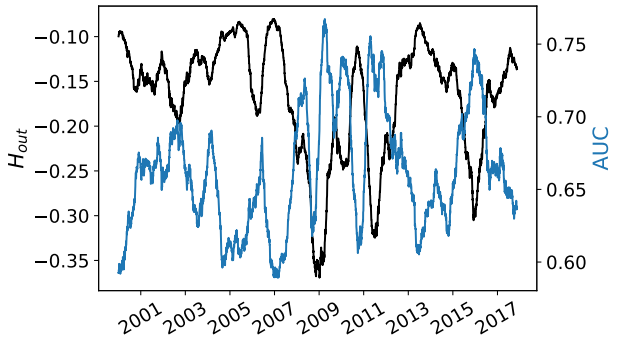
FIG. 6. (*a*) Heatmap of the difference between the average AUC of the two discrimination variables Δ and $|\Phi|$; (*b*) heatmap of the average AUC_Δ; (*c*) heatmap of the average $\text{AUC}_|\Phi|$.

and correlations better when $q < 1$. The same results hold for Hong Kong equities data (see appendix V C).

Figure 7(a) shows the evolution of the AUC for $T_{in} = T_{out} = 155$. Although Δ outperforms $|\Phi|$ most of the time, the difference between the methods is not constant. On 2008-26-06 for example, both AUCs are almost equal. Fig. 7(b) illustrates the strong anti-correlation between AUC_Δ and the out-of-sample H. Specifically, the two variables have a Pearson correlation of -0.77 and a Spearman correlation of -0.83, with a p-value close to 0. In fact, we must consider that when H increases, the total number of stable pairs decrease, and reversely. In any case, even in the worse situation, the variable Δ performs as well as Φ.



(a)



(b)

FIG. 7. (*a*) Evolution of AUC for the two model; (*b*) AUC (blue line) and H evaluated on the out-of-sample as a function of time. Both panels refer are evaluated with $T_{in} = T_{out} = 155$

III. DISCUSSION

In the high-dimensional regime, correlation matrices become pathologically noisy and the strength of their coefficients are not the best predictors of their stability. Accounting for more complex relationships between correlations makes it possible to predict the sign change of correlation coefficients deep in this regime. More precisely, dyadic relationships are better predicted from triadic relationships, as higher-order nonparametric structures exploit non-obvious structure of high-dimensional correlation matrices.

Potential applications of our method includes building better portfolios by accounting for predicted correlation sign changes, which will be addressed in a future work.

This publication stems from a partnership between CentraleSupélec and BNP Paribas.

IV. METHODS

A. Statistically Validated Networks

A binarized version of the return matrix $\mathbf{b} \in N \times T$ can be interpreted as a bipartite network. A bipartite

network is a particular network where the nodes belong to two different sets, in our case one set is composed by the N stocks and the other set by the T days. Only links among nodes of different sets are allowed. Specifically, a stock is linked to a day if its return is larger than the median return of all the stocks in such day. A typical approach to study bipartite networks is to project them into monopartite networks. A projected network is a network composed by nodes of only one set, and a link among to node is established only if those nodes share at least one common neighbors in the opposite set. However, this linkage rule is too permissive and typically leads to a very dense projected network; therefore, the resulting interaction topology could be in many cases meaningless. An alternative approach, defined in [31], is to link two nodes of the same set if the number of common neighbours they share in the opposite set cannot be explained by random chance. Specifically, one computes a p-value for each link according to the cumulative hypergeometric distribution:

$$\pi_{ij} = 1 - \sum_{x=0}^{c_{ij}-1} \frac{\binom{k_j}{k_i}\binom{T-k_j}{x-k_i}}{\binom{n}{x}}, \qquad (4)$$

where c_{ij} is the number of common neighbours of (i, j), k_i and k_j are the degree of the nodes i and j in the bipartite network respectively, and T is the number of nodes of the opposite set. Since the test is performed on every link of the projected network, a multiple-comparison correction is required to control the fraction false positive discoveries; in this work we use the False Discovery Rate (FDR) which guarantees that the proportion of false discovery is strictly less than α.

It is worth noticing that the number of common neighbours of two nodes can be evaluated with the scalar product $c_{ij} = \sum_t b_{it} b_{jt}$, and the expected number of common neighbours according with the hypergeometric distribution is $E[c_{ij}] = \frac{k_i k_j}{T} = \frac{\sum_t b_{it} \sum_t b_{jt}}{T}$. Therefore, the condition behind the statistical test can be translated into a condition of positivity of the correlation coefficient ϕ_{ij}.

[1] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[2] Richard O Michaud. The Markowitz optimization enigma: Is optimizedoptimal? *Financial Analysts Journal*, 45(1):31–42, 1989.

[3] Olivier Ledoit and Michael Wolf. Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.

[4] Matteo Marsili et al. Dissecting financial markets: sectors and states. *Quantitative Finance*, 2(4):297–302, 2002.

[5] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. Noise dressing of financial correlation matrices. *Physical review letters*, 83(7):1467, 1999.

[6] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luis A Nunes Amaral, Thomas Guhr, and H Eugene Stanley. Random matrix approach to cross correlations in financial data. *Physical Review E*, 65(6):066126, 2002.

[7] Joël Bun, Romain Allez, Jean-Philippe Bouchaud, and Marc Potters. Rotational invariant estimator for general noisy matrices. *IEEE Transactions on Information Theory*, 62(12):7475–7490, 2016.

[8] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

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

[10] Tomaso Aste, Tiziana Di Matteo, and ST Hyde. Complex networks on hyperbolic surfaces. *Physica A: Statistical Mechanics and its Applications*, 346(1-2):20–26, 2005.

[11] Michele Tumminello, Tomaso Aste, Tiziana Di Matteo, and Rosario N Mantegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426, 2005.

[12] J-P Onnela, Kimmo Kaski, and Janos Kertész. Clustering and information in correlation based financial networks. *The European Physical Journal B*, 38(2):353–362, 2004.

[13] Fritz Heider. Attitudes and cognitive organization. *The Journal of psychology*, 21(1):107–112, 1946.

[14] Dorwin Cartwright and Frank Harary. Structural balance: a generalization of heider's theory. *Psychological review*, 63(5):277, 1956.

[15] Anatol Rapoport. Mathematical models of social interaction. In D. Luce, editor, *Handbook of Mathematical Psychology*, pages 2–493. John Wiley & Sons, 1963.

[16] Claudio Altafini. Dynamics of opinion forming in structurally balanced social networks. *PloS one*, 7(6):e38135, 2012.

[17] Tibor Antal, Paul L Krapivsky, and Sidney Redner. Social balance on networks: The dynamics of friendship and enmity. *Physica D: Nonlinear Phenomena*, 224(1-2):130–136, 2006.

[18] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1361–1370. ACM, 2010.

[19] Ginestra Bianconi, Richard K Darst, Jacopo Iacovacci, and Santo Fortunato. Triadic closure as a basic generating mechanism of communities in complex networks. *Physical Review E*, 90(4):042806, 2014.

[20] L Hedayatifar, F Hassanibesheli, AH Shirazi, S Vasheghani Farahani, and GR Jafari. Pseudo paths towards minimum energy states in network dynamics. *Physica A: Statistical Mechanics and its Applications*, 483:109–116, 2017.

[21] Jukka-Pekka Onnela, Jari Saramäki, János Kertész, and Kimmo Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103, 2005.

[22] Tomaso Aste, W Shaw, and Tiziana Di Matteo. Correlation structure and dynamics in volatile markets. *New Journal of Physics*, 12(8):085009, 2010.

[23] Andrew G Haldane and Robert M May. Systemic risk in banking ecosystems. *Nature*, 469(7330):351, 2011.

[24] Marco Bardoscia, Stefano Battiston, Fabio Caccioli, and Guido Caldarelli. Pathways towards instability in financial networks. *Nature Communications*, 8:14416, 2017.

[25] Christian Borghesi, Matteo Marsili, and Salvatore Micciche. Emergence of time-horizon invariant correlation structure in financial returns by subtraction of the market mode. *Physical Review E*, 76(2):026104, 2007.

[26] Daniel J Ozer. Correlation and the coefficient of determination. *Psychological bulletin*, 97(2):307, 1985.

[27] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[28] Assaf Almog and Diego Garlaschelli. Binary versus non-binary information in real time series: empirical results and maximum-entropy matrix models. *New journal of physics*, 16(9):093015, 2014.

[29] Assaf Almog, Ferry Besamusca, Mel MacMahon, and Diego Garlaschelli. Mesoscopic community structure of financial markets revealed by price and sign fluctuations. *PloS one*, 10(7):e0133679, 2015.

[30] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, 2006.

[31] Michele Tumminello, Salvatore Micciche, Fabrizio Lillo, Jyrki Piilo, and Rosario N Mantegna. Statistically validated networks in bipartite complex systems. *PloS one*, 6(3):e17994, 2011.

[32] T Tony Cai and Weidong Liu. Large-scale multiple testing of correlations. *Journal of the American Statistical Association*, 111(513):229–240, 2016.

[33] Pierre-Alain Reigneron, Vincent Nguyen, Stefano Ciliberti, Philip Seager, and Jean-Philippe Bouchaud. The case for long-only agnostic allocation portfolios. *arXiv preprint arXiv:1906.05187*, 2019.

[34] Mark Newman. *Networks: an introduction*, page 225. Oxford university press, 2010.

[35] Christian Bongiorno, András London, Salvatore Micciché, and Rosario N Mantegna. Core of communities in bipartite networks. *Physical Review E*, 96(2):022321, 2017.

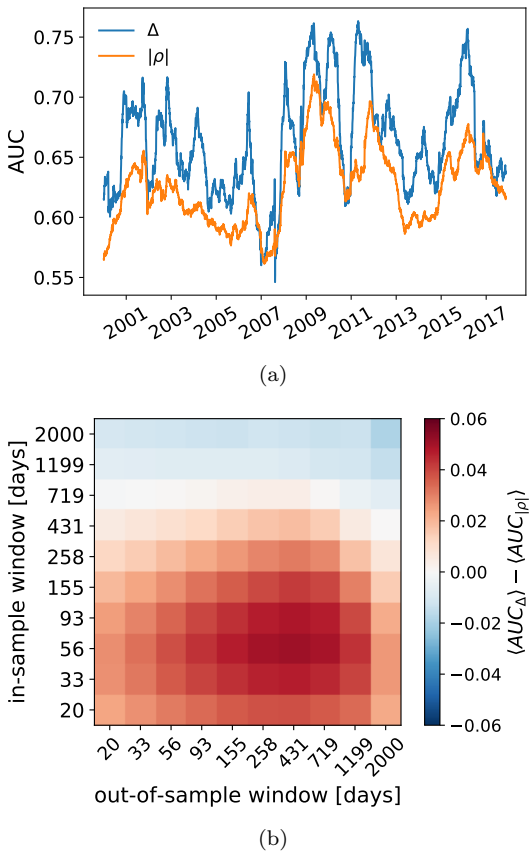
[36] Giuseppe Buccheri, Stefano Marmi, and Rosario N Mantegna. Evolution of correlation structure of industrial indices of us equity markets. *Physical Review E*, 88(1):012806, 2013.

[37] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

V. APPENDIX

A. Pearson correlations

In order to prove that our results are qualitatively independent from the binarization of the returns, we performed the same analysis from the sign of the Pearson correlation ρ_{ij} computed from partial log-returns. Figure 8 confirms that unstable triads are better predictors of correlation sign changes than correlation values themselves.



(a)

(b)

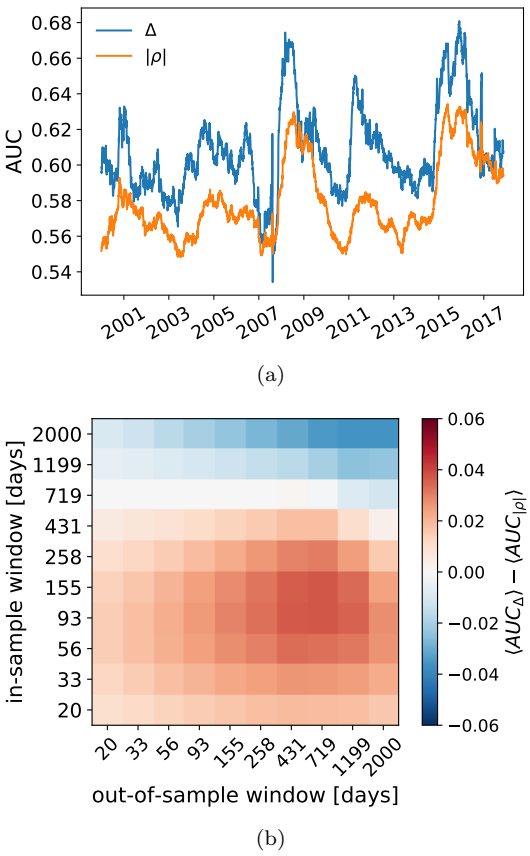
FIG. 8. (*a*) Evolution of AUC for the two model for $T_{in} = T_{out} = 155$; (*b*) Average difference of the AUC among the two model for different in-sample and out-of-sample time windows; Both panels refer to the Pearson correlation matrix among the returns minus the median **z**

B. Market mode removal: first eigenvalue

We show that our results are qualitatively stable if we remove the market mode in a different way. In this section we studied the sign of the partial Pearson correlation matrix obtained with the following equation:

$$\boldsymbol{\rho}^{(p)} = \sum_{i=2}^{N} \lambda_i \mathbf{v}_i' \, \mathbf{v}_i \tag{5}$$

where the eigenvalues and the eigenvectors are computed on the Pearson correlation matrix $\boldsymbol{\rho}$ among the original return matrix $\mathbf{r}$. As depicted in Fig. 9, results are similar to those obtained by removing the median returns; however, the nonparametric nature of the estimate is lost.



(a)

(b)

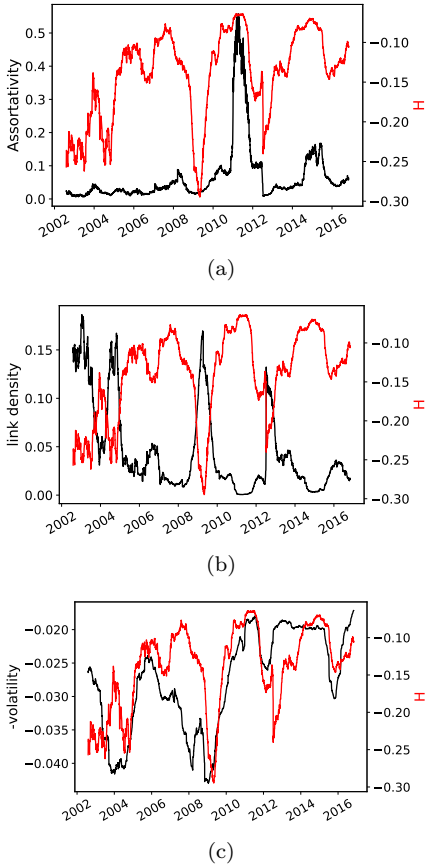
FIG. 9. (*a*) Evolution of AUC for the two model for $T_{in} = T_{out} = 155$; (*b*) Average difference of the AUC among the two model for different in-sample and out-of-sample time windows; Both panels refer to the partial Pearson correlation matrix $\boldsymbol{\rho}^{(p)}$ among the returns minus the median $\mathbf{r}$

C. Hong Kong Stock Exchange

In this section we repeated the analysis for the Hong Kong stock exchange. We build a binary return matrix $\mathbf{b}$ and the related phi matrix $\boldsymbol{\phi}$ according with the procedure illustrated in section II A.

In contrast with US equities, we do not observe a strong correlation between the sector partition and the links of the SVN, as shown in Fig. 10(a). In fact, the assortativity is very close to the random null expectation for most of the time-period. We want to stress that this does not necessarily mean that the stocks are not organized in clusters. In fact, as for US equities, H varies over time, and it is strongly correlated with the number of links of the SVN (Fig. 10(b)) and with the volatility (Fig. 10(c)). To our knowledge, the sector structure (or apparent lack thereof) of Honk Kong equities has not reported elsewhere.

In this dataset as well, unstable triads are significantly better than the absolute value of the $|\phi|$ at predicting the instability of correlation signs in the high-dimensional regime (Fig. 11(a)).



(a)

(b)

(c)

FIG. 10. Hong Kong stock exchange: (*a*) assortativity of the SVN with respect to the sector partition and H; Number links of the SVN and H; (*c*) minus Volatility and H. $T = 100$ days.

Once again, dynamical evolution of the AUC (Fig. 11(b)) is minimal in the proximity of a minima of H_{out}.

D. Negative Links

We report here the SVNs determined between binarized returns of opposite signs. Three observations stand out: i) they are mostly empty; ii) they are more likely to become non-empty in times of large volatility iii) they are generally disassortative.

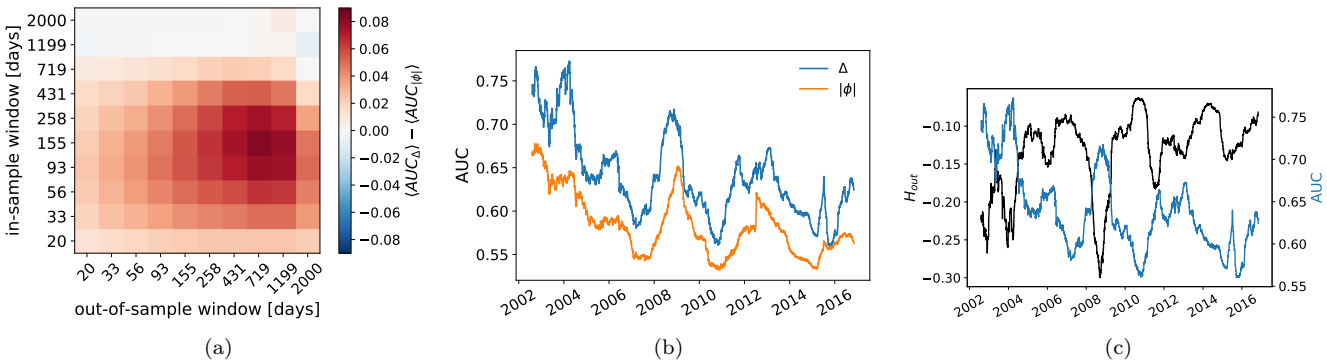


FIG. 11. Hong Kong stock exchange: (a) average difference of the AUC among the two model for different in-sample and out-of-sample time windows; (b) Evolution of AUC for the two model for $T_{in} = T_{out} = 155$; (c) AUC and H in the out-of-sample as a function of time for $T_{in} = T_{out} = 155$. All panels refer to the Hong Kong Stock Market.

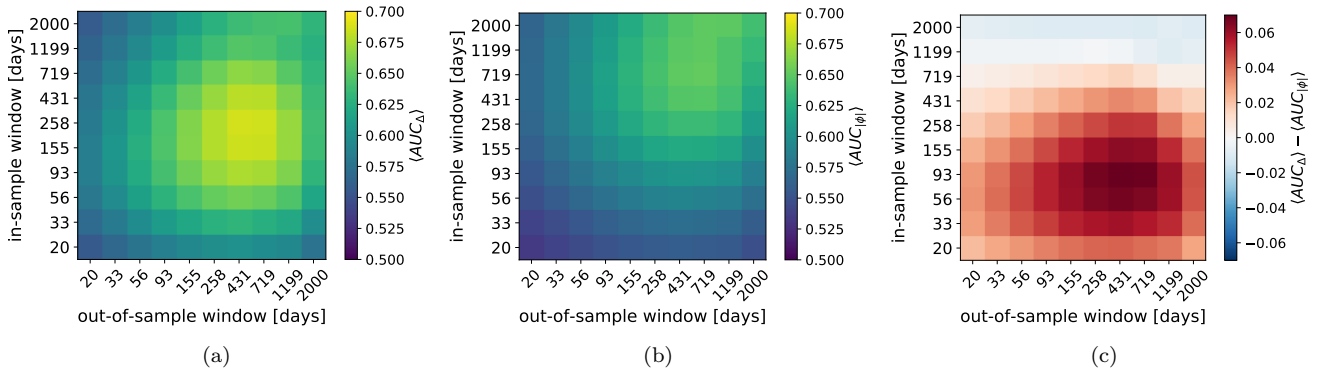


FIG. 12. (a) Heatmap of the difference between the average AUC of the two discrimination variables Δ and $|\phi|$; (b) heatmap of the average AUC_Δ; (c) heatmap of the average $\mathrm{AUC}_{|\phi|}$.

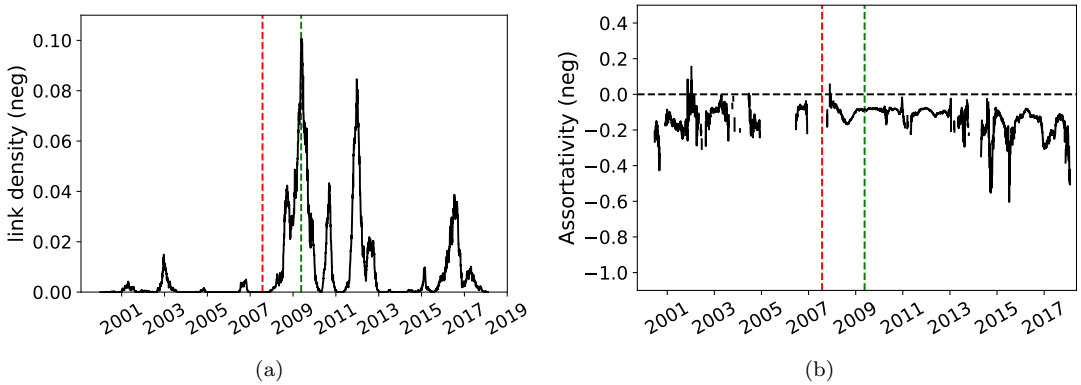


FIG. 13. (a) link density of significant negative Φ; (b) Assortativity of of significant negative Φ with respect to the sector categorization, only networks with at least 10 links are considered.