A generalization of random matrix theory and its application to statistical physics

Duan Wang,[1] Xin Zhang,[2,a)] Davor Horvatic,[3] Boris Podobnik,[1,4,5,6] and H. Eugene Stanley[1]
[1]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA*
[2]*College of Communication and Transport, Shanghai Maritime University, Shanghai 201306, China*
[3]*Physics Department, Faculty of Science, University of Zagreb, Bijenička c. 32, 10000 Zagreb, Croatia*
[4]*Faculty of Civil Engineering, University of Rijeka, Rijeka, HR 51000, Croatia*
[5]*Zagreb School of Economics and Management, Zagreb, HR 10000, Croatia*
[6]*Luxembourg School of Business, Grand-Duchy of Luxembourg, Luxembourg*

(Received 28 October 2016; accepted 19 January 2017; published online 2 February 2017)

To study the statistical structure of crosscorrelations in empirical data, we generalize random matrix theory and propose a new method of cross-correlation analysis, known as autoregressive random matrix theory (ARRMT). ARRMT takes into account the influence of auto-correlations in the study of cross-correlations in multiple time series. We first analytically and numerically determine how auto-correlations affect the eigenvalue distribution of the correlation matrix. Then we introduce ARRMT with a detailed procedure of how to implement the method. Finally, we illustrate the method using two examples taken from inflation rates for air pressure data for 95 US cities. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4975217]

The best method of studying correlations in multiple time series is random matrix theory (RMT). However, RMT assumes that there are no autocorrelations in empirical time series, and in most cases, this assumption does not hold. We analytically study the relationship between the eigenvalue distributions of the correlation matrix of uncrosscorrelated but autocorrelated time series. To take into account the influence of autocorrelations, we propose autoregressive random matrix theory (ARRMT). We use an empirical example to show that the results of ARRMT and RMT can differ greatly when autocorrelations are significant. RMT is thus invalid, and we use ARRMT when autocorrelations are significant.

I. INTRODUCTION

Cross-correlations have been widely observed in nano-devices,[1–3] and in various fields of wave physics such as ultrasonics,[4] underwater acoustics,[5] geophysics,[6,7] seismology,[8] and finance.[9–11] Numerous methods have been introduced to analyze cross-correlations between time series[9,12–16] among which random matrix theory (RMT) is one of the most popular for analyzing cross-correlations in multiple time series.[9,17–24]

The usual approach in RMT is to study the eigenvalue distribution of a Wishart matrix, which is the correlation matrix for a finite-length independent and identically distributed (*i.i.d.*) series, and to compare it with the eigenvalue distribution of the cross-correlation matrix of an empirical time series. Deviations between these two distributions might then suggest the presence of cross-correlations in the data. In this paper, we discuss the limitation of using RMT when empirical data are strongly autocorrelated, and we propose a generalization of RMT, autoregressive random matrix theory (ARRMT), to address this problem.

When cross-correlations are calculated for empirical data, the degree of cross-correlation between the two time series is usually measured by the cross-correlation coefficient, defined as $\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$, where σ_X σ_Y are the standard deviations of X and Y, respectively, and μ_X and μ_Y are the expected values of X and Y, respectively. The sample cross-correlation coefficient can be calculated by

$$r = \frac{1}{T-1}\sum_{i=1}^{T}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right). \quad (1)$$

For the Wishart matrix for N uncorrelated *i.i.d.* time series, each with length $T \geq N$, for large N the eigenvalues follow a Marchenko-Pastur distribution: $P(\lambda) = \frac{Q}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}$,[28] where $Q \equiv \frac{N}{T}$ and

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \quad (2)$$

are the maximum and minimum eigenvalues of W.

As noted above, according to RMT, the difference between the eigenvalue distributions of an empirical cross-correlation matrix and a Wishart matrix indicates the presence of cross-correlations and collective modes in the empirical time series. If cross-correlations are present in the empirical time series, we expect some eigenvalues to be larger than λ_+, where the largest eigenvalue λ_L indicates the global behavior of the multiple time series. The eigenvalues smaller than λ_+ and their corresponding eigenvectors are considered noise.

a)Electronic mail: zhangxin@shmtu.edu.cn

However, RMT has a serious limitation when applied in practice. It does not take into account that the empirical eigenvalue distributions of the cross-correlation matrix can be influenced by auto-correlations in empirical data. The *i.i.d.* time series used to calculate a Wishart matrix generally differs from the empirical time series because, in contrast to an *i.i.d.* time series, an empirical time series has (i) cross-correlations between time series pairs and (ii) auto-correlations in individual time series. Thus, the difference between the eigenvalue distributions of the empirical correlation matrix and the Wishart matrix can be caused by either cross-correlations or auto-correlations in the empirical data.

To take into account the influence of autocorrelations, we propose an autoregressive random matrix theory (ARRMT). We generate uncrosscorrelated time series that have the same autocorrelation properties as the empirical time series, and then calculate their correlation matrix and eigenvalue distribution. Then, we compare it with the eigenvalue distribution of the empirical correlation matrix. This approach takes into consideration the influence of autocorrelations on the eigenvalue distribution.

II. METHODS

A. Impact of autocorrelation on crosscorrelation coefficient

Autocorrelations change the eigenvalue distribution of the uncorrelated time series by changing the distribution of the sample correlated coefficients between each data pair. Considering two arbitrary time series X and Y, when (X, Y) has a bivariate *i.i.d* normal distribution, the Fisher transformation of r, $\frac{1}{2}\ln(\frac{1+r}{1-r})$, is approximately normally distributed with a mean of $\frac{1}{2}\ln(\frac{1+\rho}{1-\rho})$, and a standard error of $\frac{1}{\sqrt{N-3}}$.[26]

For a limit when $|r| \ll 1$ and $T \gg 1$, the distribution of the sample correlation coefficients for N *i.i.d.* series is approximated by a normal distribution with mean zero and standard error $\sqrt{\frac{1}{T}}$. However, the distribution of r will change when both X and Y are autocorrelated time series.

To simplify the derivation of the distribution of r between an autocorrelated time series, we use a standardized time series $z_t = (X_t - \langle X_t \rangle)/s_X$. The sample cross-correlation coefficient between X_t and X'_t is $r = \langle z_t z'_t \rangle = \frac{1}{T}\sum_{t=0}^{T} z_t z'_t$. We assume that X_t and X'_t are not cross-correlated, but that both X_t and X'_t are auto-correlated. Thus, r is a random variable with an expectation of zero and a variance

$$Var(r) = \frac{1}{T^2}\sum_{t}\sum_{t'} E(z_t z_{t'})E(z'_t z'_{t'}), \tag{3}$$

$$= \frac{1}{T^2}\sum_{t}\sum_{t'} d^{|t-t'|}, \tag{4}$$

where we use $E(z_t z_{t'}) = A(|t-t'|)$ and $E(z'_t z'_{t'}) = A'(|t-t'|)$, and $A(|t-t'|)$ and $(A'(|t-t'|))$ are the auto-correlations of X_t and X'_t, respectively, where $|t-t'|$ denotes the time lags.

It is straightforward to show that when $|r| \ll 1$ and $T \gg 1$[29]

$$Var(r) \approx \frac{1}{T}\left[1 + 2\sum_{\Delta t=1}^{\infty} A(\Delta t)A'(\Delta t)\right]. \tag{5}$$

Unlike an *i.i.d.* time series, the variance of sample correlation coefficients increases by $\frac{2}{T}\sum_{\Delta t=1}^{\infty} A(\Delta_t)A'(\Delta t)$. Note that Eq. (5) corresponds to an *i.i.d.* time series with a different number of observations,[30] where the effective number of observations T^* is $\frac{1}{T^*} = \frac{1}{T}\left[1 + 2\sum_{\Delta t=1}^{\infty} A(\Delta t)A'(\Delta t)\right]$. Therefore, T^* has the equivalent length of an autocorrelated time series.

To show how the presence of auto-correlations affects the eigenvalue distribution,[25] we assume that empirical time series are generated by the first-order autoregressive AR(1) process

$$X_t = \phi X_{t-1} + \epsilon_t, \tag{6}$$

where ϕ ($|\phi| < 1$) is a parameter and ϵ is an *i.i.d.* process. The auto-correlation function of an AR(1) process decays with Δt as an exponential function, $A(\Delta t) = \phi^{|\Delta t|}$.[27] Applying Eq. (5), the variance of sample correlation coefficients for two AR(1) processes, each defined by coefficients ϕ and ϕ', respectively, is

$$\mathrm{Var}(r) = \frac{1}{T}\frac{1+\phi\phi'}{1-\phi\phi'}. \tag{7}$$

We take N time series X_t that are not cross-correlated, each series having the same AR(1) coefficient ϕ. Using Eq. (7), where $\phi = \phi'$, and the corresponding expression that holds for *i.i.d.* time series of length T^*, which variance is $\frac{1}{T^*}$, we obtain $T^* = T\frac{1-\phi^2}{1+\phi^2}$. Since the eigenvalue distribution of the cross-correlation matrix generated by the *i.i.d.* time series depends only on $Q = T/N$, we can define an equivalent Q to be

$$Q^* = T^*/N = \frac{T}{N}\frac{1-\phi^2}{1+\phi^2}. \tag{8}$$

Similarly, the eigenvalue distribution becomes

$$P(\lambda) = \frac{Q^*}{2\pi}\frac{\sqrt{(\lambda^*_+ - \lambda)(\lambda - \lambda^*_-)}}{\lambda}, \tag{9}$$

where the largest and smallest eigenvalues are equal to

$$\lambda'_{\pm} = 1 + \frac{1}{Q^*} \pm 2\sqrt{\frac{1}{Q^*}}. \tag{10}$$

These results are approximate and work better when autocorrelations are weak. When autocorrelations are strong, the distribution of the sample correlation coefficients cannot be approximated using a normal distribution—and thus the equations no longer hold—but the largest eigenvalue continues to increase with the autocorrelations because of the increased variance in the crosscorrelations.

Figure 1 shows a simulation of $N = 2000$ time series each with a length $T = 4000$. Using AR(1) processes, with ϕ from 0 to 0.6, the largest eigenvalue increases from 2.9298 to 4.7257.[25,29]

The above expressions are illustrations of how the auto-correlations impact the distribution of the sample correlation

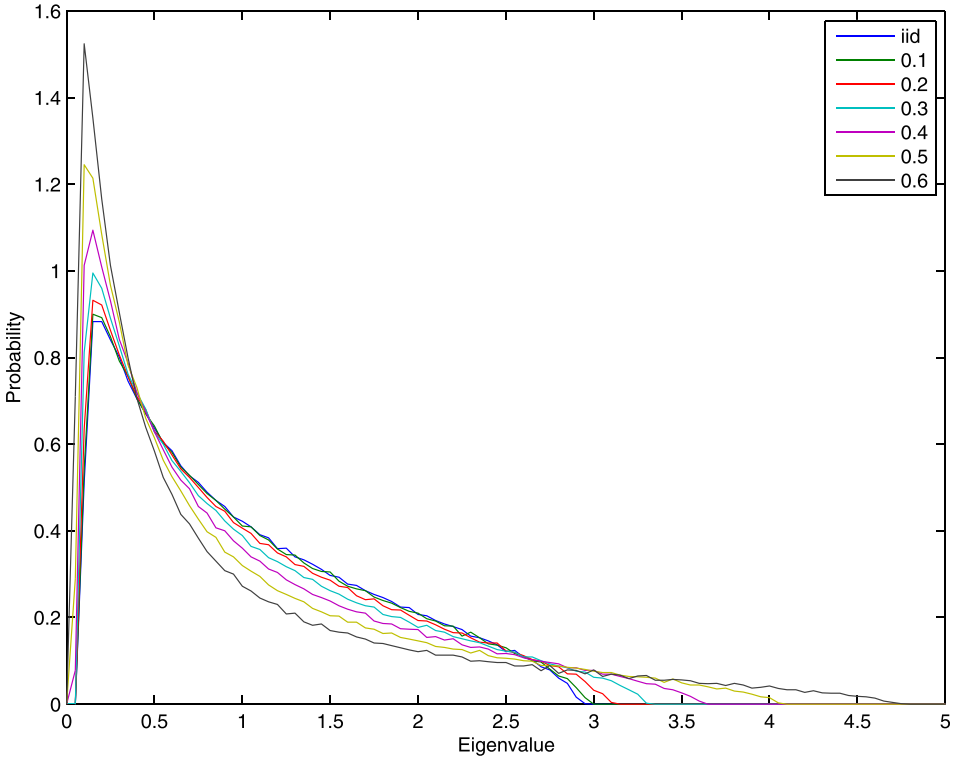

FIG. 1. Eigenvalue distribution for $N = 2000$ autocorrelate time series each with length $T = 4000$. Time series are simulated using AR(1) processes, with ϕ from 0 to 0.6. As can be seen from the figure, the largest eigenvalue increased from 2.9298 to 4.7257 when ϕ increased from 0 to 0.6.

coefficients and the eigenvalues of the correlation matrices, based on which we propose our method in the next section. A theoretical solution for the eigenvalue distribution for correlated Wishart matrices can also be derived.[31]

B. Autoregressive random matrix theory (ARRMT)

To remove the influence of auto-correlations and to study how cross-correlations affect the data, we introduce the auto-regressive Wishart matrix, which is the correlation matrix of the artificial time series $\{Y'_t\}$. This series has no cross-correlations but has the same auto-correlations as those in the empirical time series $\{Y_t\}$. By replacing the Wishart matrix W in RMT with the autoregressive Wishart matrix W′, we create the autoregressive random matrix theory (ARRMT) method and remove the influence of the auto-correlations on the eigenvalue distributions. Similarly, the difference between the eigenvalue distributions of the empirical correlation matrix C and W′ is due solely to the cross-correlations between time series $\{Y_t\}$.

The steps of the ARRMT are as follows:

(i) We test whether auto-correlations are significant among the N cross-correlated original time series $Y_{i,t}$. One of the most popular autocorrelation tests is the Ljung-Box approach.[32]

(ii) We fit each time series $Y_{i,t}$ with the auto-correlation model that best fits $Y_{i,t}$. Based on this fitting, we assign to each series i a set of model parameters (e.g., ϕ_i, θ_i,…). The simplest model is AR(1), but higher orders of autoregressive and moving-average (ARMA) models,

$$Y_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \tag{11}$$

or non-linear models like threshold autoregressive (TAR)[33] can also be used if AR(1) does not fit the auto-correlations. In econometrics, people use time dependent volatility models like generalized autoregressive conditional heteroskedasticity (GARCH) for better fit of the time series data with auto-correlations in the volatilities. However, in ARRMT it is not necessary to use these volatility models because they do not change the distribution of the sample correlation coefficients. According to Eq. (refvar1), the variance of the sample correlation coefficients depends only on the auto-correlation of X_t, not on the volatility of X_t.

(iii) Using the fitted model from (ii), we simulate N time series $Y'_{i,t}$, each characterized by the same coefficients (ϕ_i, θ_i,…) we found in the original time series $Y_{i,t}$. Then, $Y'_{i,t}$ has the same auto-correlation properties as the original time series $Y_{i,t}$.

(iv) We calculate the cross-correlation matrix W′ of the generated time series $Y'_{i,t}$. Then, we calculate the largest eigenvalue λ'_+ of W′.

(v) Finally, we compare the largest eigenvalue λ'_+ with the eigenvalues of the correlation matrix C of the empirical time series. Eigenvalues larger than λ'_+ are related to significant factors.

When N and T are small, the variance of the simulated λ'_+ will be large. Thus, the last two steps in the procedure become:

- We repeat steps (ii) and (iii) n times and calculate the 95th percentile of λ'_+, denoted $\lambda'_{+0.95}$.
- Finally, we compare $\lambda'_{+0.95}$ with the largest eigenvalue λ_L of the correlation matrix C of the empirical time series. Eigenvalues larger than λ'_+ are related to significant factors.

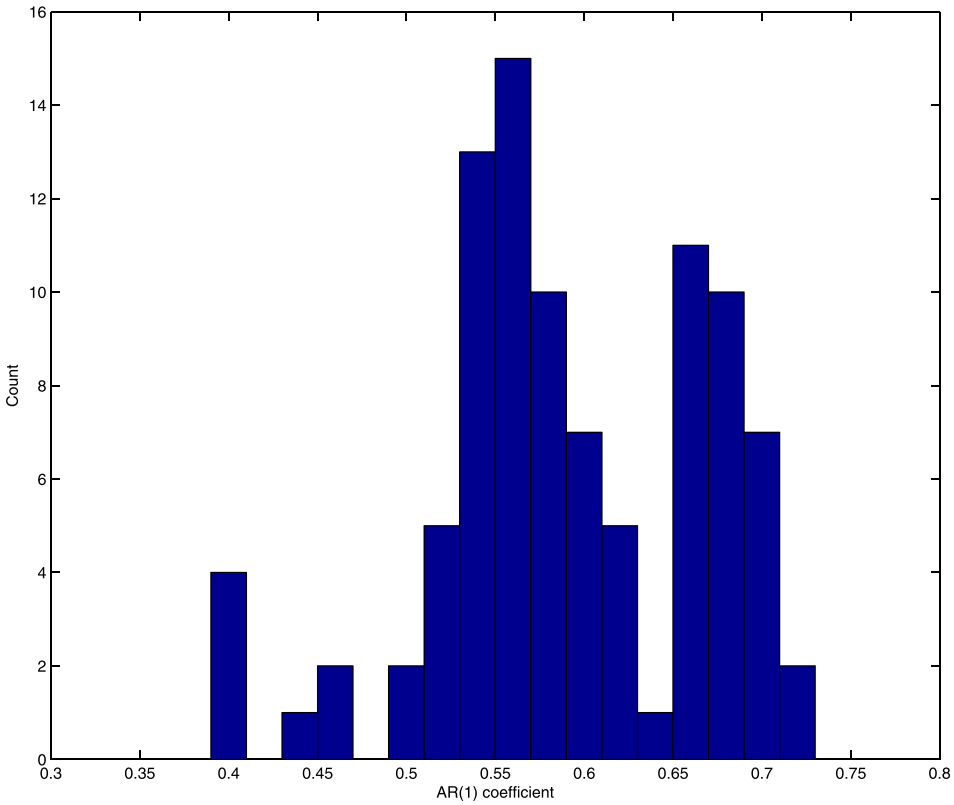

FIG. 2. Distribution of AR(1) coefficients of the 95 air pressure changes time series. The distribution indicates that most of the air pressure change time series have strong positive autocorrelations.

III. RESULTS

To illustrate the ARRMT method, we apply both RMT and ARRMT to multiple time series characterized by both cross-correlations and auto-correlations, data comprising 649 daily changes in atmospheric pressure $P_{i,t}$ for 95 different cities in the US, and defined as

$$R_{i,t} = P_{i,t} - P_{i,t-1}. \tag{12}$$

To demonstrate the advantage of using ARRMT over RMT, we apply RMT to air pressure changes and calculate the 95th percentile of the largest eigenvalues $\lambda_{+0.95} = 1.9174$ of the Wishart matrix using Eq. (2). Then we calculate the correlation matrix of empirical time series and the empirical

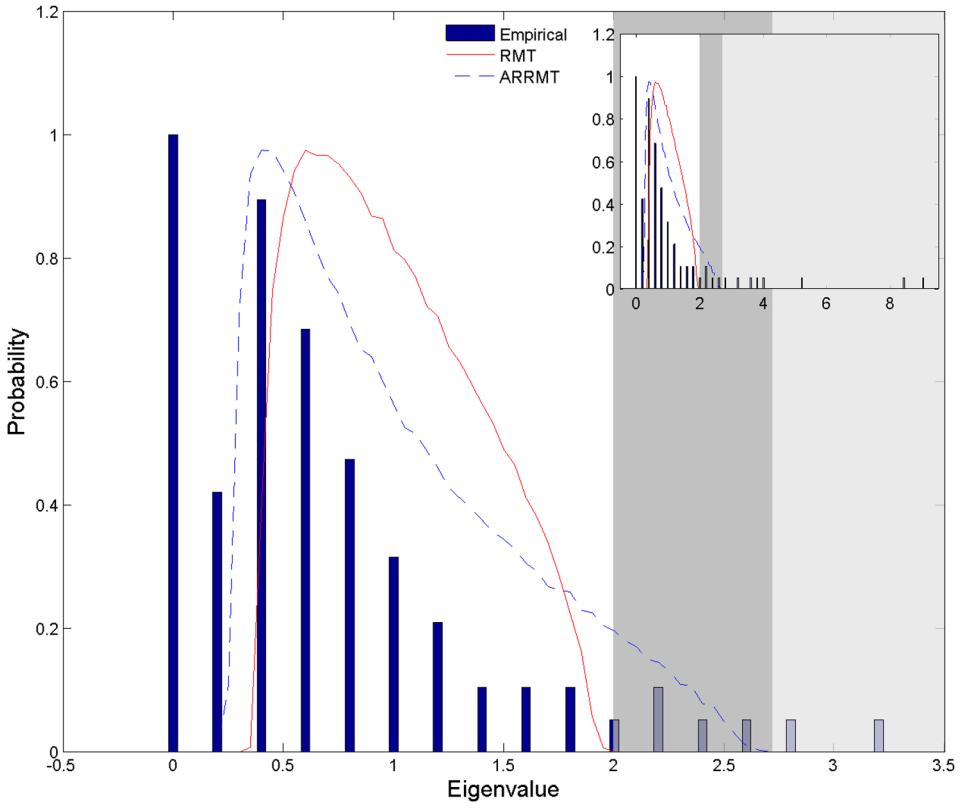

FIG. 3. Bar eigenvalue distribution of the correlation matrix for the air pressure changes of 95 US cities. Red solid line: the eigenvalue distribution of 95 simulated random time series repeated 1000 times. Blue dashed line: eigenvalue distribution of 95 simulated uncorrelated time series which has the same autocorrelations as the empirical time series repeated 1000 times. Using ARRMT, we find 8 empirical eigenvalues larger than λ'_+, indicating that there are only 8 factors that accounts for the air pressures in 95 US cities, as compared to 13 from RMT.

eigenvalues, among which the largest eigenvalue is $\lambda_L = 8.9740$ ($\gg\lambda_+$), indicating the existence of cross-correlations. We find that among the 20 eigenvalues there are 13 eigenvalues larger than λ_+, indicating that there are 13 significant factors influencing the air pressure in the 95 cities.

We next apply ARRMT by assuming that AR(1) of Eq. (6) is an appropriate candidate for modeling auto-correlations in the data. Using Eq. (6), we fit each of the 95 air pressure change time series R_t of Eq. (12). For each series $R_{i,t}$, we obtain the AR(1) coefficient ϕ_i. Figure 2 shows the distribution of AR(1) coefficients, which indicates that the auto-correlations are significant in most of the time series. We then generate 95 time series $Y'_{i,t}$ using the AR(1) model, each with a fitted value of ϕ_i. We next calculate the correlation matrix $\mathbf{W}'$ of the 95 generated time series $Y'_{i,t}$ and the eigenvalue distribution.

Figure 3 shows the largest eigenvalue for the Wishart matrix $\mathbf{W}$ and the autoregressive Wishart matrix $\mathbf{W}'$. As expected, due to the presence of auto-correlations in the data, we find that the 95th percentile of the largest eigenvalues of matrix $\mathbf{W}'$, $\lambda'_{+0.95} = 2.5922$, is larger than $\lambda_{+0.95} = 1.9174$ calculated for the Wishart matrix $\mathbf{W}$. We then compare $\lambda'_{+0.95}$ of Eq. (10) with the largest eigenvalues obtained for the empirical correlation matrix of the inflation rates and find that ARRMT reveals that there are only eight significant eigenvalues larger than λ'_+. Thus taking into account the presence of auto-correlations in the data, ARRMT finds that there are only eight factors that affect the changes in air pressure in the 95 cities.

In practice, when empirical data exhibit long memory auto-correlations AR(1) must be replaced by the more general AR(n) process $X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i}$. Here, we fit each time series with a higher-order AR(n) model and find that AR(10) fits the data better than AR(1). Applying the AR(10) model, we find that the largest eigenvalue is $\lambda'_+ = 2.823$. Although it is larger than the 2.592 value obtained by AR(1), the number of significant factors remains eight.

IV. CONCLUSIONS

In conclusion, we find that auto-correlations can significantly influence the eigenvalue distribution of the correlation matrix, and that RMT is therefore unreliable when analyzing cross-correlations in multiple time series when there are strong auto-correlations. To take into account the presence of auto-correlations in cross-correlated time series, we introduce the auto-regressive random matrix theory (ARRMT).

In ARRMT, we use a modified Wishart matrix that takes into account the auto-correlations commonly found in empirical data. The difference between the eigenvalue distributions of the empirical correlation matrix and modified Wishart matrix indicates the existence of the cross-correlations. We show that ARRMT is much more reliable method when auto-correlations exist in the studied atmospheric pressure data for 95 US cities.

ACKNOWLEDGMENTS

This work was supported by Project No. 71601112 by National Science Foundation of China, and Shanghai Pujiang Program via Grant No. 15PJC061. The Boston University work was supported by DOE Contract No. DE-AC07-05Id14517, and by NSF Grant Nos. CMMI 1125290, PHY 1505000, and CHE-1213217.

[1]P. Samuelsson *et al.*, Phys. Rev. Lett. **91**, 157002 (2003).
[2]A. Cottet *et al.*, Phys. Rev. Lett. **92**, 206801 (2004).
[3]I. Neder *et al.*, Phys. Rev. Lett. **98**, 036803 (2007).
[4]R. L. Weaver and O. I. Lobkis, Phys. Rev. Lett. **87**, 134301 (2001).
[5]P. Roux, K. G. Sabra, W. A. Kuperman, and A. Roux, J. Acoust. Soc. Am. **117**, 79 (2005).
[6]K. Yamasaki *et al.*, Phys. Rev. Lett. **100**, 228501 (2008).
[7]K. Wapenaar, Phys. Rev. Lett. **93**, 254301 (2004).
[8]M. Campillo and A. Paul, Science **299**, 547 (2003).
[9]L. Laloux *et al.*, Phys. Rev. Lett. **83**, 1467 (1999); V. Plerou *et al.*, *ibid.* **83**, 1471 (1999).
[10]R. N. Mantegna, Eur. Phys. J. B **11**, 193 (1999).
[11]L. Kullmann *et al.*, Phys. Rev. E **66**, 026125 (2002).
[12]I. T. Jolliffe, *Principal Component Analysis* (Springer-Verlag, New York, 1986).
[13]S. Arianos and A. Carbone, J. Stat. Mech. **03**, P03037 (2009).
[14]T. Guhr and B. Kalber, J. Phys. A **36**, 3009 (2003).
[15]B. Podobnik and H. E. Stanley, Phys. Rev. Lett. **100**, 084102 (2008).
[16]X. Zhang, B. Podobnik, D. Kenett, and H. E. Stanley, Physica A **415**, 43 (2014);.
[17]E. Wigner, Ann. Math. **62**, 548 (1955).
[18]M. L. Mehta, *Random Matrices* (Academic Press, 2004).
[19]V. Plerou *et al.*, Phys. Rev. E **65**, 066126 (2002).
[20]B. Podobnik *et al.*, Europhys. Lett. **90**, 68001 (2010).
[21]D. Wang *et al.*, Phys. Rev. E **83**, 046121 (2011).
[22]S. R. Bahcall, Phys. Rev. Lett. **77**, 5276 (1996).
[23]V. S. Rychkov *et al.*, Phys. Rev. Lett. **103**, 066602 (2009).
[24]T. Guhr *et al.*, Phys. Rep. **299**, 189 (1998).
[25]A. M. Sengupta and P. P. Mitra, Phys. Rev. E **60**, 3389 (1999).
[26]R. A. Fisher, Biometrika **10**(4), 507 (1915).
[27]J. D. Hamilton, *Time Series Analysis* (Princeton, New Jersey, 1994).
[28]V. A. Marchenko and L. A. Pastur, Mat. Sb. (N.S.) **72**(4), 507 (1967).
[29]M. S. Bartlett, Suppl. J. R. Stat. Soc. **8**(1), 27 (1946).
[30]G. V. Bayley and J. M. Hammersley, Suppl. J. R. Stat. Soc. **8**(2), 184 (1946).
[31]S. H. Simon and A. L. Moustakas, Phys. Rev. E **69**, 065101(R) (2004).
[32]G. M. Ljung and G. E. P. Box, Biometrika **65**(2), 297 (1978).
[33]H. Tong and K. S. Lim, J. R. Stat. Soc., Ser. B **42**, 245 (1980).