Physica A 400 (2014) 109–123

Contents lists available at ScienceDirect

Physica A

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

Visualization of a stock market correlation matrix

Alethea Rea [a], William Rea [b,*]

[a] Data Analysis Australia, Perth, Australia

[b] Department of Economics and Finance, University of Canterbury, Christchurch, New Zealand

HIGHLIGHTS

- We present the first application of Neighbor-Nets to financial data.
- Neighbor-Nets splits graphs are used to visualize a correlation matrix.
- We compare Neighbor-Nets splits graphs with hierarchical clustering trees.
- We compare Neighbor-Nets splits graphs with minimum spanning trees.
- Our results suggest new way to select small diversified stock portfolios.

ARTICLE INFO

Article history:
Received 29 August 2013
Received in revised form 2 January 2014
Available online 18 January 2014

Keywords:
Visualization
Neighbor-Nets
Correlation matrix
Minimum spanning trees
Hierarchical clustering trees

ABSTRACT

This paper presents a novel application of Neighbor-Net, a clustering algorithm developed for constructing a phylogenetic network in the field of evolutionary biology, to visualizing a correlation matrix. We apply Neighbor-Net as implemented in the SplitsTree software package to 48 stocks listed on the New Zealand Stock Exchange. We show that by visualizing the correlation matrix using a Neighbor-Net splits graph and its associated circular ordering of the stocks that some of the problems associated with understanding the large number of correlations between the individual stocks can be overcome. We compare the visualization of Neighbor-Net with that provided by hierarchical clustering trees and minimum spanning trees. The use of Neighbor-Net networks, or splits graphs, yields greater insight into how closely individual stocks are related to each other in terms of their correlations and suggests new avenues of research into how to construct small diversified stock portfolios.

1. Introduction

The basic problem of vast amounts of numerical data overwhelming human capacity to understand them is common to many data-intensive fields of study. The human ability to comprehend large amounts of data when presented in graphical form has long been known and is encapsulated in the old adage "a picture is worth a thousand words". There is a burgeoning field of data visualization, see for example Refs. [1–4] among many others.

Our primary interest in this paper is to further investigate into correlations in stock markets using visualizations. Previously several visualization methods have been applied to correlation matrices of both financial assets and other data types. Hierarchical trees have been applied to stock market data by Mantegna [5] and to exchange rate data by Naylor et al. [6]. Minimum spanning trees have been applied to stock market data by Mantegna, Onnela et al. and Bonanno et al. [5,7,8] and to exchange rate data by Naylor et al. [6]. So-called *h*-plots have been used by Trosset [9] to visualize correlations among gene expression profiles. Partial correlation threshold networks and partial correlation planar graphs have been applied to

* Corresponding author. Tel.: +64 33643474.
E-mail addresses: alethea.rea@gmail.com (A. Rea), bill.rea@canterbury.ac.nz, rea.william@gmail.com (W. Rea).

0378-4371/$ – see front matter
http://dx.doi.org/10.1016/j.physa.2014.01.017

stock market data by Kenett et al. [10]. All of these have yielded useful insights based on their respective visualization methods.

The purpose of our paper is to introduce a visualization method developed for evolutionary biology, Neighbor-Nets [11], to the study of stock market correlations. We show that Neighbor-Nets' circular ordering of stocks and structure of the network offers some distinct advantages over tree based visualization methods in gaining an understanding of the relationships between stocks based on the correlations among their returns.

The remainder of the paper is structured as follows. Section 2 describes the methods used in this paper, namely Neighbor-Net networks and two other methods previously used to study correlations matrices in finance (hierarchical clustering and minimum spanning trees). Section 3 applies these methods to data from the New Zealand Stock Exchange. Section 4 contains the discussion and our conclusions. For those interested in the origin of Neighbor-Net and the technical details on the construction of Neighbor-Net splits graphs, Appendix A provides an overview of the Neighbor-Net clustering algorithm. Appendix B provides a list of three letter stock codes, company names and their exchange-assigned industry group for the stocks discussed in this paper.

2. Methods

A typical stock market correlation matrix for n stocks is of full rank which means that it can only be represented fully in an $(n-1)$-dimensional space. In visualization, the high dimensional data space is collapsed to a much lower dimensional space so that the data can be represented on 2-dimensional surface such as a page or computer screen. When collapsing the $(n-1)$-dimensional space to a 2-dimensional space it is important to make the reduction in dimension in such a way as to retain as much information as possible from the original high-dimensional space.

This section discusses three dimension reduction methods based on clustering: hierarchical clustering trees (HCT, Section 2.2) minimum spanning trees (MST, Section 2.3) and Neighbor-Net networks (Section 2.4). These three techniques are first described individually then compared in Section 2.5, particularly highlighting the advantages of Neighbor-Net networks.

Because each of these three visualization methods requires the relationship between the variables to be expressed as distances, or weights, we begin by describing four potential metrics for converting correlations to distances before choosing one for use in the remainder of the paper.

2.1. Distance metrics

We need to convert the numerical values in the correlation matrix to a measure which can be construed to be a distance. There are at least four ways this can be done. They are

$$d_{ij} = 1 + \rho_{ij} \tag{1}$$

$$d_{ij} = 1 - \rho_{ij} \tag{2}$$

$$d_{ij} = \sqrt{2(1 - \rho_{ij})} \tag{3}$$

$$d_{ij} = \arccos(\rho_{ij}) \tag{4}$$

where d_{ij} is the estimated distance and ρ_{ij} is the estimated correlation between stocks i and j. A simple variation of Eq. (4) is to multiply the result by $2/\pi$ so that it has a range between 0 and 2 as do the other metrics.

Eq. (1) is a measure of dissimilarity. A distance of 0 corresponds to two stocks which have a perfect negative correlation and so are maximally dissimilar. Because the correlation takes values in the interval -1 to $+1$, in this metric the distance between the stocks also grows so that two stocks which are perfect substitutes for each other ($\rho_{ij} = 1$) are considered to be maximally distant. A graph of this conversion from correlation to distance is presented in Fig. 1a.

Eqs. (2)–(4) are all measures of similarity, but they differ in their details. For each of these three metrics, a zero distance corresponds to a perfect correlation, hence no distance between the two stocks because they are perfect substitutes for each other. Similarly, the maximum distance (2 for Eqs. (2) and (3), and π for Eq. (4)), corresponds to a perfect negative correlation.

Eq. (2) is a simple linear transformation from correlation to distance, essentially the opposite of Eq. (1). A graph of this conversion is presented in Fig. 1b.

Eq. (3) is a non-linear transformation which gives a much more rapid separation of stocks with high correlations than does Eq. (2). For example, using Eq. (2) a change of correlation from $\rho_{ij} = 0.95$ to $\rho_{ij} = 0.85$ is a change in distance from 0.05 to 0.15 while for Eq. (3) it is a change from 0.2236 to 0.3873. Eq. (3) is known as the ultrametric and has been used by Mantegna, Naylor et al. and Tumminello et al. [5,6,12] among many others. A graph of this conversion is presented in Fig. 2a.

Eq. (4) is a fundamentally different metric because it converts the correlation into an angular distance. Eq. (4) was used by Trosset [9]. He plotted his variables as radii or diameters on a unit circle using the angles generated by this metric, see Ref. [9] for details. A graph of this conversion is presented in Fig. 2b.

While the reader should be aware that alternative distance metrics have been used by other authorities it is not the purpose of this paper to investigate their relative merits. We will use Eq. (3) throughout to facilitate comparison with the existing literature because it is the most widely used metric.