相関行列からのネットワーク推定
Network estimation from correlation matrix

相馬 亘, 福田 健介, 増川 純一
Wataru Souma, Kensuke Fukuda, Jun-ichi Maskawa

NiCT/ATR CIS 応用ネットワーク科学, souma@atr.jp
NiCT/ATR CIS Applied Network Science Lab.
国立情報学研究所, kensuke@nii.ac.jp
National Institute of Informatics
福山平成大学/NiCT, maskawa@heisei-u.ac.jp
Fukuyama Heisei Univ./NiCT

概 要

Correlation matrix is constructed from the time series data of the Internet traffic. Eigenvalues and eigenvectors are calculated for the correlation matrix. By comparing distribution of eigenvalues obtained from the correlation matrix with that of random matrix theory, the meaningful part of the correlation is extracted. The network corresponding to the meaningful part of the correlation is estimated by the method of minimum spanning tree.

Keywords

Correlation matrix, Random matrix theory, Network estimation,

1 はじめに

本稿は, インターネット・トラフィックの相関構造の解析に, 経済物理学などで開発されてきた手法を応用し, 新たな知見を得ることを目的としている. 特に, 高頻度な株価データの解析に用いられてきた, ランダム行列理論によるノイズ分離の手法 [1, 2, 3, 4] である. また, 多くのネットワークではリンクデータが完全に収集できるわけではなく, 欠損リンクが存在する. そのような場合, ネットワークを推定することが重要である. そこで, 相関行列を元にしてネットワークを推定する手法を議論する.

本稿では, 学術情報ネットワーク (Science Information NETwork; SINET) [5] のトラフィックデータを解析する. ネットワークを研究する場合, そのトポロジーとネットワーク・フローやノード・ダイナミクスの双方のデータが解析できることは重要である. SINET はこれらのデータが入手できる数少ない研究対象の 1 つである.

2 相関行列の分解

SINET を構成するノード i の時刻 t でのトラフィックを $x_{i,t}$ と書き, その対数変化率

$$r_{i,t} := \log(x_{i,t+1}) - \log(x_{i,t}) \tag{1}$$

を考える. ここで $i = 1, \ldots, N$ で $t = 1, \ldots, T$ であり, $N = 494$ は SINET のノード数で, $T = 672$ はトラフィック時系列の長さである. いま, $r_{i,t}$ を規格化して, トラフィック行列を

$$\mathbf{G} := \begin{bmatrix} g_{1,1} & \cdots & g_{1,T} \\ \vdots & & \vdots \\ g_{N,1} & \cdots & g_{N,T} \end{bmatrix}, \ g_{i,t} := \frac{r_{i,t} - \langle r_i \rangle}{\sigma_i} \tag{2}$$

で定義する. ここで, $\langle r_i \rangle$ と σ_i はそれぞれ $r_{i,t}$ の平均と標準偏差である. これを用いると, 相関行列 $\mathbf{C}$

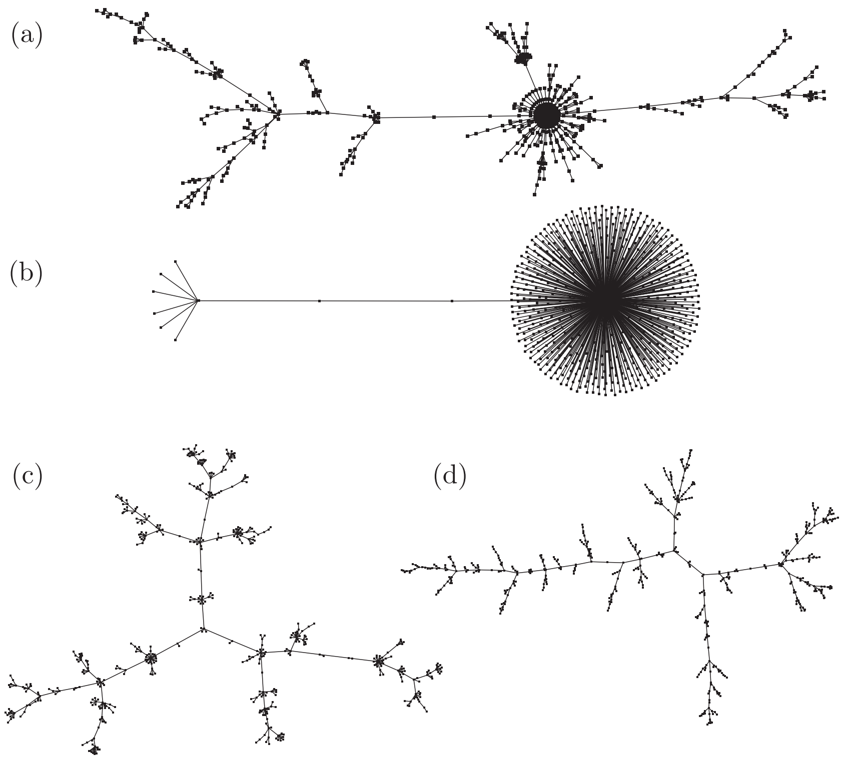

図 1: $\mathbf{C}, \mathbf{C}^{\mathrm{a}}, \mathrm{a} \in \{\mathrm{t.p.}, \mathrm{g.p.}, \mathrm{n.p.}\}$ から構成された MST

は,

$$\mathbf{C} := \frac{1}{T}\mathbf{G}\mathbf{G}^{\mathsf{T}} \tag{3}$$

で定義される.

このCに対して固有値と固有ベクトルを計算し, その結果とランダム行列理論の結果を比較することによって

$$\mathbf{C} = \mathbf{C}^{\mathrm{t.p.}} + \mathbf{C}^{\mathrm{g.p.}} + \mathbf{C}^{\mathrm{n.p.}} \tag{4}$$

と分割できることが分かる [2]. ここで, 式 (4) 右辺の $\mathbf{C}^{\mathrm{t.p.}}, \mathbf{C}^{\mathrm{g.p.}}, \mathbf{C}^{\mathrm{n.p.}}$ はそれぞれ, SINET 全体, グループ構造, ノイズに関する擬相関行列である.

3 ネットワーク推定

相関行列 $\mathbf{C}$ を用いて最小広域木（Minimum Spanning Tree; MST）を作ると, 図 1(a) が得られる. このネットワークは, 木であるけれども塊と判断できそうなものが存在していることと, ハブの存在が特徴である. また, $\mathbf{C}^{\mathrm{t.p.}}$ だけを用いると, 図 1(b) に示されるように巨大なハブを抽出でき, $\mathbf{C}^{\mathrm{g.p.}}$ だけを用いると図 1(c) のようにグループ構造が抽出できる. 一方, $\mathbf{C}^{\mathrm{n.p.}}$ だけを用いた場合は図 1(d) が得られるが, 特徴的な性質は見られない.

参考文献

[1] V. Plerou, et al.: Random matrix approach to cross correlations in financial data. *Phys. Rev. E*, Vol. 65 (2002), 066126.

[2] D.-H. Kim, and H. Jeong: Systematic analysis of group identification in stock markets. *Phys. Rev. E*, Vol. 72 (2005), 046133.

[3] 相馬亘: 経済物理学の実用化に向けて–ランダム行列理論からのアプローチ–. 応用数理, Vol. 15, No. 3 （2005）, 45–59.

[4] 相馬亘, 藤原義久, 尹熙元: 経済物理とランダム行列–株式市場にある本質的な構造の抽出–. 数理科学, 2007 年 2 月号, 44–49.

[5] http://www.sinet.ad.jp/