Quality & Quantity
https://doi.org/10.1007/s11135-023-01746-0

ORIGINAL PAPER

An empirical comparison of correlation-based systemic risk measures

Caterina Pastorino[1] · Pierpaolo Uberti[2]

Accepted: 5 September 2023

Abstract

Despite the growing attention in the last years on the topic of systemic risk, a widely accepted definition of systemic crisis is missing. We use a theoretical scheme to subjectively define a systemic event. This permits the analysis of a financial crisis as a standard binary classification problem, providing an intuitive and useful framework to compare systemic risk measures defined in very different fields. Then we focus the empirical analysis on the comparison of the performance of correlation-based systemic risk measures using the standard tools for the evaluation of binary classifiers as the receiver operating characteristic (ROC) curve and the area under the curve (AUC). We show that the binary classification framework is useful but unable to capture some significant differences among the measures under comparison. The experimental approach, developed on real financial data, is divided in an in-sample exercise, able to evaluate the descriptive power of the different systemic risk measures, and an out-of-sample application to evaluate the capacity of the measures in preventing and predicting systemic events. The forecasting ability of a measure can be fundamental for policy makers and investors respectively to stabilize market fluctuations and to reduce the losses.

Keywords Systemic risk · Financial crises · Correlation-based systemic risk measures · Binary classification

1 Introduction

In recent years, starting from the 2008 subprime global financial crisis, economic crises and financial distress have become more frequent and severe, impacting the stability of the whole economic system. In addition, previous crises have shaken the world economy and illustrated the importance of systemic risk. As a result of this sense of uncertainty and instability, a current broad stream of research has addressed the notion of systemic risk.

✉ Caterina Pastorino
c.pastorino@campus.unimib.it

Pierpaolo Uberti
pierpaolo.uberti@unimib.it

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

[2] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy

Published online: 23 September 2023

The enormous social and political cost of this type of risk requires the design and implementation of specific tools to prevent financial crises.

At the state of the art, a formal widely accepted definition of systemic crisis is missing in the literature. Consequently, systemic risk, as the risk of a systemic event to occur, is a concept that remains vague and elusive. On one hand, the absence of a definition could be a limitation to the development of a unified theoretical framework. On the other hand, the size and the complexity of the financial system suggest that systemic risk assessment must be approached from a wide variety of perspectives, implying that more than one risk measure should be necessary. Therefore, a single consensus on a measure of systemic risk may neither be possible nor desirable; for a comprehensive review of the systemic risk measures proposed in the literature see Sect. 2.

Generally speaking, a systemic event occurs when many market participants are simultaneously affected by severe losses, which then spread through the whole system impacting its stability. Thus, systemic risk can be observed as a set of events that threaten the stability or public confidence in the financial system. The European Central Bank, see Hartmann et al. (2009), page 134, associates systemic risk to the concept of financial instability "so widespread that it impairs the functioning of a financial system to the point where economic growth and welfare suffer materially". The goal of our research is to present an empirical comparison of different systemic risk measures, showing that the standard binary classification model, opportunely adapted to the present context, is a useful framework to compare different measures. Through the application on correlation-based systemic risk measure, the contribution of our paper is to identify a class of measures that over-perform the other measures both in a descriptive and in a predictive context.

Aiming at preserving the variegate universe of systemic risk measures proposed in the literature, we start our analysis from a very natural definition of systemic crisis by fixing its duration and severity in terms of average loss on a given period. This provides a simple unified framework, the standard binary classification model, where it is possible to compare different systemic risk measures, independently from the economic variables and the models that are used to calculate them. The independent variable is represented by the systemic risk measure while the binary dependent variable becomes the occurrence/not occurrence of a systemic crisis. It is then natural to use the ROC curve and the AUC as simple and intuitive tools to evaluate the performance of the alternative measures, see Hand and Till (2001) for a detailed review. We also show that some important differences among the measures remain hidden when working in the standard classification problem, suggesting that the theoretical framework is too simple to highlight some fundamental details. Moreover, since one main issue regarding systemic risk measures is their ability to prevent future crises, we show how to adapt the proposed framework to study the forecasting power of the different measures. This permits identifying which systemic risk measures could be suitable to be used by policy makers and investors as early warning indicators to prevent instability and losses in global financial markets.

The main focus of the present research is on the empirical comparison of the performance of correlation-based systemic risk measures. Starting from the claim "correlation-based measures are widespread, yet they measure only pairwise association and are largely wed to linear, Gaussian thinking, making them of limited value in financial-market contexts", see Diebold and Yılmaz (2014), page 119, we argue three main results. First, we show how the correlation-based measures under comparison work fine in a descriptive context, with limited but significant differences among each other. Second, we illustrate that elementary correlation-based measures like the average correlation have a poor predictive power, resulting in substantially limited help as early warning indicators. Third, we highlight through empirical evidence that more sophisticated correlation-based systemic risk measures show interesting performance also in the out-of-sample forecasting framework. For example, the measures that depend on the eigenvalues of the correlation matrix, as the family proposed in Maggi et al. (2020), are able to overcome the limitation on the pairwise structure of the correlation, controlling linear dependency among each possible portfolio created starting from the original variables. In order to test the robustness of our findings, we perform a sensitivity analysis varying the parameters of duration and severity that define the systemic event and we use two different financial indexes, the S &P500 and the Eurostoxx600, to approximate the whole economic system. No significant differences in the results are obtained when the settings of the experiment vary, supporting that our findings do not mainly depend on arbitrary choices.

The paper is organized as follows: Sect. 2 contains a comprehensive review of the recent literature on systemic risk measurement, Sect. 3 discusses the methodological proposal to compare systemic risk measures, Sect. 4 provides the empirical experiments and it is divided into three sub-Sections with respectively the data description and the definitions of the risk measures under comparison, the in-sample and the out-of-sample applications; Sect. 5 concludes the paper while the appendix provides a further empirical example performed on a different database to testify the reproducibility of the results.

2 Literature review

Although systemic risk is universally recognised as a threat to financial stability, providing a unique definition is hard. A robust framework for monitoring and managing financial stability should be able, at the same time, to incorporate many different perspectives and to adapt systemic risk measures to the ongoing evolution of the financial system. Since these features are relevant both for policymakers and speculators, academic research can give an important contribution to the understanding of the concept of systemic risk. Forecasting ability is one of the most important requirements a systemic risk measures should have, because financial crises have been one of the major causes of economic distress. Of course, an accurate prediction of crises through the study of predictive indicators of systemic risk may allow the management of market losses. In the economic literature many survey papers collect the multitude of systemic risk measures and related conceptual frameworks that have been proposed over the past several years. These papers enumerate and classify the indicators of systemic risk according to some convenient criteria predetermined by the authors. For example, Rodríguez-Moreno and Peña (2013) estimate and compare two groups of high-frequency market-based systemic risk measures called macro and micro. The measures that belong to the first group provide information about the financial sector, while the measures in the second group depend on the information from individual institutions. Silva et al. (2017) present an analysis of the literature on systemic risk by ranking 266 articles which were published no later than September 2016; this approach makes it possible to identify gaps in the literature on systemic risk and to select the most influential articles in the field. In Bisias et al. (2012) a selection of 31 quantitative measures of systemic risk studied in the economic literature are listed. This classification of systemic risk measures considers several taxonomies: the first taxonomy deals with data requirements, the second looks at supervisory scopes, which is of particular interest for policy makers while the third considers what could be easier for researchers to use, allowing them to quickly identify common themes, algorithms and data structures within each category.

In the following we provide a comprehensive enumeration of the most important studies on systemic risk published in the last ten years, grouping them for the similarities in the approaches, with the objective of depicting the variegate and intricate universe of systemic risk measures.

A part of the current literature on systemic risk concerns macroeconomic models of systemic risk; for example, Giglio et al. (2016) study how systemic risk and financial market distress affect the distribution of shocks to real economic activity. They develop a new systemic risk index based on an out-of-sample predictive quantile regression approach. Duca and Peltonen (2013) propose a financial stress index to identify the onset date of systemic financial crises; they also suggest a model that combines both domestic and global indicators of macro-financial vulnerability to predict systemic financial crises.

Since the aggregate measurement of risks and imbalances does not capture everything, another branch of literature considers the analysis of contagion and the spread of a potential shock through the system. A large body of analysis is based on granular foundations and network measurements: Acemoglu et al. (2015) and Elliott et al. (2014) develop studies on network analysis and systemic financial linkages, Mezei and Sarlin (2016) build a network analysis that exploits the so called fuzzy cognitive map. In order to estimate systemic risk with graphical network models, Cerchiello and Giudici (2017) propose a framework based on two different sources, financial markets and financial tweets, and suggests a way to combine them, using a Bayesian approach. Starting from the concept of connectedness, some papers study the phenomenon from the point of view of network connectedness, see Diebold and Yılmaz (2014) and Demirer et al. (2018). Various network-based approaches have been proposed to analyze the contribution of financial firms to systemic risk, given the network interdependence among firms' tail risk exposures, see Hautsch et al. (2015) and Betz et al. (2016). Härdle et al. (2016) derive an approach that allow to rank the systemic risk receivers and systemic risk emitters in the US financial market named Tail Event driven network (TENET). The TENET approach is also develop in Wang et al. (2018). Billio et al. (2012) suggest two econometric measures of systemic risk that capture the interconnectedness among the monthly returns of hedge funds, banks, brokers, and insurance companies based on principal components analysis (PCA) and Granger causality tests. A similar approach has also been studied by Zheng et al. (2012) and Zhang and Broadstock (2020).

A further line of research evaluates systemic risk using of prospective measures, see among the others (Allen et al. 2012). However, studies exploiting PCA can also be developed following a prospective approach, see Zheng et al. (2012).

Furthermore, Diebold and Yilmaz (2009) provide a simple and intuitive measure of interdependence of asset returns and/or their volatility. The authors formulate a quantitative measure of such interdependence, referred to as a spillover index. This background is also developed using a generalized vector autoregressive model in which forecast-error variance decompositions are invariant with respect to the ordering of the variables. Therefore, Diebold and Yilmaz (2012) proposed measures of both total and directional volatility spillovers.

In terms of systemic risk monitoring, stress tests are useful as a special case of forward-looking analysis. Stress testing is codified in regulation and international standards, including the Basel accord, see Pederzoli and Torricelli (2017).

Following Bisias' classification, a complementary philosophy to the predictive measures is the cross-sectional measure: this approach aims at examine the co-dependence of institutions on each other's "health". Based on the Adrian and Brunnermeier (2011) analysis, many researches focus on conditional value at risk as a measure of systemic risk (Exposure CoVaR), see Laeven et al. (2016), Bernal et al. (2014) and López-Espinosa et al. (2012);

Lopez-Espinosa et al. (2015). The CoVaR approach can also be developed through a multivariate GARCH analysis, see Girardi and Ergün (2013). Reboredo and Ugolini (2015) assess systemic risk in European sovereign debt markets before and after the onset of the Greek debt crisis by taking conditional value at risk (CoVaR), characterized and calculated using copulas. Moreover, Bierth et al. (2015) implement a method based on CoVaR and panel regression. To evaluate the systemic risk in a cross-sectional dimension (Black et al. 2016) calculate a distress insurance premium which integrates the characteristics of bank size, probability of default, and correlation. Unlike these papers, Acharya et al. (2017) have implemented market-based systemic distress indexes that consider a bank's expected capital shortfall conditioned to systemic events, named systemic expected shortfall (SES) and marginal expected shortfall (MES). Acharya et al. (2012) focus on firms' expected capital shortfall in the event of a crisis and is inspired by the SRISK measure defined as the capital a firm will need in the event of another financial crisis, see Brownlees and Engle (2017). This approach was also investigated in Engle et al. (2015).

In order to measure the systemic risk posed by individual institutions, Varotto and Zhao (2018) define a hybrid systemic risk indicator called rSYR. An interesting analysis is proposed by Sedunov (2016), which compares the performance of three institution-level measures of systemic risk exposure: CoVaR, SES and Granger causality. Instead, Cai et al. (2018) develop a new measure of bank interconnectedness using syndicated corporate loan portfolios. Empirical results show that the interconnectedness is positively correlated with several bank-level measures of systemic risk, including systemic capital shortfall (SRISK), distressed insurance premium (DIP) and CoVaR. Finally, it is important to look at systemic risk as measures of illiquidity and insolvency, see López-Espinosa et al. (2013).

Given this long and probably incomplete list of possible approaches to systemic risk, it seems safe to assume that more than one risk measure is needed to capture the overall complex nature of the phenomenon.

3 The methodological approach

We propose to use a comprehensive global financial index as a valuable proxy of the whole economic system. Operationally, given a time frame and a window length w, the dependent variable ${}^{w}y_t$ at time t is the average return of the index on the period from $t - w + 1$ to t, i.e. ${}^{w}y_t = \frac{1}{w}\sum_{i=t-w+1}^{t} y_i$. The measure of systemic risk at time t is indicated with m_t. Without loss of generality, we assume that the measure m_t is normalized, $0 \leq m_t \leq 1\ \forall t$; this assumption permits to directly interpret the value of the systemic risk measure as the probability of a systemic event to occur. In the present framework, a systemic event is defined as follows.

Definition 1 (*Systemic event*) Given a positive integer w and a threshold loss l, a systemic event on the period from $t - w + 1$ to t occurs when ${}^{w}y_t \leq l$.

In words, a systemic event occurs when the average return of the global referring index on a given period of length w is lower than the chosen threshold loss l. This definition is simple, intuitive and subjective, because it requires to set the duration of a systemic event, the amount of the average loss and the index that represents the whole economic system. Moreover, Definition 1 potentially transforms ${}^{w}y_t$ into a binary variable, permitting to use the standard techniques developed for the evaluation of binary classifiers. To highlight the

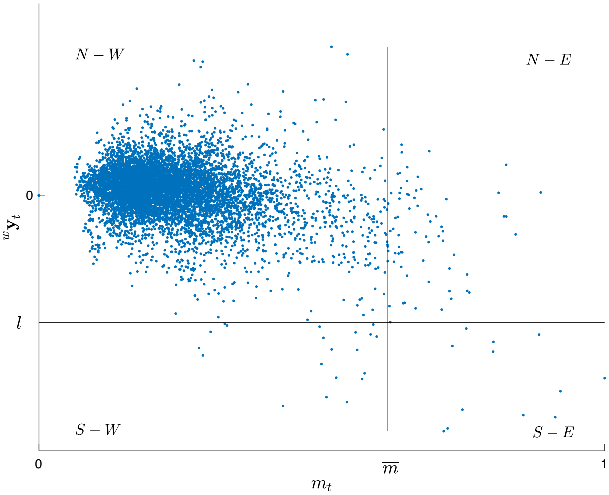


Fig. 1 A graphical example of a scatter plot for the points $(m_t, ^w y_t)$ and their classification

intuition, we plot in Fig. 1 the points $(m_t, ^w y_t)$ for some values of t, putting the dependent variable $^w y_t$ on the vertical axis and the systemic risk measure m_t on the horizontal axis. In Fig. 1 the horizontal line represents the threshold l given in Definition 1 while the vertical line is a threshold $\overline{m}$ on the probability of a systemic event.[1] The two lines divide the plane in four regions, called respectively N–W, N–E, S–W and S–E, that can immediately be related to the entries of the standard confusion matrix used in binary classification problems. While the entries of the confusion matrix are the absolute frequencies of the four possible results in a binary classification problem, in the present context, they correspond to the number of points that belongs to each of the four regions. The perfect classifications are given by the sum of the points in N–W and S–E, corresponding respectively to the case of low values of m_t and absence of a systemic event and high values of m_t and the occurrence of a systemic event. The values of m_t are considered high or low with respect to the choice of the threshold $\overline{m}$. The remaining two regions of the plane correspond to the classification errors or miss-classifications. N–E contains the so-called false positives: in this case the value of m_t is high but no systemic event occurs. The region S–W contains the false negatives, the situation in which the value of m_t is low while a systemic event occurs.

One peculiar difficulty in the description and prediction of systemic events, as it is clear from Fig. 1, is that a suitable measure of systemic risk should be able to overcome the issue that severe systemic events are, hopefully, extremely rare with respect to the periods where the economy and the markets behave normally. In general, the two alternatives of the classification problem are very unbalanced in terms of frequency. This phenomenon is usual in many binary classification applications; for example, in credit risk detection, see Figini and Uberti (2010), the number of clients of a bank not returning a loan is usually very limited with respect to the total number of clients. In the present framework, it is possible to partially overcome the structural problem of the unbalanced frequencies by changing the parameters l and w in Definition 1. One further peculiarity of binary classification models

[1] As in the standard classification framework, the dependent variable is a binary variable, in our case the presence/absence of a systemic event. The independent variable is a continuous variable, in this case the probability of a systemic event to occur. To binarize the probability a standard approach is to set a threshold value.

is the difference between the two classification errors. In the present case, the false negative is a severe error in terms of potential economic impact since it represents the situation in which the risk measure remains silent when a crisis occurs. On the opposite, the impact of a false positive is limited, generating only an opportunity cost: the risk measure provides a positive signal inducing to shed against a potential systemic event that does not occur.

To compare different risk measures in terms of discriminating capacities it is then natural to use classic tools as the ROC curve and the AUC. The ROC curve is obtained through the calculation of the confusion matrix for some given values of the threshold $\overline{m} \in [0, 1]$. This permits a global evaluation of the classifier avoiding an arbitrary choice of $\overline{m}$, see Hand and Till (2001). Given l, $\overline{m}$ and defining $Card(\cdot)$ as the function that counts the number of elements of a set, the confusion matrix can be written as:

$$\begin{bmatrix} Card\left\{(m_t,{}^w\mathbf{y}_t) : m_t \leq \overline{m},{}^w\mathbf{y}_t \geq l\right\} & Card\left\{(m_t,{}^w\mathbf{y}_t) : m_t > \overline{m},{}^w\mathbf{y}_t \geq l\right\} \\ Card\left\{(m_t,{}^w\mathbf{y}_t) : m_t \leq \overline{m},{}^w\mathbf{y}_t < l\right\} & Card\left\{(m_t,{}^w\mathbf{y}_t) : m_t > \overline{m},{}^w\mathbf{y}_t < l\right\} \end{bmatrix}.$$

Varying $\overline{m}$ and computing the correspondent confusion matrix is then possible to draw the ROC curve plotting the True positive rate against the False positive rate. Consequently, the AUC is calculated as the area under the ROC curve.

Finally, if we want to evaluate the forecasting power of a given systemic risk measure, it is then sufficient to replace the backward-looking dependent variable with its forward-looking version y_t^w, where $y_t^w = \frac{1}{w}\sum_{i=t+1}^{t+w} y_i$.

4 Empirical analysis

In this section we perform an empirical comparison to test the differences among alternative systemic risk measures, applying the methodological framework described in Sect. 3. The section is divided in three sub-sections: the first sub-section contains a brief data description and the enumeration of the systemic risk measures under comparison. The other two sub-sections provide respectively the in-sample and the out-of-sample experiments.

4.1 Data and measures

We choose to use the S &P500 as the comprehensive financial index able to represent the whole economic system.[2] The systemic event is then investigated on the base of the returns of the constituents of the S &P500 grouped in sectors. The implicit assumption is that during a systemic event high severity losses in each sector of the economic system occur simultaneously causing a global drop. The data set is composed by the time series of the daily returns from January 3, 1990 to February 23, 2021 of the S &P500 index and its sector sub-indexes: FINANCIALS, INFORMATION TECHNOLOGY, TELECOMMUNICATION SERVICES, HEALTH CARE, INDUSTRIALS,

[2] The choice of the S &P500 as a valuable proxy of the whole economic system is as standard as arbitrary. The economic system is unique, what changes is the proxy to represent it. Therefore we perform an analogous empirical experiment choosing the EUROSTOXX600 index as the approximation of the economic system to test the robustness of our findings. The results with respect to the EUROSTOXX600 are collected in the appendix.

CONSUMER DISCRETIONARY, ENERGY, CONSUMERS STAPLES, UTILITIES, MATERIALS.

Let A be the $T \times n$ matrix where each column A^i for $i = 1, \dots, n$ contains the time series of the normalized returns of the ith sector. In the present application $T = 8125$ and $n = 10$. The returns of the S &P500 are collected in the column vector y with $T = 8125$ observations. We refer to $\rho(A^i, A^j)$ as the Pearson correlation coefficient between A^i and A^j. The quantities $\sigma_1, \dots, \sigma_n$ are the n singular values of matrix A, taken in decreasing order.

We list in the following the definitions of the systemic risk measures under comparison.

- The *average correlation* (*AC*) is a real-valued function of the matrix A defined as

$$AC(A) = \frac{2}{n(n-1)} \sum_{i,j=1, i \neq j}^{n} |\rho(A^i, A^j)|, \tag{1}$$

- The *cumulative risk fraction* (*CRF*), see Billio et al. (2012), is a real-valued function of the matrix A defined as

$$CRF_k(A) = \frac{\sum_{j=1}^{k} \sigma_j^2(A)}{\sum_{j=1}^{n} \sigma_j^2(A)} \quad k = 1, \dots, n \tag{2}$$

- The *market rank indicator* (*MRI*), see Figini et al. (2020), is a real-valued function of the matrix A defined as

$$MRI_k(A) = \frac{\sigma_1(A)}{\left(\prod_{j=1}^{k} \sigma_{n-j+1}(A)\right)^{\frac{1}{k}}} \quad k = 1, \dots, n \tag{3}$$

- The *condition number* (*CN*), see Golub and Van Loan (1989), is a real-valued function of the matrix A defined as

$$CN(A) = \frac{\sigma_1(A)}{\sigma_n(A)} \tag{4}$$

- The *arithmetic rank indicator* (*ARI*), see Maggi et al. (2020), is a real-valued function of the matrix A defined as

$$ARI_k(A) = \frac{\sigma_1(A)}{\frac{1}{k}\sum_{j=1}^{k} \sigma_{n-j+1}(A)} \quad k = 1, \dots, n \tag{5}$$

- The *variance inflation factors* (*VIF*), see Belsley et al. (2005), of A are defined as

$$VIF_j(A) = \frac{1}{1 - R_j^2}, \quad j = 1, \dots, n \tag{6}$$

where R_j^2 is the coefficient of determination of the linear regression of A^j with respect to $\left\{A^i \mid i = 1, \dots, n, i \neq j\right\}$. The *maximum variance inflation factor* (*M-VIF*) is defined as

$$M\text{-}VIF(A) = \max\left\{VIF_1(A), \dots, VIF_n(A)\right\}. \tag{7}$$

- The *Mahalanobis distance* (d_M), see Mahalanobis (1936), is a real-valued function of the matrix A defined as

$$d_M(A) = \sqrt{A_{\$}(A'A)^{-1}A'_{\$}}, \tag{8}$$

where $A_{\$}$ is the last row of matrix A.

Remark 1 We note that, thank to the assumption on matrix A, in particular thank to the fact the columns of A are assumed to be standardized, all of the systemic risk measures listed above depend on the correlation between the sectors constituting the S &P500 index. In particular, the AC is a function of the correlation coefficients among the columns of A; the CRF as the MRI, the CN and the ARI are functions of the singular values of A, that are the square roots of the eigenvalues of $A'A$. In the present case $A'A$ is the correlation matrix. The VIF depends on the R^2 of the linear regressions between the columns of A; the R^2 directly depends on the correlation between the two variables under investigation since it is equal to 0 or 1 respectively when there is null correlation or a perfect linear correlation. The Mahalanobis distance is a Euclidean weighted distance that depends on the inverse of the covariance matrix; in the present case, as already highlighted, the covariance matrix coincides with the correlation matrix. For these reasons all the considered measures can be interpreted as correlation-based systemic risk measures.

Remark 2 The totality of the correlation-based systemic risk measures under comparison belong to the family of the proper measures of connectedness, as introduced in Maggi et al. (2020); we refer to that paper for the proofs of the theoretical properties of the measures.

Remark 3 Some of the correlation-based systemic risk measures listed above, the *MRI* and the *ARI*, depend on the value of an extra parameter k. Note that, for $k = 1$, $MRI = ARI = CN$. The parameter k measures how many dimensions of the space of the economic system represented by the columns of A are at risk of being lost in terms of diversification opportunities for the investors. While in *CN* the parameter $k = 1$, the *MRI* and the *ARI* are obtained setting $k = 3$, highlighting that financial crisis can correspond to situations in which $\text{rank}(A) < n - 1$; the economic interpretation of the algebraic result on the matrix A is that during financial crises it becomes very difficult to effectively diversify because all of the activities tend to positive correlate and show similar behaviors.

Starting from the definitions given above, we briefly resume the strengths and weaknesses of the measures under comparison. The *AC* depends exclusively on the entries of the correlation matrix; it is very simple and intuitive but it suffers from the fact that the dependence structure is pairwise and linear. The *CRF*, *MRI*, *CN* and *ARI* are functions of the singular values of the correlation matrix; thanks to this peculiarity, they overcome the issue related to the pairwise structure of the correlation while they are still calculated in a linear dependence context. The idea behind the *CRF* is opposite with respect to the one inspiring *MRI*, *CN* and *ARI*: the *CRF* depends on the weight of the first principal component while *MRI*, *CN* and *ARI* are calculated on the basis of the smaller principal components. Note that even if, usually, the increase of the first principal component corresponds to a decrease of the last components, it can also happen that a change in the weight of the first principal component is not reflected by a correspondent opposite variation in the weight of the last component. For this reason the measures can react differently to market

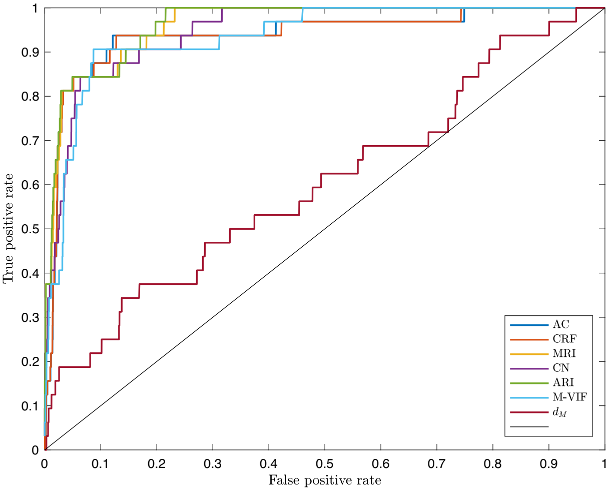


Fig. 2 ROC curves for the S &P500 index, window length $w = 20$

changes. The *M-VIF* is a function of the R^2 of specific linear regressions as pointed out in the definition; it is calculated in a linear framework as all the other measures but it represents an alternative way to overcome the issue related to the pairwise structure of the correlation. Finally, the Mahalanobis distance (d_M) is completely different from the other measures; its main drawbacks have to be found in the relation with the Normal distribution and in the fact that it depends only on the last available observation

4.2 In-sample descriptive comparison

In order to evaluate the descriptive power of the different systemic risk measures we propose an in-sample exercise.

With respect to Definition 1 we set $w = 20$ and a threshold loss $l = -0.01$; in words, a systemic event occurs when for $w = 20$ consecutive days, approximately one working month, the average daily return of the S &P500 index is less or equal than $l = -0.01$. The analysis is performed through a rolling-window procedure: given the $T = 8125$ observations dataset as described in Sect. 4.1, we set the window length equal to w.[3] Then, starting from $w + 1$, the previous w observations in the vector y are used to check the occurrence of the systemic event while the first w rows of matrix A are used to calculate the correlation-based systemic risk measures as defined in Sect. 4.1. The described process iteratively continues dropping the first return in vector y and the first row in matrix A and adding the returns of the subsequent period to update the values of the risk measures and check the occurrence of the systemic event. This iterative procedure performed on the entire dataset produces the couples $(m_t,^w y_t)$ to be analyzed as described in Sect. 3. Figure 2 contains the ROC curves obtained using the systemic risk measures listed in the previous section as the independent variables; in Table 1 are summarized the values of the AUC for each measure.

[3] In this case, for simplicity, the length of the rolling window used for the calculation of the risk measure coincides to the length of the window used for the definition of the systemic event; in general it is not necessary for the two windows to have the same length.

Table 1 AUCs

Risk measure	AUC
AC	0.9386
CRF	0.9381
MRI	0.9614
CN	0.9474
ARI	0.9633
M-VIF	0.9361
d_M	0.6008

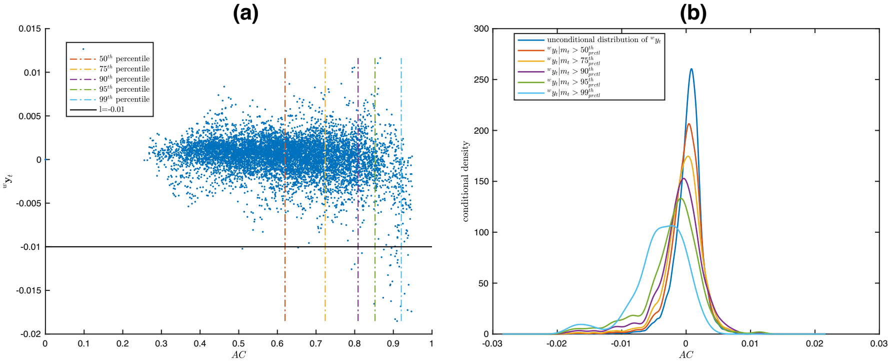


Fig. 3 Scatter plot with respect to the *AC* and the correspondent conditional distributions of returns

On the basis of Fig. 2 and Table 1 the following comments are evident. First, in the descriptive framework almost all the proposed systemic risk measures show a very similar behavior. The only exception is the Mahalanobis distance that is associated with a low value of AUC close to 50% and a ROC curve close to the diagonal, revealing a limited classification power that makes its performance almost indistinguishable from the one of a random classifier. The other measures are characterized by high values of AUC between 0.93 and 0.96 and ROC curves that are close to the perfect classification. This means that the considered correlation-based systemic risk measures seem to be able to correctly classify the presence or absence of a systemic event in a descriptive context. Moreover, considering that all the ROC curves intersect, it is impossible, in the present context, to identify a systemic risk measure that is generally preferable with respect to the others. While the overall analysis shows a substantial equivalence in terms of classification performance of the correlation-based systemic risk measures under comparison, a more detailed investigation can highlight significant differences within the measures. Let compare, for example, the *AC* and the *MRI*. Figure 3a, contains the scatter plot with respect to the *AC*; the vertical lines in the graph correspond to different values of $\overline{m}$. The conditional densities in Fig. 3b are obtained applying the MatLab Gaussian kernel smoothing function, see Peter (1985), to the points $({}^{w}y_t, m_t > \overline{m})$. Figure 4 is the analogous to 3 with respect to *MRI*. Looking at the results, when conditioning to high values of the risk measure, the conditional densities in Figs. 3b and 4b move leftwards and flatten for both the measures. However, the *MRI* clearly over-performs the *AC* in terms of classification accuracy.

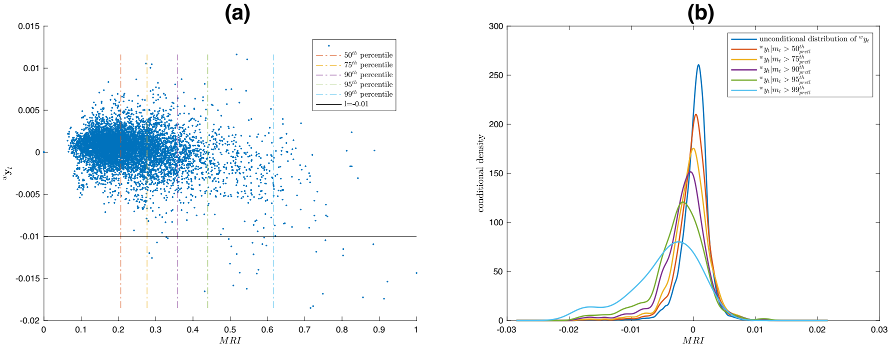


Fig. 4 Scatter plot with respect to the *MRI* and the correspondent conditional distributions of returns

The scatter plots for the two measures depicted in Figs. 3 and 4 strongly qualitatively differ. Considering the *AC*, almost all the points belong to a unique cloud of points uniformly distributed on the graph. On the opposite, if we look at the distribution of the points with respect to the *MRI*, we can notice that most of the points are clustered on the top-left of the graph, while few points are located on the bottom-right. This qualitative consideration highlights that the *MRI* shows a better discriminating power than the *AC*. In this context, the discriminating power is the capacity to separate the small number of observations corresponding to the rare severe systemic events from the large majority of points where the economic system behaves normally.

In our opinion, the fact that this significant difference between the measures remains hidden when comparing them through the ROC curves and the AUCs depends on many components. First, the time series employed in our analysis contain more than 8000 daily observations. We believe that the request to classify a huge number of daily observations can worsen the issue of handling an imbalanced database. Second, regarding the evaluation of the best performing classifier, the experimental approach developed on the base of the ROC curve and related concepts, only partially captures the complex behavior of each individual measure and it is probably useful only for a first global gross assessment. In our opinion, this depends on the fact that the ROC curves and related tools are developed to compare binary classifiers with the goal of making the comparison independent from the arbitrary choice of the threshold used for the binarization of the independent variable. In the present case, to highlight the fundamental differences in terms of classification performance between the different measures, it is necessary to focus on what happens for extremely high values of m_t, when the severe systemic events are expected to occur. The comparison between the *AC* and the *MRI*, even though all the measures are normalized, also suggests that in practical applications each systemic risk measure requires to choose a specific suitable threshold value.

With respect to the conditional distributions depicted in Figs. 3 and 4, Table 2 shows the mean, standard deviation and Value at Risk at the 1% (V@R) for each of the correlation-based risk measures under comparison. The V@R as risk indicator is employed because it is widely used by practitioners and it explicitly accounts for the risk of losses. More specifically, three effects can be recognized from Table 2 while increasing the threshold ($\overline{m}$) on the measure: the mean of the conditional distributions decreases, the standard deviation increases and the V@R of the market index increases. This first

Table 2 Daily average return, standard deviation and Value at Risk at 1% significance level of the returns distributions conditioned to given percentiles of the correlation-based risk measures under comparison

	uncond	${}^{w}y_t\|m_t\ 50^{\text{th}}_{prctl}$	${}^{w}y_t\|m_t\ 90^{\text{th}}_{prctl}$	${}^{w}y_t\|m_t\ 95^{\text{th}}_{prctl}$	${}^{w}y_t\|m_t\ 99^{\text{th}}_{prctl}$
AC					
mean	0.0003	– 0.0001	– 0.0011	– 0.0022	– 0.0039
stdev	0.0022	0.0026	0.0038	0.0043	0.0040
$V@R_{1\%}$	0.0067	0.0088	0.0155	0.0170	0.0178
CRF					
mean	0.0003	– 0.0001	– 0.0013	– 0.0023	– 0.0052
stdev	0.0022	0.0026	0.0037	0.0044	0.0051
$V@R_{1\%}$	0.0066	0.0088	0.0155	0.0171	0.0198
MRI					
mean	0.0003	– 0.0001	– 0.0014	– 0.0024	– 0.0046
stdev	0.0022	0.0027	0.0037	0.0043	0.0054
$V@R_{1\%}$	0.0067	0.0089	0.0155	0.0170	0.0200
CN					
mean	0.0003	– 0.0001	– 0.0013	– 0.0021	– 0.0040
stdev	0.0022	0.0027	0.0038	0.0041	0.0053
$V@R_{1\%}$	0.0067	0.0089	0.0155	0.0167	0.0195
ARI					
mean	0.0003	– 0.0001	– 0.0014	– 0.0024	– 0.0045
stdev	0.0022	0.0026	0.0037	0.0043	0.0053
$V@R_{1\%}$	0.0067	0.0089	0.0155	0.0170	0.0198
M-VIF					
mean	0.0003	– 0.0001	– 0.0014	– 0.0019	– .0037
stdev	0.0022	0.0027	0.0038	0.0042	0.0052
$V@R_{1\%}$	0.0067	0.0089	0.0155	0.0169	0.0191
d_M					
mean	0.0003	0.0001	– 0.0002	– 0.0005	– 0.0008
stdev	0.0022	0.0023	0.0027	0.0029	0.0034
$V@R_{1\%}$	0.0067	0.0075	0.0099	0.0118	0.0152

macroscopic evidence suggests that the measures are suitable indicators for discriminating between the presence or absence of a systemic crisis.

The results resumed in Table 2 are very interesting. The first evidence is how the different measures behave very similarly when we compare the distributions of returns of the index conditioned to small values of the threshold. In our opinion this is not an argument against the necessity of using different measures, but the empirical evidence of the fact that the data are strongly unbalanced. With respect to a length of the time series of $T = 8125$, the events that verify the conditions of Definition 1 with the chosen levels of w and l are approximately 50. If we focus on the last column of Table 2 it is possible to notice very interesting and significant differences between the measures. The measure associated with the most extremes values of the conditional distribution is the MRI, showing the best performance in terms of classification power. We want to underline how, in this context, apparently small differences between the results hide a considerable economic impact. For example, if we compare the *CRF* and the

Table 3 S &P500 index, in-sample sensitivity analysis: value at risk at 1% significance level of the returns distributions conditioned to 99th percentile and AUC for different levels of *l* and *w*

	$V@R_{1\%}$	l = −0.005	l = −0.01	l = −0.015	l = −0.02
w = 15					
AC	0.0211	0.7834	0.9188	0.9772	0.9707
CRF	**0.0223**	0.7981	0.9180	**0.9871**	**0.9888**
MRI	**0.0223**	0.8187	0.9251	0.9762	0.9793
CN	0.0208	0.7817	0.9031	0.9443	0.9712
ARI	0.0225	**0.8217**	**0.9258**	0.9805	0.9809
M-VIF	0.0188	0.7772	0.8857	0.9329	0.9694
d_M	0.0133	0.5951	0.6079	0.5012	0.7792
w = 20					
AC	0.0178	0.7977	0.9386	0.9823	[–]
CRF	0.0198	0.8184	0.9466	**0.9901**	[–]
MRI	**0.0200**	**0.8501**	0.9614	0.9878	[–]
CN	0.0195	0.8388	0.9474	0.9802	[–]
ARI	0.0198	0.8492	**0.9633**	0.9885	[–]
M-VIF	0.0191	0.8366	0.9361	0.9748	[–]
d_M	0.0151	0.5874	0.6008	0.7681	[–]
w = 25					
AC	0.0147	0.8178	0.9583	0.9847	[–]
CRF	0.0165	0.8345	0.9679	**0.9991**	[–]
MRI	0.0167	0.8798	**0.9796**	0.9977	[–]
CN	**0.0173**	0.8807	0.9747	0.9967	[–]
ARI	0.0167	0.8758	0.9790	0.9976	[–]
M-VIF	0.0167	**0.8903**	0.9761	0.9956	[–]
d_M	0.0132	0.6486	0.7193	0.8071	[–]

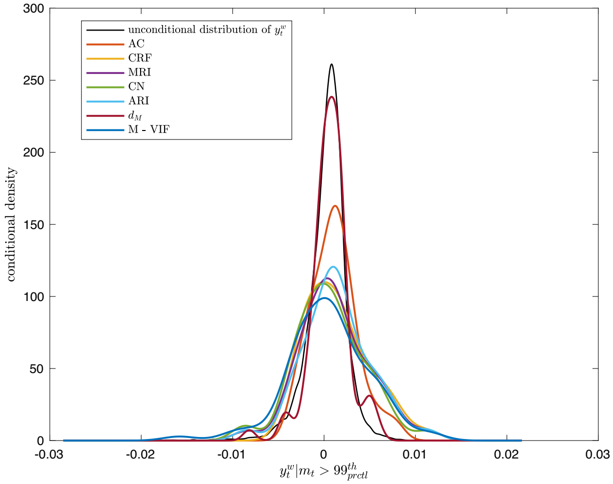
The bold is used to highlight the highest values within the competing measures on the given time window w and loss l, while [-] denotes the missing values

MRI with respect to the $V@R_{1\%}$ the difference is equal to ≈ 0.0022. Recalling that the empirical exercise is built on daily data, this apparently small difference of detected daily expected loss at a significance level of 1% corresponds to a monthly difference of $0.0022 * 20 \approx 4.4\%$.

Considering the arbitrariness of the choice of the parameters that define the systemic event, $l = -0.01$ and $w = 20$, in the previous example, we perform an analogous experiment for different values of $w = 15, 20, 25$ and $l = -0.005, -0.01, -0.015, -0.02$ aiming at supporting the robustness of our findings. Each combination of w and l results in a different binarization of the original data. The results are collected in Table 3.

We provide some comments on the results in Table 3. First, the missing values depend on the fact that for some specific choices of w and l there are no events; for example, in our database, there does not exist a period of $w = 20$ consecutive days associated with a daily average loss 0.02. The results in Table 3 confirm the findings: in the in-sample framework, the main differences among the measures can be seen in terms of conditional $V@R$ while, apparently, the measures look very similar in terms of AUC. There is no single measure that can be identified as definitively the best;

Fig. 5 S &P500 returns distribution conditioned to the 99th percentile



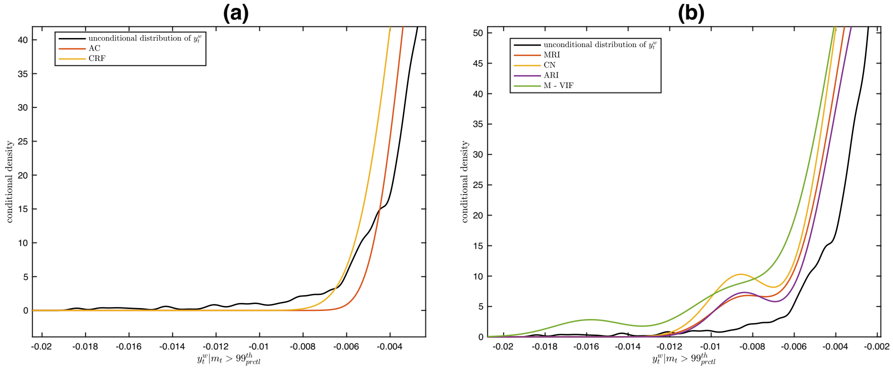


Fig. 6 Left tails of S &P500 returns distribution conditioned to the 99th percentile

nevertheless, the measures showing better performance are the ones that depend on the eigenvalues of the correlation matrix while the d_M is permanently the worse.

4.3 Out-of-sample forecasting comparison

The out-of-sample exercise compares the alternative measures of systemic risk in terms of predictive power with the aim of understanding which of them are more suitable to be used in a forecasting context.

The values of the parameters are preserved with respect to the definition of crisis, $w = 20$ and $l = -0.01$. The analysis is performed through a rolling-window procedure: given the $T = 8125$ observations dataset as described in Sect. 4.1, we set the window length equal to w. Then, starting from $w + 1$, the next w observations in the vector y are used to check the future occurrence of the systemic event while the first w rows of matrix A are used to calculate the correlation-based systemic risk measures as defined in Sect. 4.1. The described process iteratively continues dropping the first return in vector y and the first row in matrix A and adding the returns of the subsequent period to recalculate the risk

Table 4 AUCs for different level of w_e

w_e	0	5	20	60
AC	0.6903	0.6131	0.3908	0.4513
CRF	0.7213	0.6867	0.6072	0.5507
MRI	0.7844	0.7694	0.6122	0.5436
ARI	0.7775	0.7614	0.6098	0.5326
CN	0.7882	0.7713	0.5980	0.6076
d_M	0.6363	0.5863	0.4196	0.5155
M-VIF	0.7825	0.7362	0.6799	0.6303

measures and check the occurrence of the systemic event on y. This iterative procedure performed on the entire dataset produces the couples (m_t, y_t^w) to be analyzed as described in Sect. 3.

In Fig. 5 the returns distributions conditioned to the 99th percentile are depicted against the unconditional distribution of returns. This comprehensive representation only provides a first general impression on the fact that the Mahalanobis distance does not work and that the discriminating power of all the other measures is lower in the forecasting framework if compared to the descriptive performance.

Focusing on the left tail of the conditioned distribution of returns, we are able to highlight interesting differences among the measures. In particular, Fig. 6 shows two opposite behaviors: in sub-figure (a) it is possible to notice how the left tails of the returns distribution conditioned to the AC and CRF are lighter compared to the left tail of the unconditional distribution. This evidence empirically shows that the AC and the CRF are not useful in forecasting systemic events. In sub-figure (b) the left tails of the returns distributions conditioned to the other measures are depicted. It is evident how, even if with different levels, the MRI, the CN, the ARI and the M-VIF are able to discriminate extreme events also in a forecasting framework. From a graphical point of view, the measure that shows the better performance is the M-VIF.

Considering the importance of the predictive power of systemic risk measures, we introduce a shift parameter w_e to evaluate how in advance a measure is able to provide the warning signal. We underline that early warnings could be extremely helpful in having the time to shed against future possible losses or financial turbulence. We set four values for the parameter $w_e = 0, 5, 20$ or 60 to investigate four stylized scenarios in which the measures are tested for their capacity to anticipate a systemic event that is going to occur respectively tomorrow, in one week, in one month or in a one trimester. The AUCs for the different measures in the four scenarios are resumed in Table 4.

The first evidence is that the forecasting power of the measures decreases with the increase of the shift parameter w_e. This phenomenon is expected and it is clear both from the values of the AUCs collected in Table 4 and from the graphical behavior of the ROC curves in Fig. 7. The values of AUC approach approximately 0.5 when w_e increases and the ROC curves flatten on the diagonal making the results indistinguishable from a random classifier.

Comparing the performance of the single measures we can notice that few of them show very interesting forecasting power on a short term horizon ($w_e = 0$): in particular the *MRI*, the *ARI* and the *M-VIF* are the best with associated AUC values close to 80%. When the forecasting horizon gets longer the above mentioned measures remain the best among the

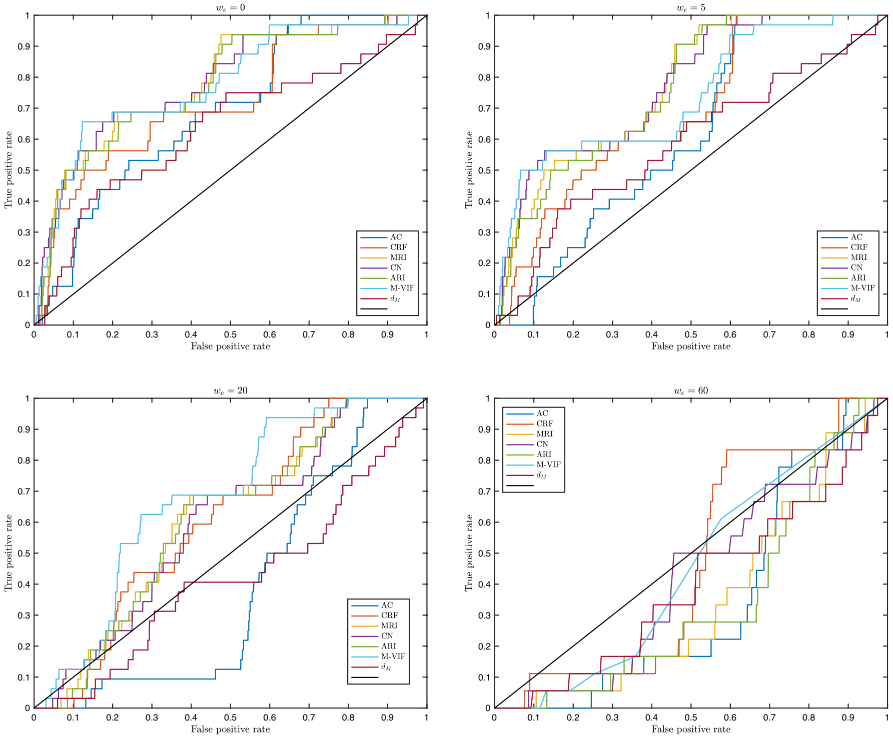


Fig. 7 ROC curves with respect to the correlation-based risk measures under comparison, shift parameter $w_e = 0, 5, 20, 60$

considered measures; the *MRI* and the *ARI* maintain very interesting performance on the weekly horizon while the *M-VIF* seems to over-perform in longer term horizons.

In Table 5 we present the results of a sensitivity analysis performed in the out-of-sample framework for $w_e = 0, l = -0.005, -0.01, -0.015, -0.02$ and $w = 15, 20, 30$.

The results of the sensitivity analysis confirm what obtained in the previous experiment. Considering the AUC, many measures look very similar while they strongly differ if we consider the conditional V@R. One striking example is given by the *AC* that is the best measure in terms of *AUC* for $w = 15, 20$ while it significantly under-perform in detecting the big losses as showed by the V@R.

5 Conclusions

Policy makers and investors need to describe and predict systemic events in order to prevent or, at least, reduce the negative impact of financial crises and market downturns. Then, it is essential to compare the performance of the different systemic risk measures proposed in the literature.

In this paper we define a systemic event in a very natural way through the specification of its duration and the associated average loss. After choosing a comprehensive financial index that represents the whole economic system, we perform an empirical

Table 5 S &P500 index, out-of-sample sensitivity analysis for $w_e = 0$: value at risk at 1% significance level of the returns distributions conditioned to 99th percentile and AUC for different levels of l and w

	$V@R_{1\%}$	l = −0.005	l = −0.01	l = −0.015	l = −0.02
w = 15					
AC	0.0121	**0.6574**	0.7919	0.8831	0.7359
CRF	**0.0144**	0.6537	0.7889	**0.9192**	0.8240
MRI	0.0119	0.6550	0.8126	0.9083	0.8501
CN	0.0084	0.6420	0.7621	0.8455	0.8339
ARI	0.0117	0.6566	**0.8226**	0.9166	**0.8538**
M-VIF	0.0083	0.6372	0.7416	0.8217	0.8300
d_M	0.0135	0.5144	0.5017	0.6218	0.6907
w = 20					
AC	0.0045	**0.6680**	0.6903	0.5265	[–]
CRF	0.0055	0.6610	0.7213	0.6051	[–]
MRI	0.0087	0.6677	0.7844	0.7030	[–]
CN	0.0096	0.6557	**0.7882**	**0.7126**	[–]
ARI	0.0089	0.6659	0.7775	0.6905	[–]
M-VIF	**0.0144**	0.6522	0.7825	0.6929	[–]
d_M	0.0077	0.5414	0.6363	0.5185	[–]
w = 25					
AC	0.0039	0.6206	0.5166	0.3778	[–]
CRF	0.0041	0.6320	0.6286	0.4076	[–]
MRI	0.0043	0.6402	0.7027	0.5215	[–]
CN	0.0060	0.6475	**0.7269**	**0.6050**	[–]
ARI	0.0043	0.6379	0.6951	0.4985	[–]
M-VIF	**0.0120**	**0.6501**	0.7023	0.5345	[–]
d_M	0.0065	0.5750	0.5536	0.0350	[–]

The bold is used to highlight the highest values within the competing measures on the given time window w and loss l, while [-] denotes the missing values

Fig. 8 ROC curves

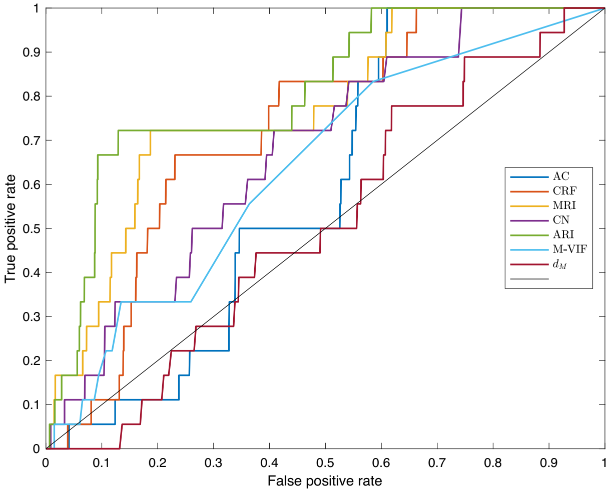


Table 6 AUCs

Risk measure	AUC
AC	0.8026
CRF	0.8865
MRI	0.9003
CN	0.8482
ARI	0.9133
M-VIF	0.8409
d_M	0.6866

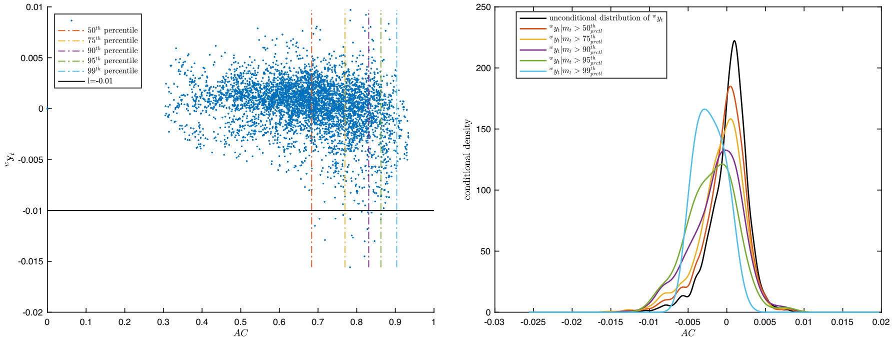


Fig. 9 Scatter plot with respect to the *AC* and the correspondent conditional distributions of returns

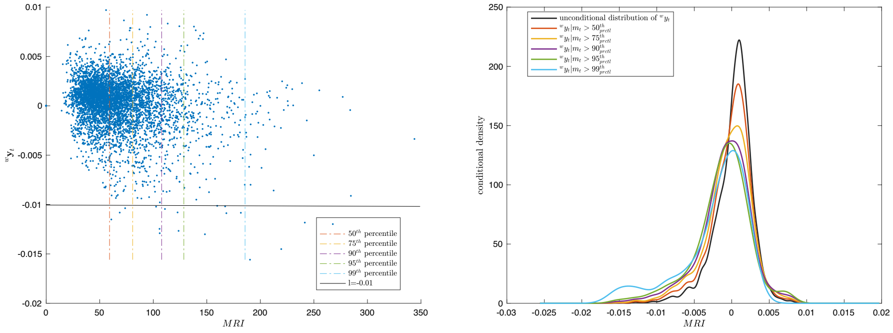


Fig. 10 Scatter plot with respect to the *MRI* and the correspondent conditional distributions of returns

comparison of alternative correlation-based systemic risk measures. From a descriptive perspective and in a standard binary classification framework, the measures based on the correlation of markets' sectors show a considerably high power in correctly discriminate periods of crisis and financial turbulence. Although the overall comparison does not highlight significant differences among the measures, except for the Mahalanobis distance that shows a very peculiar behavior, a more detailed analysis of

Table 7 Daily average return, standard deviation and Value at Risk at 1% significance level of the returns distributions conditioned to given percentiles of the correlation-based risk measures under comparison

	uncond	${}^{w}y_t\|m_t\ 50^{\text{th}}_{prctl}$	${}^{w}y_t\|m_t\ 90^{\text{th}}_{prctl}$	${}^{w}y_t\|m_t\ 95^{\text{th}}_{prctl}$	${}^{w}y_t\|m_t\ 99^{\text{th}}_{prctl}$
AC					
mean	0.0003	– 0.0003	– 0.0013	– 0.0019	– 0.0022
stdev	0.0025	0.0028	0.0033	0.0032	0.0018
$V@R_{1\%}$	0.0080	0.0092	0.0099	0.0099	0.0065
CRF					
mean	0.0003	– 0.0003	– 0.0016	– 0.0021	– 0.0022
stdev	0.0025	0.0029	0.0037	0.0037	0.0020
$V@R_{1\%}$	0.0080	0.0092	0.0125	0.0130	0.0070
MRI					
mean	0.0003	– 0.0001	– 0.0009	– 0.0013	– 0.0023
stdev	0.0025	0.0029	0.0035	0.0038	0.0045
$V@R_{1\%}$	0.0080	0.0092	0.0122	0.0136	0.0162
CN					
mean	0.0003	0.0000	– 0.0001	0.0001	– 0.0010
stdev	0.0025	0.0028	0.0030	0.0031	0.0040
$V@R_{1\%}$	0.0080	0.0090	0.0104	0.0118	0.0146
ARI					
mean	0.0003	– 0.0002	– 0.0012	– 0.0016	– 0.0030
stdev	0.0025	0.0029	0.0036	0.0039	0.0043
$V@R_{1\%}$	0.0080	0.0092	0.0127	0.0137	0.0158
M-VIF					
mean	0.0003	0.0000	– 0.0001	0.0000	– 0.0004
stdev	0.0025	0.0027	0.0031	0.0031	0.0036
$V@R_{1\%}$	0.0080	0.0089	0.0104	0.0119	0.0132
d_M					
mean	0.0003	0.0002	0.0000	– 0.0001	0.0000
stdev	0.0025	0.0027	0.0029	0.0029	0.0031
$V@R_{1\%}$	0.0080	0.0086	0.0100	0.0108	0.0127

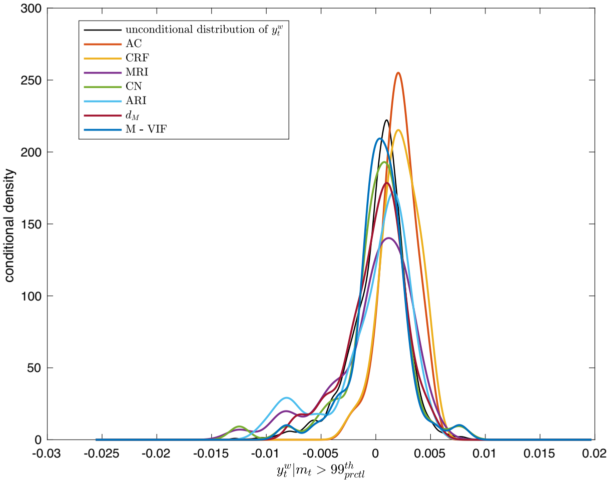
the performance focusing on the left tail of the returns distribution conditioned to high values of the measures reveals significant differences in terms of discriminating power. In particular, the measures based on the eigenvalues of the correlation matrix over-perform the other correlation-based measures; in our opinion, this is due to the fact that the eigenvalues of the correlation matrix carry the information on the dependence among all the constituents of the whole economic system, overcoming the pairwise structure of the correlation. The out-of-sample experiment confirms and reinforces the results of the descriptive exercise. The measures based on the eigenvalues of the correlation matrix show non-negligible discriminating power when used in a predictive context. On the opposite, the average correlation, which performed attractively in the descriptive context, seems to lose all its utility when used in a forecasting framework. Surprisingly, even if based on the eigenvalues of the correlation matrix, also the *CRF* shows a poor discriminating power. On the basis of the empirical results, the measures

Table 8 Eurostoxx index, in-sample sensitivity analysis: value at risk at 1% significance level of the returns distributions conditioned to 99th percentile, AUCs for different level of threshold *l* and window *w*

	$V@R_{1\%}$	l = −0.005	l = −0.01	l = −0.015	l = −0.02
w = 20					
AC	0.0065	0.7871	0.8026	0.7816	[–]
CRF	0.0070	**0.8339**	0.8865	0.9308	[–]
MRI	**0.0162**	0.7272	0.9003	0.9913	[–]
CN	0.0146	0.6368	0.8482	0.8961	[–]
ARI	0.0158	0.7655	**0.9133**	**0.9992**	[–]
M-VIF	0.0132	0.6142	0.8409	0.8850	[–]
d_M	0.0127	0.5750	0.6866	0.3885	[–]
w = 25					
AC	0.0052	0.8155	0.8580	[–]	[–]
CRF	0.0053	**0.8667**	0.9560	[–]	[–]
MRI	**0.0128**	0.8485	**0.9741**	[–]	[–]
CN	0.0124	0.8045	0.9617	[–]	[–]
ARI	0.0123	0.8528	0.9717	[–]	[–]
M-VIF	0.0097	0.7977	0.9154	[–]	[–]
d_M	0.0117	0.6507	0.9008	[–]	[–]
w = 30					
AC	0.0041	0.8251	0.8265	[–]	[–]
CRF	0.0064	**0.8937**	0.9413	[–]	[–]
MRI	**0.0120**	0.8792	0.9445	[–]	[–]
CN	0.0090	0.8470	**0.9546**	[–]	[–]
ARI	**0.0120**	0.8807	0.9422	[–]	[–]
M-VIF	0.0083	0.8364	0.8989	[–]	[–]
d_M	0.0110	0.6952	0.9210	[–]	[–]

The bold is used to highlight the highest values within the competing measures on the given time window w and loss l, while [-] denotes the missing values

Fig. 11 Eurostoxx returns distribution contioned to the 99th percentile



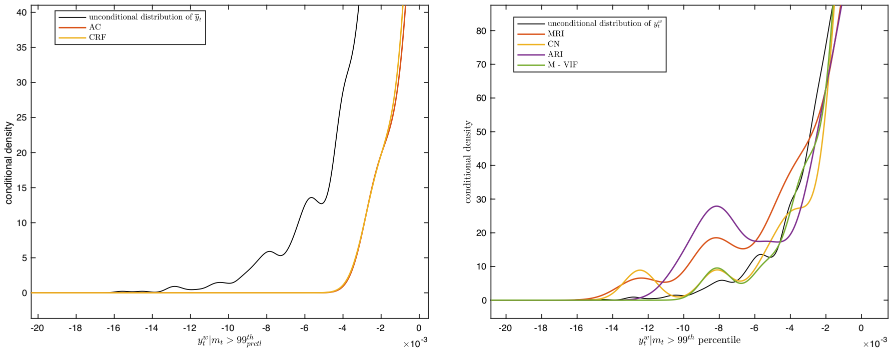


Fig. 12 Left tails of Eurostoxx returns distribution contioned to the 99th percentile

Table 9 AUCs for different level of w_e

w_e	0	5	20	60
AC	0.5884	0.5281	0.3271	0.3490
CRF	0.7251	0.7112	0.5527	0.4680
MRI	0.7696	0.7977	0.6059	0.3522
ARI	0.8103	0.7942	0.6233	0.3457
CN	0.6771	0.7743	0.5725	0.4234
d_M	0.5108	0.5071	0.5434	0.3973
M-VIF	0.6378	0.7200	0.5895	0.4492

that depend on the eigenvalues of the correlation matrix seem to be preferable in practice, both in a descriptive and in a forecasting framework.

Appendix

In order to strengthen the empirical analysis developed in Sect. 4, we perform a similar empirical exercise only changing the referring global financial index. In this application, we use the Eurostoxx index (EUROSTOXX600) as a proxy of the whole economic system. Precisely, we consider the daily returns of the EUROSTOXX600 and its 18 sectoral sub-indexes: AUTOS, BANKS, BASIC RES, CHEMICAL, CONSTRUCT, ENERGY, FIN SERV, FOOD &BEV, HLTHCARE, INDUSTRIAL, INSURERS, MEDIA, PGOODS, RETAIL, TECH, TELCOS, TRAVEL &LESR, UTILITIES.

In-sample descriptive comparison EUROSTOXX600

The EUROSTOXX600 in-sample exercise confirms the results obtained with the S &P500. In general, looking at the ROC curves and related AUC values in Fig. 8 and Table 6, the correlation-based systemic risk measures perform well in a descriptive context.

Moreover, the scatter plots and conditional densities with respect to the *AC* and the *MRI* are compared in Figs. 9 and 10. Looking at the results, the discriminating power of *MRI* is

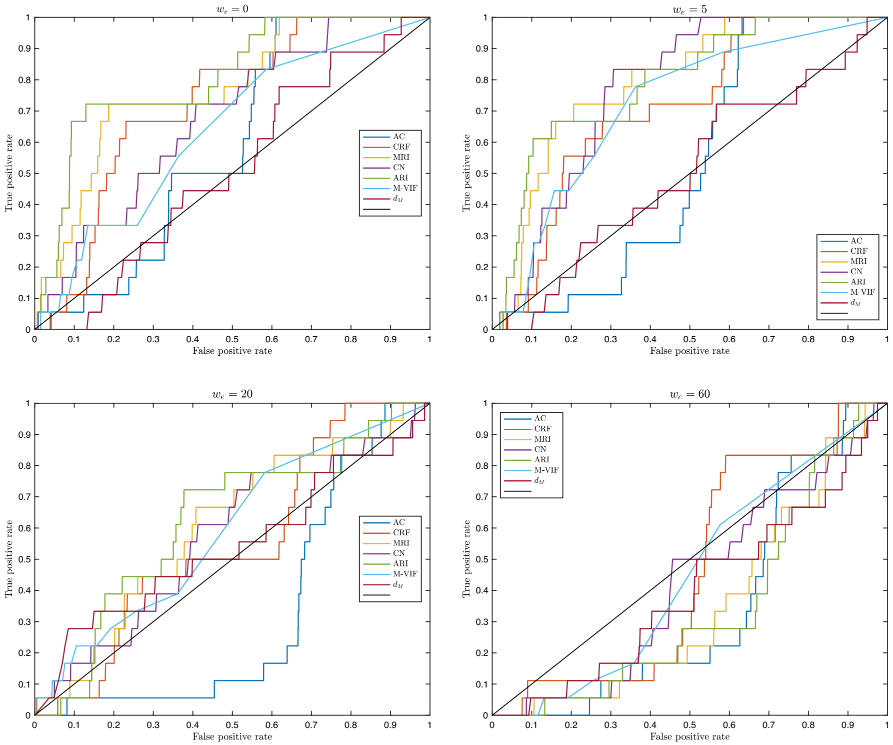


Fig. 13 ROC curves for Eurostoxx index, shift parameter $w_e = 0, 5, 20, 60$

significantly higher than the one of *AC* in term of descriptive performance. This evidence is highlighted and resumed in Table 7, where for each threshold we determine the value of mean, standard deviation and $V@R_{1\%}$ related to each measure.

In Table 8 we provide a sensitivity analysis considering different parameters of severity and duration.

Out-of-sample forecasting comparison EUROSTOXX600

Let consider the predictive power of the systemic risk measures under comparison (Fig. 11). The left tails of the returns distribution conditioned to the *AC* and *CRF* show that these measures are unable to discriminate the presence/absence of financial crises in a predictive context, whereas other measures seem to be useful in forecasting systemic events, as shown in Fig. 12.

In Table 9 we test the predictive power of each measure for different values of the shift parameter w_e. In general, predictive power decreases as the time horizon increases. Considering $w_e = 5$ and $w_e = 20$ we note that the predictive power decreases, but the *MRI* and the *ARI* outperform the other measures showing interesting forecasting power (Fig. 13).

The following results suggest a further consideration: *M-VIF* does not outperform the other measures in long-term horizons as in the previous empirical analysis of S &P500 index.

Table 10 Eurostoxx index, out-of-sample sensitivity analysis for $w_e = 0$: value at risk at 1% significance level of the returns distributions conditioned to 99th percentile, AUCs for different level of threshold l and window w

	$V@R_{1\%}$	l = −0.005	l = −0.01	l = −0.015	l = −0.02
w = 20					
AC	0.0025	0.6090	0.5884	0.4722	[–]
CRF	0.0029	**0.6568**	0.7251	0.8603	[–]
MRI	**0.0125**	0.5981	0.7696	0.8817	[–]
CN	**0.0125**	0.5649	0.6771	0.7674	[–]
ARI	0.0104	0.6145	**0.8103**	**0.9303**	[–]
M-VIF	0.0082	0.5311	0.6378	0.6880	[–]
d_M	0.0078	0.5226	0.5108	0.7333	[–]
w = 25					
AC	0.0019	0.5547	0.5324	[–]	[–]
CRF	0.0018	**0.6228**	0.8732	[–]	[–]
MRI	0.0065	0.6156	**0.9766**	[–]	[–]
CN	0.0118	0.6070	0.9697	[–]	[–]
ARI	0.0080	0.6158	0.9726	[–]	[–]
M-VIF	0.0058	0.6024	0.9609	[–]	[–]
d_M	**0.0134**	0.5648	0.4529	[–]	[–]
w = 30					
AC	0.0003	0.4916	0.3145	[–]	[–]
CRF	0.0004	0.5814	0.7313	[–]	[–]
MRI	0.0056	0.5924	**0.8637**	[–]	[–]
CN	**0.0115**	0.5698	0.8451	[–]	[–]
ARI	0.0055	**0.5959**	**0.8637**	[–]	[–]
M-VIF	0.0043	0.5633	0.8120	[–]	[–]
d_M	0.0061	0.5332	0.8319	[–]	[–]

The bold is used to highlight the highest values within the competing measures on the given time window w and loss l, while [-] denotes the missing values

Table 10 contains a sensitivity analysis for different values of w and l, $w_e = 0$, when the eurostoxx index is used as representative of the whole economic system

In conclusion, this second empirical application based on the use of the EUROSTOXX600 as the global referring financial index substantially confirms the results obtained with respect to S &P500.

Author contributions The authors equally contributed to the design and implementation of the research, the analysis of the results and the writing of the manuscript.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. No funding was received for conducting this study.

Data availibility The datasets analyzed in the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Acemoglu, D., Ozdaglar, A., Tahbaz-Salehi, A.: Systemic risk and stability in financial networks. Am. Econ. Rev. **105**(2), 564–608 (2015)

Acharya, V., Engle, R., Richardson, M.: Capital shortfall: a new approach to ranking and regulating systemic risks. Am. Econ. Rev. **102**(3), 59–64 (2012)

Acharya, V., Pedersen, L.H., Philippon, T., et al.: Measuring systemic risk. Rev. Financial Stud. **30**(1), 2–47 (2017)

Adrian, T., Brunnermeier, M.K.: Covar. Tech. rep., National Bureau of Economic Research (2011)

Allen, L., Bali, T.G., Tang, Y.: Does systemic risk in the financial sector predict future economic downturns? Rev. Financial Stud. **25**(10), 3000–3036 (2012)

Belsley, D.A., Kuh, E., Welsch, R.E.: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, New York (2005)

Bernal, O., Gnabo, J.Y., Guilmin, G.: Assessing the contribution of banks, insurance and other financial services to systemic risk. J. Bank. Finance **47**, 270–287 (2014)

Betz, F., Hautsch, N., Peltonen, T.A., et al.: Systemic risk spillovers in the European banking and sovereign network. J. Financial Stab. **25**, 206–224 (2016)

Bierth, C., Irresberger, F., Weiß, G.N.: Systemic risk of insurers around the globe. J. Bank. Finance **55**, 232–245 (2015)

Billio, M., Getmansky, M., Lo, A.W., et al.: Econometric measures of connectedness and systemic risk in the finance and insurance sectors. J. Financial Econ. **104**(3), 535–559 (2012)

Bisias, D., Flood, M., Lo, A.W., et al.: A survey of systemic risk analytics. Annu. Rev. Financ. Econ. **4**(1), 255–296 (2012)

Black, L., Correa, R., Huang, X., et al.: The systemic risk of European banks during the financial and sovereign debt crises. J. Bank. Finance **63**, 107–125 (2016)

Brownlees, C., Engle, R.F.: SRISK: A conditional capital shortfall measure of systemic risk. Rev. Financial Stud. **30**(1), 48–79 (2017)

Cai, J., Eidam, F., Saunders, A., et al.: Syndication, interconnectedness, and systemic risk. J. Financial Stab. **34**, 105–120 (2018)

Cerchiello, P., Giudici, P.: Categorical network models for systemic risk measurement. Qual. Quant. **51**(4), 1593–1609 (2017)

Demirer, M., Diebold, F.X., Liu, L., et al.: Estimating global bank network connectedness. J. Appl. Economet. **33**(1), 1–15 (2018)

Diebold, F.X., Yilmaz, K.: Measuring financial asset return and volatility spillovers, with application to global equity markets. Econ. J. **119**(534), 158–171 (2009)

Diebold, F.X., Yilmaz, K.: Better to give than to receive: Predictive directional measurement of volatility spillovers. Int. J. Forecast. **28**(1), 57–66 (2012)

Diebold, F.X., Yılmaz, K.: On the network topology of variance decompositions: measuring the connectedness of financial firms. J. Econom. **182**(1), 119–134 (2014)

Duca, M.L., Peltonen, T.A.: Assessing systemic risks and predicting systemic events. J. Bank. Finance **37**(7), 2183–2195 (2013)

Elliott, M., Golub, B., Jackson, M.O.: Financial networks and contagion. Am. Econ. Rev. **104**(10), 3115–53 (2014)

Engle, R., Jondeau, E., Rockinger, M.: Systemic risk in Europe. Rev. Finance **19**(1), 145–190 (2015)

Figini, S., Uberti, P.: Model assessment for predictive classification models. Commun. Stat. Theory Methods **39**(18), 3238–3244 (2010)

Figini, S., Maggi, M., Uberti, P.: The market rank indicator to detect financial distress. Econom. Stat. **14**, 63–73 (2020)

Giglio, S., Kelly, B., Pruitt, S.: Systemic risk and the macroeconomy: an empirical evaluation. J. Financial Econ. **119**(3), 457–471 (2016)

Girardi, G., Ergün, A.T.: Systemic risk measurement: multivariate GARCH estimation of COVAR. J. Bank. Finance **37**(8), 3169–3180 (2013)

Golub, G., Van Loan, C.: Special Linear Systems. Matrix Computations, 2nd edn. John Hopkins University Press, Baltimore (1989)

Hand, D.J., Till, R.J.: A simple generalisation of the area under the roc curve for multiple class classification problems. Mach. Learn. **45**(2), 171–186 (2001)

Härdle, W.K., Wang, W., Yu, L.: Tenet: tail-event driven network risk. J. Econom. **192**(2), 499–513 (2016)

Hartmann, P., De Bandt, O., Molyneux, P., et al.: The concept of systemic risk. Financial Stab. Rev. 134–142 (2009)

Hautsch, N., Schaumburg, J., Schienle, M.: Financial network systemic risk contributions. Rev. Finance **19**(2), 685–738 (2015)

Laeven, L., Ratnovski, L., Tong, H.: Bank size, capital, and systemic risk: Some international evidence. J. Bank. Finance **69**, S25–S34 (2016)

López-Espinosa, G., Moreno, A., Rubia, A., et al.: Short-term wholesale funding and systemic risk: a global COVAR approach. J. Bank. Finance **36**(12), 3150–3162 (2012)

López-Espinosa, G., Rubia, A., Valderrama, L., et al.: Good for one, bad for all: determinants of individual versus systemic risk. J. Financial Stab. **9**(3), 287–299 (2013)

Lopez-Espinosa, G., Moreno, A., Rubia, A., et al.: Systemic risk and asymmetric responses in the financial industry. J. Bank. Finance **58**, 471–485 (2015)

Maggi, M., Torrente, M.L., Uberti, P.: Proper measures of connectedness. Ann. Finance **16**(4), 547–571 (2020)

Mahalanobis, P.C.: On the generalized distance in statistics. National Institute of Science of India (1936)

Mezei, J., Sarlin, P.: Aggregating expert knowledge for the measurement of systemic risk. Decis. Support Syst. **88**, 38–50 (2016)

Pederzoli, C., Torricelli, C.: Systemic risk measures and macroprudential stress tests: an assessment over the 2014 EBA exercise. Ann. Finance **13**(3), 237–251 (2017)

Peter, D.H.: Kernel estimation of a distribution function. Commun. Stat. Theory Methods **14**(3), 605–620 (1985)

Reboredo, J.C., Ugolini, A.: Systemic risk in European sovereign debt markets: a CoVaR-copula approach. J. Int. Money Finance **51**, 214–244 (2015)

Rodríguez-Moreno, M., Peña, J.I.: Systemic risk measures: the simpler the better? J. Bank. Finance **37**(6), 1817–1831 (2013)

Sedunov, J.: What is the systemic risk exposure of financial institutions? J. Financial Stab. **24**, 71–87 (2016)

Silva, W., Kimura, H., Sobreiro, V.A.: An analysis of the literature on systemic financial risk: a survey. J. Financial Stab. **28**, 91–114 (2017)

Varotto, S., Zhao, L.: Systemic risk and bank size. J. Int. Money Finance **82**, 45–70 (2018)

Wang, G.J., Jiang, Z.Q., Lin, M., et al.: Interconnectedness and systemic risk of China's financial institutions. Emerg. Mark. Rev. **35**, 1–18 (2018)

Zhang, D., Broadstock, D.C.: Global financial crisis and rising connectedness in the international commodity markets. Int. Rev. Financial Anal. **68**(101), 239 (2020)

Zheng, Z., Podobnik, B., Feng, L., et al.: Changes in cross-correlations as an indicator for systemic risk. Sci. Rep. **2**(1), 1–8 (2012)