Neurocomputing 264 (2017) 50–56

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Twitter data models for bank risk contagion

Paola Cerchiello*, Paolo Giudici, Giancarlo Nicola

Department of Economics and Management, University of Pavia, Italy

ARTICLE INFO

Article history:
Received 31 March 2016
Revised 6 September 2016
Accepted 16 October 2016
Available online 21 June 2017

Keywords:
Bayesian models
Financial Contagion
Graphical Gaussian models

ABSTRACT

A very important and timely area of research in finance is systemic risk modelling, which concerns the estimation of the relationships between different financial institutions, with the aim of establishing which of them are more contagious/subject to contagion. The aim of this paper is to develop a systemic risk model which, differently from existing ones, employs not only the information contained in financial market prices, but also big data coming from financial tweets. From a methodological viewpoint, we propose a new framework, based on graphical Gaussian models, that can estimate systemic risks with stochastic network models based on two different sources: financial markets and financial tweets, and suggest a way to combine them, using a Bayesian approach. From an applied viewpoint, we present the first systemic risk model based on big data, and show that such a model can help predicting the default probability of a bank, conditionally on the others.

1. Introduction

Systemic risk models address the issue of interdependence between financial institutions and, specifically, measure how bank default risks are transmitted among banks.

The study of bank defaults is important for two reasons. First, an understanding of the factors related to bank failure enables regulatory authorities to supervise banks more efficiently. If supervisors can detect problems early enough, regulatory actions can be taken, to prevent a bank from failing and, therefore, to reduce the costs of its bail-in, faced by shareholders, bondholders and depositors; or those of its bail-out, faced by governments and, ultimately, by the taxpayers. Second, the failure of a bank very likely induces failures of other banks or of parts of the financial system. Understanding the determinants of a single bank failure may thus help to understand the determinants of financial systemic risks, were they due to microeconomic idiosyncratic factors or to macroeconomic imbalances. When problems are detected, their causes can be removed or isolated, to limit "contagion effects".

Most research papers on bank failures are based on financial market models, that originate from the seminal paper of Merton [21], in which the market value of bank assets is matched against bank liabilities. Due to its practical limitations, Merton's model has been evolved into a reduced form (see e.g. Vasicek, [25]), leading to a widespread diffusion of the resulting approach, and the related implementation in regulatory models.

The last few years have witnessed an increasing research literature on systemic risk, with the aim of identifying the most contagious institutions and their transmission channels. Specific measures of systemic risk have been proposed for the banking sector; in particular, by Acharya et al. [1], Adrian and Brunnermeier (2011), Brownlees and Engle (2012), Acharya et al. (2012), Dumitrescu and Banulescu (2014) and Hautsch et al. (2015). On the basis of market prices, these authors calculate the quantiles of the estimated loss probability distribution of a bank, conditional on the occurrence of an extreme event in the financial market.

The above approach is useful to establish policy thresholds aimed, in particular, at identifying the most systemic institutions. However, it is a bivariate approach, which allows to calculate the risk of an institution conditional on another (or on a reference market), but it does not address the issue of how risks are transmitted between different institutions in a multivariate framework.

Trying to address the multivariate nature of systemic risk, researchers have proposed a network modelling approach, following the idea in Diamond and Dybvig [12] and the seminal papers of Sheldon and Maurer [23], Eisenberg and Noe [13], Boss et al. [5], Upper and Worms [24]. In this literature, interconnectedness is related to the detection of the most central players in a network that describes financial flows between agents. While the simplest way of measuring the centrality of a node in the network is by counting the number of neighbours that it has, more sophisticate measures of centrality have been applied, including that shown in Battiston et al. (2012) who develop a network algorithm -the DebtRank- starting from Google's PageRank algorithm.

* Corresponding author.
E-mail addresses: paola.cerchiello@unipv.it (P. Cerchiello), giancarlo.nicola01@universitadipavia.it (G. Nicola).

http://dx.doi.org/10.1016/j.neucom.2016.10.101
0925-2312/

A different type of network models, recently proposed, are based on correlations (or distances) between financial descriptors of agents, such as their stock market prices, bond interest rate spreads or corporate default spreads. The first contributions in this framework are Mantegna [20], Onnela et al. [22], Tumminello et al. (2004) and, recently, Billio et al. [2] and Diebold and Yilmaz (2014), who propose measures of connectedness based on Granger-causality tests and variance decompositions. Barigozzi and Brownlees (2013), Ahelegbey et al. (2015) and Giudici and Spelta (2016) have extended the approach introducing stochastic graphical models.

Here we shall follow this latter approach, and add a stochastic framework, based on graphical models. We will thus be able to derive, on the basis of market price data on a number of financial institutions, the network model that best describes their interrelationships and, therefore, explains how systemic risk is transmitted among them.

It is well known that market prices are formed in complex interaction mechanisms that often reflect speculative behaviours, rather than the fundamentals of the companies to which they refer. Market models and, specifically, financial network models based on market data may, therefore, reflect “spurious” components that could bias systemic risk estimation. This weakness of the market suggests to enrich financial market data with data coming from other, complementary, sources. Indeed, market prices are only one of the evaluations that are carried out on financial institutions: other relevant ones include ratings issued by rating agencies, reports of qualified financial analysts, and opinions of influential media.

Most of the previous sources are private, not available for data analysis. However, summary reports from them are now typically reported, almost in real time, in social networks and, in particular, in tweets. In parallel with these developments, seminal papers on the statistical analysis of such data have recently appeared: see, for example, Bollen et al. [3], Bordino et al. [4], Choi and Varian [10], Feldman [14], Cerchiello and Giudici [9], Andersen (2016)), who all show the added value of tweets and, more generally, of textual data, in economics and finance.

Indeed twitter data offers the opportunity to extract data that can complement market prices and that can, in addition, “replace” market information when not available (as it occurs for banks that are not listed).

To extract from tweets data that can be assimilated to market prices, their text has to be preprocessed using semantic analysis techniques. In our context, if financial tweets on a number of banks are collected daily, semantic analysis allows to obtain a daily “sentiment” that expresses, for each day, how each considered bank is, on average, being evaluated by twitterers.

In this paper, we propose to build graphical Gaussian models using daily variation of bank “sentiment”, and to integrate them with graphical models based on market data, by means of a Bayesian approach. This allows to obtain a comprehensive measurement framework of bank interconnectedness, that can be employed to understand contagion effects.

The novelty of this paper is twofold. From a methodological viewpoint, we propose a framework, based on graphical Gaussian models, that can estimate systemic risks with models based on two different sources: financial markets and financial tweets, and suggest a way to combine them, using a Bayesian approach.

From an applied viewpoint, we propose a novel usage of big data contained in financial tweets, and show that such data can shed further light on the interrelationships between financial institutions.

The rest of the paper is organised as follows: in Section 2 we introduce our proposal; in Section 3 we apply our proposal to financial and tweet data on the Italian banking market and, finally, in Section 4 we present some concluding remarks.

2. Methodology

We first introduce the graphical network models that will be used to estimate relationships between banks, both with market and tweet data.

Relationships between banks can be measured by their partial correlation, that expresses the direct influence of a bank on another. Partial correlations can be estimated assuming that the observations follow a graphical Gaussian model, in which Σ is constrained by the conditional independences described by a graph (see e.g. Lauritzen, [19]).

More formally, let $X = (X_1, \ldots, X_N) \in R^N$ be a N-dimensional random vector distributed according to a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. Without loss of generality, we will assume that the data are generated by a stationary process, and, therefore, $\mu = 0$. In addition, we will assume throughout that the covariance matrix Σ is not singular.

Let $G = (V, E)$ be an undirected graph, with vertex set $V = \{1, \ldots, N\}$, and edge set $E = V \times V$, a binary matrix, with elements e_{ij}, that describe whether pairs of vertices are (symmetrically) linked between each other ($e_{ij} = 1$), or not ($e_{ij} = 0$). If the vertices V of this graph are put in correspondence with the random variables X_1, ..., X_N, the edge set E induces conditional independence on X via the so-called Markov properties (see e.g. Lauritzen, [19]).

In particular, the pairwise Markov property determined by G states that, for all $1 \le i < j \le N$:

$$e_{ij} = 0 \Longleftrightarrow X_i \perp X_j | X_{V \setminus \{i,j\}}; \tag{1}$$

that is, the absence of an edge between vertices i and j is equivalent to independence between the random variables X_i and X_j, conditionally on all other variables $x_{V \setminus \{i,\, j\}}$.

Let the elements of Σ^{-1}, the inverse of the variance-covariance matrix, be indicated as $\{\sigma^{ij}\}$, Whittaker [26] proved that the following equivalence also holds:

$$X_i \perp X_j | X_{V \setminus \{i,j\}} \Longleftrightarrow \rho_{ijV} = 0 \tag{2}$$

where

$$\rho_{ijV} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}} \tag{3}$$

denotes the ijth partial correlation, that is, the correlation between X_i and X_j, conditionally on the remaining variables $X_{V \setminus \{i,\, j\}}$.

Therefore, by means of the pairwise Markov property, and given an undirected graph $G = (V, E)$, a graphical Gaussian model can be defined as the family of all N-variate normal distributions that satisfies the constraints induced by the graph on the partial correlations, as follows:

$$e_{ij} = 0 \Longleftrightarrow \rho_{ijV} = 0 \tag{4}$$

for all $1 \le i < j \le N$.

Stochastic inference in graphical models may lead to two different types of learning: structural learning, which implies the estimation of the graphical structure G that best describes the data, and quantitative learning, that aims at estimating the parameters of a graphical model, for a given graph.

Structural learning can be achieved choosing the graphical structure with maximal likelihood. To this aim, we now recall the expression of the likelihood of a graphical Gaussian model.

For a given graph G, consider a sample X of size n. For a subset of vertices $A \subset N$, let Σ_A denote the variance-covariance matrix of the variables in X_A, and define with S_A the corresponding observed variance-covariance sub-matrix.