This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2015.2433920, IEEE Access

Massive MIMO as a Big Data System: Random Matrix Models and Testbed

Changchun Zhang, *Student Member, IEEE*, Robert C. Qiu, *Fellow, IEEE*

Abstract

The paper has two parts. The first one deals with how to use large random matrices as building blocks to model the massive data arising from the massive (or large-scale) MIMO system. As a result, we apply this model for distributed spectrum sensing and network monitoring. The part boils down to the streaming, distributed massive data, for which a new algorithm is obtained and its performance is derived using the central limit theorem that is recently obtained in the literature. The second part deals with the large-scale testbed using software-defined radios (particularly USRP) that takes us more than four years to develop this 70-node network testbed. To demonstrate the power of the software defined radio, we reconfigure our testbed quickly into a testbed for massive MIMO. The massive data of this testbed is of central interest in this paper. It is for the first time for us to model the experimental data arising from this testbed. To our best knowledge, we are not aware of other similar work.

Index Terms

Massive MIMO, 5G Network, Random Matrix, Testbed, Big Data.

I. Introduction

Massive or large-scale multiple-input, multiple output (MIMO), one of the disruptive technologies of the next generation (5G) communications system, promises significant gains in wireless data rates and link reliability [1] [2]. In this paper, we deal with the massive data aspects of the massive MIMO system. In this paper, we use two terms (massive data and big data) interchangeably, following the practice from National Research Council [3].

The benefits from massive MIMO are not only limited to the higher data rates. Massive MIMO techniques makes green communications possible. By using large numbers of antennas at the base station, massive MIMO helps to focus the radiated energy toward the intended direction while minimizing the intra and intercell interference. The energy efficiency is increased dramatically as the energy can be focused with extreme sharpness into small regions in space [4]. It is shown in [5] that, when the number of base station (BS) antennas M grows without bound, we can reduce the transmitted power of each user proportionally to $1/M$ if the BS has perfect channel state information (CSI), and proportionally to $\frac{1}{\sqrt{M}}$ if CSI is estimated from uplink pilots. Reducing the transmit power of the mobile users can drain the batteries slower. Reducing the RF power of downlink can cut the electricity consumption of the base station.

Massive MIMO also brings benefits including inexpensive low-power components, reduced latency, simplification of MAC layer, etc [4]. Simpler network design could bring lower complexity computing which save more energy of the network to make the communications green.

Currently, most of the research of massive MIMO is focused on the communications capabilities. In this paper, we promote an insight that, very *naturally*, the massive MIMO system can be regarded as a big data system. Massive waveform data—coming in a streaming manner—can be stored and processed at the base station with a large number of antennas, while not impacting the communication capability. Especially, the random matrix theory can be well mapped to the architecture of large array of antennas. The random matrix theory data model has ever been validated by [6] in a context of distributed sensing. In this paper, we extend this work to the massive MIMO testbed. In particular, we studied the function of multiple non-Hermitian random matrices and applied the variants to the experimental data collected on the massive MIMO testbed. The product of non-Hermitian random matrices shows encouraging potential in signal detection, that is motivated for spectrum sensing and network monitoring. We also present two concrete applications that are demonstrated on our testbed using the massive MIMO system as big data system. From the two applications, we foresee that, besides signal detection, the random-matrix based big data analysis will drive more mobile applications in the next generation wireless network.

II. Modeling for Massive Data

Large random matrices are used models for the massive data arising from the monitoring of the massive MIMO system. We give some tutorial remarks, to facilitate the understanding of the experimental results.

Robert C. Qiu and Changchun Zhang are with the Department of Electrical and Computer Engineering, Center for Manufacturing Research, Tennessee Technological University, Cookeville, TN, 38505, e-mail: czhang42@students.tntech.edu; rqiu@ieee.org.

A. Data Modeling with Large Random Matrices

Naturally, we assume n observations of p-dimensional random vectors $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{C}^{p\times 1}$. We form the data matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbb{C}^{p\times n}$, which naturally, is a random matrix due to the presence of ubiquitous noise. In our context, we are interested in the practical regime $p = 100 - 1,000$, while n is assumed to be arbitrary. The possibility of arbitrary sample size n makes the classical statistical tools infeasible. We are asked to consider the asymptotic regime [7]–[10]

$$p \to \infty, n \to \infty, p/n \to c \in (0, \infty)\,, \tag{1}$$

while the classical regime [11] considers

$$p \text{ fixed}, n \to \infty, p/n \to 0. \tag{2}$$

Our goal is to reduce massive data to a few statistical parameters. The first step often involves the covariance matrix estimation using the sample covariance estimator

$$\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}^H = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^H \in \mathbb{C}^{p\times p}, \tag{3}$$

that is a sum of rank-one random matrices [12]. The sample covariance matrix estimator is the maximum likelihood estimator (so it is optimal) for the classical regime (2). However, for the asymptotic regime (1), this estimator is *far from optimal*. We still use this estimator due to its special structure. See [7]–[10] for modern alternatives to this fundamental algorithm. For brevity, we use the sample covariance estimator throughout this paper.

B. Non-Hermitian Free Probability Theory

Once data are modeled as large random matrices, it is natural for us to introduce the non-Hermitian random matrix theory into our problem at hand. Qiu's book [9] gives an exhaustive account of this subject in an independent chapter, from a mathematical view. This paper is complementary to our book [9] in that we bridge the gap between theory and experiments. We want to understand how accurate this theoretical model becomes for the real-life data. See Section V for details.

Roughly speaking, large random matrices can be treated as free matrix-valued random variables. "Free" random variables can be understood as independent random variables. The matrix size must be so large that the asymptotic theoretical results are valid. It is of central interest to understand this finite-size scaling in this paper.

III. Distributed Spectrum Sensing

Now we are convinced that large random matrices are valid for experimental data modeling. The next natural question is to test whether the signal or the noise is present in the data. Both networking monitoring and spectrum sensing can be formulated as a matrix hypothesis testing problem for anomaly detection.

A. Related Work

Specifically, consider the n samples $\mathbf{y}_1, ..., \mathbf{y}_n$, drawn from a p-dimensional complex Gaussian distribution with covariance matrix $\mathbf{\Sigma}$. We aim to test the hypothesis:

$$\mathcal{H}_0 : \mathbf{\Sigma} = \mathbf{I}_p.$$

This test has been studied extensively in classical settings (i.e., p fixed, $n \to \infty$), first in detail in [13]. Denoting the sample covariance by $\mathbf{S}_n = \frac{1}{n}\sum\limits_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^H$, the LRT is based on the linear statistic (see Anderson (2003) [11, Chapter 10])

$$L = \mathrm{Tr}\left(\mathbf{S}_n\right) - \ln\left(\det \mathbf{S}_n\right) - p. \tag{4}$$

Under $\mathcal{H}_0$, with p fixed, as $n \to \infty$, nL is well known to follow a χ^2 distribution. However, with high-dimensional data for which the dimension p is large and comparable to the sample size n, the χ^2 approximation is no longer valid. A correction to the LRT is done in Bai, Jiang, Yao and Zheng (2009) [14] on large-dimensional covariance matrix by random matrix theory. In this case, a better approach is to use results based on the double-asymptotic given by Assumption 1. Such a study has been done first under $\mathcal{H}_0$ and later under the spike alternative $\mathcal{H}_1$. More specifically, under $\mathcal{H}_0$, this was presented in [14] using a CLT framework established in Bai and Silverstein (2004) [15]. Under "$\mathcal{H}_1 : \mathbf{\Sigma}$ has a spiked covariance structure as in Model A", this problem was addressed only very recently in the independent works, [16] and [17]. We point out that [16] (see also [18]) considered a generalized problem which allowed for multiple spiked eigenvalues. The result in [16] was again based on the CLT framework of Bai and Silverstein (2004) [15], with their derivation requiring the calculation of contour integrals. The same result was presented in [17], in this case making use of sophisticated tools of contiguity and Le Cam's first and third lemmas [19].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2015.2433920, IEEE Access

B. Spiked Central Wishart Matrix

Our problem is formulated as

$$\begin{aligned}&\mathcal{H}_0 : \mathbf{\Sigma} = \mathbf{I}_p\\ &\mathcal{H}_1 : \mathbf{\Sigma} \in \text{Model A: Spiked central Wishart.}\end{aligned} \tag{5}$$

Model A: Spiked central Wishart: Matrices with distribution $\mathcal{CW}_p\left(n, \mathbf{\Sigma}, \mathbf{0}_{p\times p}\right)(n \geqslant p)$, where $\mathbf{\Sigma}$ has multiple distinct "spike" eigenvalues $1+\delta_1 > \cdots > 1+\delta_r$, with $\delta_r > 0$ for all $1 \leq k \leq r$, and all other eigenvalues equal to 1.

Assumption 1. $n, p \to \infty$ such that $n/p \to c \geqslant 1$.

Theorem III.1 (Passemier, McKay and Chen (2014) [20])**.** *Consider Model A and define*

$$a = \left(1-\sqrt{c}\right)^2, \quad b = \left(1+\sqrt{c}\right)^2. \tag{6}$$

Under Assumption 1, for an analytic function $f : \mathcal{U} \to \mathbb{C}$ where $\mathcal{U}$ is an open subset of the complex plane which contains $[a, b]$, we have

$$\sum_{i=1}^{p} f\left(\frac{\lambda_i}{p}\right) - p\mu \xrightarrow{\mathcal{L}} \mathcal{N}\left(\sum_{\ell=1}^{r} \bar{\mu}\left(z_{0,\ell}\right), \sigma^2\right),$$

where

$$\mu = \frac{1}{2\pi}\int_a^b f(x)\,\frac{\sqrt{(b-x)(x-a)}}{x}dx \tag{7}$$

$$\sigma^2 = \frac{1}{2\pi^2}\int_a^b \frac{f(x)}{\sqrt{(b-x)(x-a)}}\left[\mathcal{P}\int_a^b \frac{f'(y)\sqrt{(b-y)(y-a)}}{x-y}dy\right]dx \tag{8}$$

with these terms independent of the spikes. The spike-dependent terms $\bar{\mu}\left(z_{0,\ell}\right), 1 \leqslant \ell \leqslant r$ admit

$$\bar{\mu}\left(z_{0,\ell}\right) = \frac{1}{2\pi}\int_a^b \frac{f(x)}{\sqrt{(b-x)(x-a)}}\left[\frac{\sqrt{\left(z_{0,\ell}-a\right)\left(z_{0,\ell}-b\right)}}{z_{0,\ell}-x} - 1\right]dx \tag{9}$$

where

$$z_{0,\ell} = \begin{cases}\frac{(1+c\delta_\ell)(1+\delta_\ell)}{\delta_\ell}, & \textit{for Model A}\\ \frac{(1+\nu_\ell)(1+\nu_\ell)}{\nu_\ell}, & \textit{for Model B}\end{cases}.$$

The branch of the square root $\sqrt{\left(z_{0,\ell}-a\right)\left(z_{0,\ell}-b\right)}$ is chosen.

As an application of Theorem III.1 for Model A, we consider the classical LRT that the population covariance matrix is the identity, under a rank-one spiked population alternative.

Here, we will adopt our general framework to recover the same result as [16] and [17] very efficiently, simply by calculating a few integrals. Under $\mathcal{H}_1$, as before we denote by $1+\delta$ the spiked eigenvalue of $\mathbf{\Sigma}$. Since $n\mathbf{S}_n \sim \mathcal{CW}_p\left(n, \mathbf{\Sigma}, \mathbf{0}_{p\times p}\right)$, we now apply Theorem III.1 for the case of Model A to the function

$$f_L(x) = \frac{x}{c} - \ln\left(\frac{x}{c}\right) - 1.$$

Let $\lambda_i, 1 \leqslant i \leqslant p$, be the eigenvalues of $n\mathbf{S}_n$. Since the domain of definition of f_L is $(0, \infty)$, we assume that $c > 1$ to ensure $a > 0$ (see (6)). Then, under Assumption 1,

$$L = \sum_{i=1}^{p} f_L\left(\frac{\lambda_i}{p}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(p\mu + \bar{\mu}, \sigma^2\right),$$

$$L = \sum_{i=1}^{p} f_L\left(\frac{\lambda_i}{p}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(p\mu + \bar{\mu}\left(z_{0,1}\right), \sigma^2\right),$$

where $r = 1$ is used for one spike to obtain

$$\mu = 1 + (c-1)\ln\left(1-c^{-1}\right), \quad \sigma^2 = -c^{-1}\ln\left(1-c^{-1}\right)$$

with the spike-dependent term

$$\bar{\mu} = \delta_1 - \ln\left(1+\delta_1\right).$$

The special case of one spike is also considered in [21]. These results are in agreement with [16] and [17].

C. Distributed Streaming Data

For each server, equation (5) formulates the testing problem. How do we formulate this problem when the data are spatially distributed across N servers? Our proposed algorithm is as follows: **Algorithm 1**

1) The i-th server computes the sample covariance matrix $\mathbf{S}_i, i = 1, ..., N$.
2) The i-th server computes the linear statistic
$$L_i = \mathrm{Tr}\left(\mathbf{S}_i\right) - \ln\left(\det \mathbf{S}_i\right) - p, i = 1, ..., N.$$
3) The i-th server communicates the linear statistic $L_i, i = 1, ..., N$ to one server that acts as the coordinator.
4) Finally, the coordinator server obtains the linear statistic $L_i, i = 1, ..., N$ via communication and sum up the values $L_D = L_1 + \cdots + L_N$.
5) All the above computing and communication are done in in parallel.

The communication burden is very low. The central ingredient of Algorithm 1 is to exploit the Central Limit Theorem of the used linear statistic L defined in (4). By means of Theorem III.1, we have

$$L = \sum_{i=1}^{p} f\left(\frac{\lambda_i}{p}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(p\mu + \sum_{\ell=1}^{r} \bar{\mu}\left(z_{0,\ell}\right), \sigma^2\right).$$

Since $L_1, ..., L_N$ are Gaussian random variables, the sum of Gaussian random variables are also Gaussian; thus $L_D = L_1 + \cdots + L_N$ is also Gaussian, denoted as $\mathcal{N}\left(\mu_D, \sigma_D^2\right)$.

The false alarm probability for the linear statistic can be obtained using standard procedures. If $L_D > \gamma$, the signal is present; otherwise, the signal does not exist. The false alarm probability is

$$\begin{aligned} P_{fa} = \mathbb{P}\left(L > \gamma \left| \mathcal{H}_0 \right.\right) &= \mathbb{P}\left(\tfrac{L-\mu_D}{\sigma_D} > \tfrac{\gamma-\mu_D}{\sigma_D} \left| \mathcal{H}_0 \right.\right) \\ &= \int_{\frac{L-\mu_D}{\sigma_D}}^{\infty} \tfrac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right) dt \\ &= Q\left(\tfrac{L-\mu_D}{\sigma_D}\right) \end{aligned}$$

where $Q\left(x\right) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right) dt$. For a desired false-alarm rate ε, the associated threshold should be chosen such that

$$\gamma = \mu_D + \frac{1}{\sigma_D} Q^{-1}\left(\varepsilon\right).$$

To predict the detection probability, we need to know the distribution of ξ under $\mathcal{H}_1$, which has been obtained using Theorem III.1. The detection probability is calculated as

$$\begin{aligned} P_d = \mathbb{P}\left(L_D > \gamma \left| \mathcal{H}_1 \right.\right) &= \mathbb{P}\left(\tfrac{\xi-\mu_D}{\sigma_D} > \tfrac{\gamma-\mu_D}{\sigma_D} \left| \mathcal{H}_1 \right.\right) \\ &= Q\left(\tfrac{L_D-\mu_D}{\sigma_D}\right). \end{aligned}$$

IV. Massive MIMO Testbed and Data Acquisition

A. System Architecture and Signal Model

The system architecture of the testbed is as Fig. 1.

The general software-defined radio (SDR) universal software radio peripheral (USRP) platform is used to emulate the base station antenna in our testbed. We deployed up to 70 USRPs and 30 high performance PCs to work collaboratively as an large antenna array of the massive MIMO base station. These USRPs are well clock synchronized by an AD9523 clock distribution board. The system design of this testbed can be found in [22].

Our testbed has demonstrated initial capabilities as below:

a) Channel Reciprocity for Channel Measurement: Channel matrix measurement is a critical task for Multi-User Massive MIMO system. For the antenna i and j, if the uplink and downlink work in TDD mode, the channel reciprocity will be useful for the pre-coding in MIMO system. Channel reciprocity means $h_{i,j} = h_{j,i}$ if $h_{i,j}$ represents the air channel from antenna i to antenna j and vice versa.

Given the h is the air channel between antenna i and j, the measured channel $h_{i,j}$ and $h_{j,i}$ follow the model depicted as Fig. 2, where where $T\left(i\right)$, $R\left(j\right)$, $R\left(i\right)$, $T\left(j\right)$ represent the effect from circuits like upper/down conversion, filters, etc., for both the upper and down links.

Thus we have

$$\begin{aligned} h_{i,j} &= T\left(i\right) \cdot h \cdot R\left(j\right) \\ h_{i,j} &= R\left(i\right) \cdot h \cdot T\left(j\right) \end{aligned} \tag{10}$$

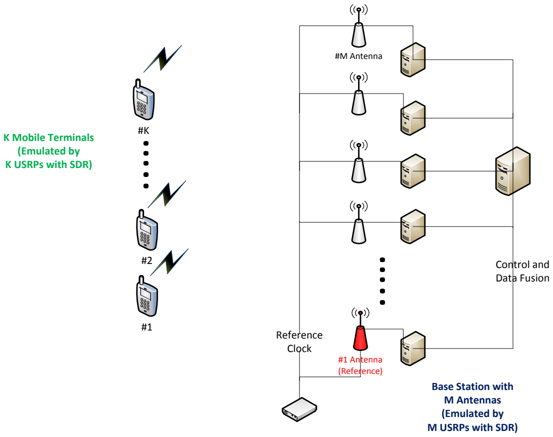

Fig. 1: System Architecture of Multi-User Massive MIMO Testbed.

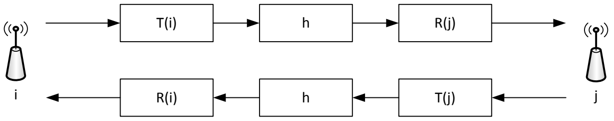

Fig. 2: Reciprocity mode for TDD channel

Usually, the relative calibration is sufficient for the pre-coding as we have

$$\frac{h_{i,j}}{h_{j,i}} = \frac{T(i) \cdot R(j)}{R(i) \cdot T(j)} \tag{11}$$

which is constant in ideal situation.

Channel reciprocity described above includes the circuits impact. Our measurement shows that ratio $h_{i,j}/h_{j,i}$ between the downlink and uplink channel frequency response for antenna i and j is almost constant. For example, we collect 3 rounds of data within a time duration that the channel can be regarded as static. Thus 3 such ratios are obtained for a specified link between USPR node transmitting antenna 3 and receiving antenna 2. Three absolute values of the sampled ratio for $h_{3,2}/h_{2,3}$ are 1.2486, 1.22, 1.2351 respectively.

b) Massive Data Acquisition for Mobile Users or Commercial Networks: Consider the time evolving model described as following. Let N be the number of antennas at base station. All the antennas start sensing at the same time. Every time, on each antenna, a time series with samples length T is captured and denoted as $x_i \in \mathbb{C}^{1\times T}, i = 1, \ldots, N$. Then a random matrix from N such vectors are formed as:

$$\mathrm{X}_j = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}_{N\times T} \tag{12}$$

where, $j = 1, \cdots, L$. Here L means we repeat the sensing procedure with L times. Then L such random matrices are obtained. In the following sections, we are interested in variant random matrix theoretical data models, including the product of the L random matrices and their geometric/arithmetic mean. We call it time evolving approach.

Besides the time evolving approach, we can also use a different data format to form random matrix. Suppose we select n receivers at Massive MIMO base station. At each receiver, we collect $N \times T$ samples to get a random matrix $\mathrm{X}_i \in \mathbb{C}^{N\times T}$ with $i = 1, \cdots, n$. Similarly, we are interested in the functions of these random matrices. We call it space distributed approach.

In the next section, we specify which approach is used to form the random matrix when a certain theoretical model is used.

V. Random Matrix Theoretical Data Model and Experimental Validation

We are interested in the eigenvalue distribution for every data model. The results obtained from the experimental data are compared with theoretical distributions (if exists). The experimental data come from noise-only case and signal-present case. Our testbed captures the commercial signal data at $869.5MHz$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2015.2433920, IEEE Access

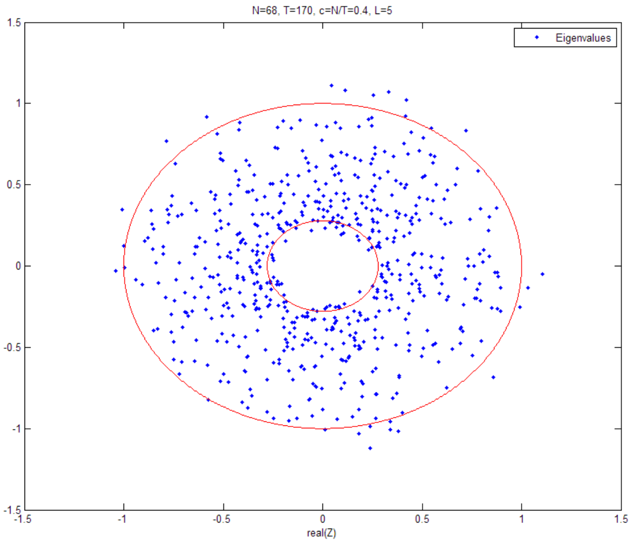

Fig. 3: The eigenvalue distribution for product of non-Hermitian random matrices, noise only, all snapshots.

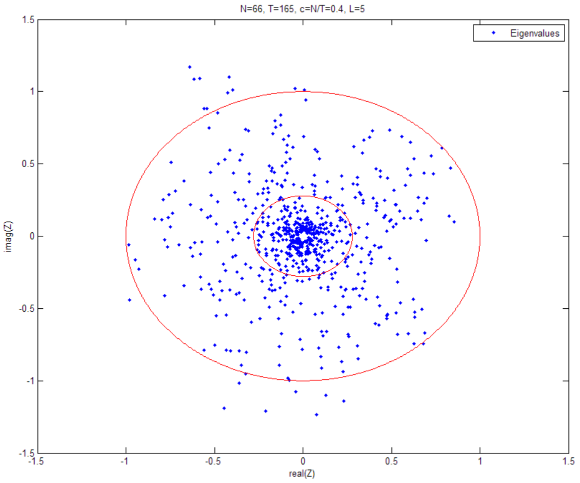

Fig. 4: The eigenvalue distribution for product of non-Hermitian random matrices, signal present, all snapshots.

A. Product of non-Hermitian random matrix

The eigenvalue distribution for the product of non-Hermitian random matrix, so far, gives us the best visible information to differentiate the situations of noise only and signal present. Here the timing evolving approach is used. Denote the product of non-Hermitian random matrix as:

$$\mathrm{Z} = \prod_{j=0}^{L} \mathrm{X}_j \tag{13}$$

In the experiment, L is adjustable. In addition, a number of such Z are captured with time evolving, to investigate if the pattern is changing or not with time. Every Z could be regarded as one snapshot. For both the noise and signal experiments, we took 10 snapshots. All the 10 snapshots are put together to show eigenvalue distribution more clearly.

1) Eigenvalue Distributions for Noise-Only and Signal-Present: Firstly, we visualize the eigenvalue distribution on the complex plane to see the difference between the cases of noise-only and signal-present.

Noise Only: If the eigenvalue distribution for all the snapshots are put together, we see Fig. 3, in which the red circle represents the "Ring Law".

Signal Present: If putting together the eigenvalues of all snapshots, we see Fig. 4, in which the inner radius of the eigenvaule distribution is smaller than that of the ring law.

We also use the probability density diagram to show the difference between noise only and signal present cases, with different L. The theorem V.1 actually gives the theoretical values of the inner radius and outer radius of the ring law.

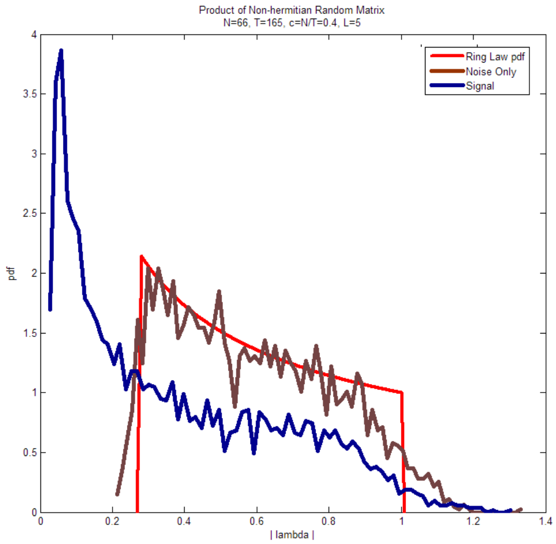

Fig. 5: Probability of eigenvalue for product of the non-Hermitian random matrix, both cases, with $L = 5$.

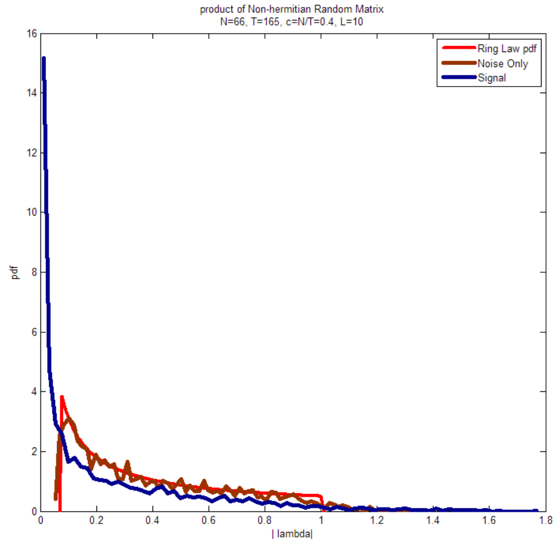

Fig. 6: Probability of eigenvalue for product of the non-Hermitian random matrix, both cases, with $L = 10$.

Theorem V.1. *The empirical eigenvalue distribution of $N \times T$ matrix $\prod_{i=1}^{L} \mathrm{X}_i$ converge almost surely to the same limit given by*

$$f_{\prod_{i=1}^{L} \mathrm{X}_i}(\lambda) = \begin{cases} \frac{2}{\pi c L} |\lambda|^{2/L-2} & (1-c)^{L/2} \leqslant r \leqslant 1 \\ 0 & \textit{elsewhere} \end{cases}$$

as N, $n \to \infty$ with the ratio $c = N/n \leqslant 1$ fixed.

We are interested in the probability density of $|\lambda|$. Let $r = |\lambda|$, which is described in Eq. 14, derived from the Theorem V.1.

$$f_{\prod_{i=1}^{L} \mathrm{X}_i}(r) = \begin{cases} \frac{2}{cL} r^{\frac{2}{L}-1} & (1-c)^{L/2} \leqslant r \leqslant 1 \\ 0 & \textit{elsewhere} \end{cases} \tag{14}$$

The PDF is also shown in Fig. 5 and Fig. 6 with different L.

The above results show that eigenvalue distribution follows the ring law in this model for noise only case. The signal present case also has the ring law while the inner radius is much smaller than the noise only case, especially when L is large.

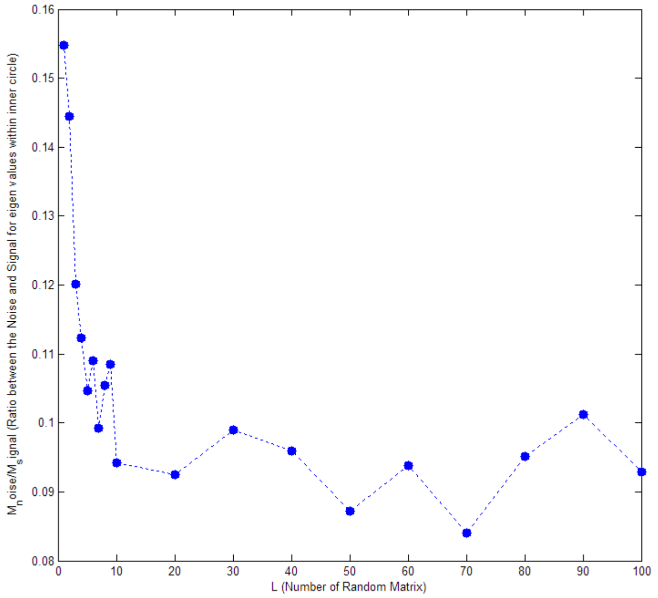

Fig. 7: Shrinking eigenvalue ratio within the ring law inner circle between the noise only and signal present cases.

2) Empirical Effect of L to Differentiate Cases of Noise only and Signal Present: Regarding the product of non-Hermitian random matrices, the main difference observed in cases of noise only and signal present, is about the inner circle radius of the eigenvalue distribution.

According to the ring law, the inner circle radius of the eigenvalue distribution for the noise only case, is constrained by Eq. 15, which is a fixed value for determined L and c.

$$r_{\text{inner}} = (1-c)^{\frac{L}{2}} \tag{15}$$

Meanwhile, the radius shrinks for the case of the signal being present. In addition, for both cases, the inner circle radius decreases with increasing L. The question is whether it is easier to differentiate the two cases when increasing the value L?

For the same L, we define $M_{\text{noise}}\left(L\right)|_{r<r_{\text{inner}}}$ as the number of eigenvalues falling within the ring law inner circle, measured for the noise only case, and the $M_{\text{signal}}\left(L\right)|_{r<r_{\text{inner}}}$ as the number of eigenvalues falling within the inner circle of the ring law, measured for signal present case. Thus, we have a ratio denoted as

$$\rho\left(L\right) = \frac{M_{\text{noise}}\left(L\right)|_{r<r_{\text{inner}}}}{M_{\text{signal}}\left(L\right)|_{r<r_{\text{inner}}}} \tag{16}$$

to represent the impact of L.

Fig. 7 show the trend of the ratio with increasing L. Generally, the ratio decreases with the increasing L, indicating that the larger L brings better distance between the cases of noise only and signal present. However, the trend is very similar with the negative exponential function of L. When L is greater than 10, the ratio does not change much.

B. Geometric Mean

Using the same data as last paragraph, the geometric mean of the non-Hermitian random matrix can be obtained as:

$$\mathrm{Z} = \left(\prod_{j=0}^{L} \mathrm{X}_j\right)^{1/L} \tag{17}$$

Time evolving approach is used here. In this experiment, we adjust the L and the convergence is observed when L is increased. All the diagrams below include 10 snapshots of data results. Basically, in this case, the eigenvalues converge to the outer unit circle and are not changing much with increasing L.

Noise Only: Fig. 8 to Fig. 10 show the eigenvalue distribution of the geometric mean for noise only situation.

Signal Present: Fig. 11 to Fig. 13 show the eigenvalue distribution of the geometric mean for signal situation. Different with noise case, the convergence of the eigenvalue is sensitive to the value of L. With bigger L, the distribution converges more to the unit circle.

We also show the PDF of the eigenvalue absolute values for geometric mean, in Fig. 14 and Fig. 15 with different L.

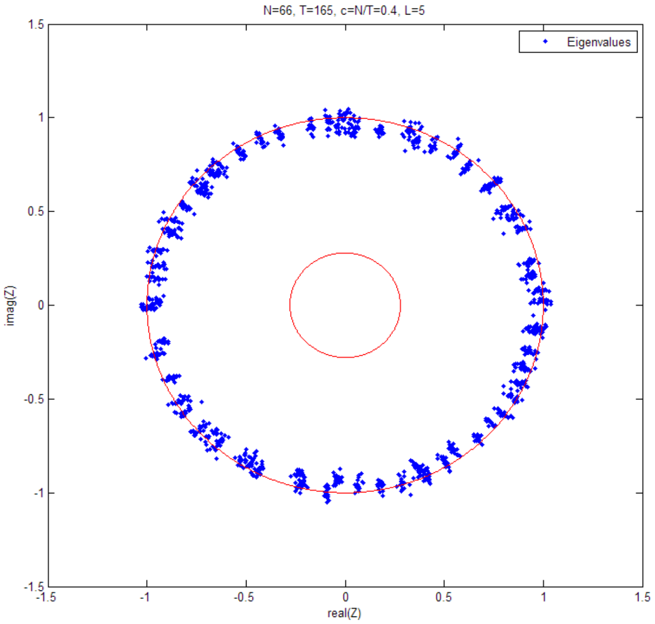

Fig. 8: The eigenvalue distribution for geometric mean of non-Hermitian random matrix, noise only, all snapshots, L=5.

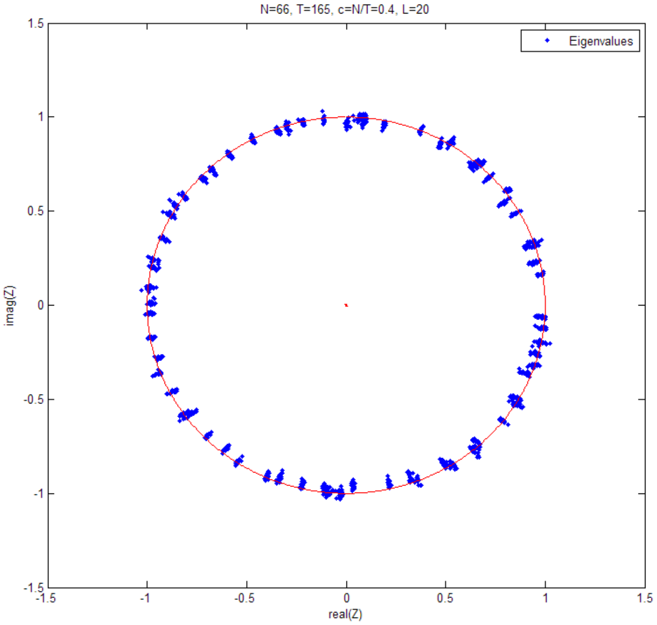

Fig. 9: The eigenvalue distribution for geometric mean of non-Hermitian random matrix, noise only, all snapshots, L=20.

From all the visualized results for the Geometric mean model, we see that

- the eigenvalue distribution is similar to the ring law, but the radius is not the same as product of non-Hermitian random matrices.
- the difference between inner radius and the outer radius, for the signal-present case, is larger than that for noise-only case.
- with L increased, the "ring" is converged more to the outer circle. The absolute difference between noise-only and signal-present is actually not get larger with increasing L.

C. Arithmetic Mean:

The arithmetic mean of the non-Hermitian random matrix is defined as

$$\mathrm{Z} = \frac{1}{L}\left(\sum_{j}^{L} \mathrm{X}_j\right) \tag{18}$$

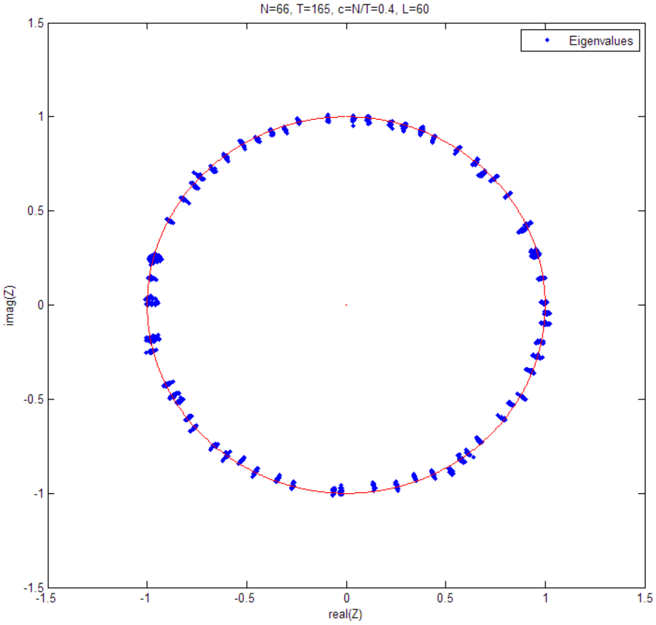

Fig. 10: The eigenvalue distribution for geometric mean of non-Hermitian random matrix, noise only, all snapshots, L=60.

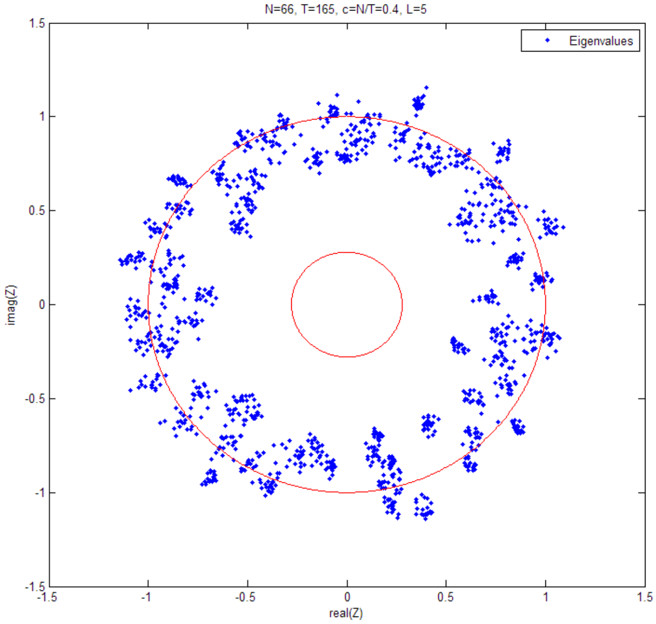

Fig. 11: The eigenvalue distribution for geometric mean of non-Hermitian random matrix, signal present, all snapshots, L=5.

For both the noise-only and signal-present cases, we adjust the value of L to see the effect. We select $L = 5, 20, 100$.

Noise Only: Fig. 16 to Fig. 18 show the eigenvalue distribution of the arithmetic mean of the L non-Hermitian random matrix, for the noise only case.

Signal Present: Fig. 19 to Fig. 21 show the eigenvalue distribution of the arithmetic mean of the L non-Hermitian random matrix, for the signal present case.

The corresponding PDFs of the eigenvalue absolute values of arithmetic mean are also shown Fig. 22 and Fig. 23.

From the visualized results of the eigenvalue distribution for Arithmetic mean model, we see

- The eigenvalue distribution for either noise-only and signal-present is following a similar ring law.
- The width of the ring, for signal-present, is larger than that for noise-only.
- We cannot get extra benefit by increasing L, as the width of the ring is not impacted by L.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2015.2433920, IEEE Access

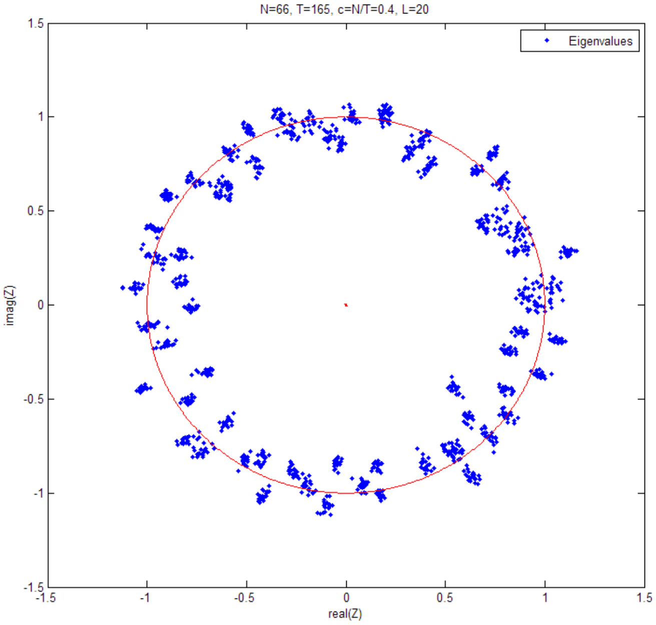

Fig. 12: The eigenvalue distribution for geometric mean of non-Hermitian random matrix, signal present, all snapshots, L=20.

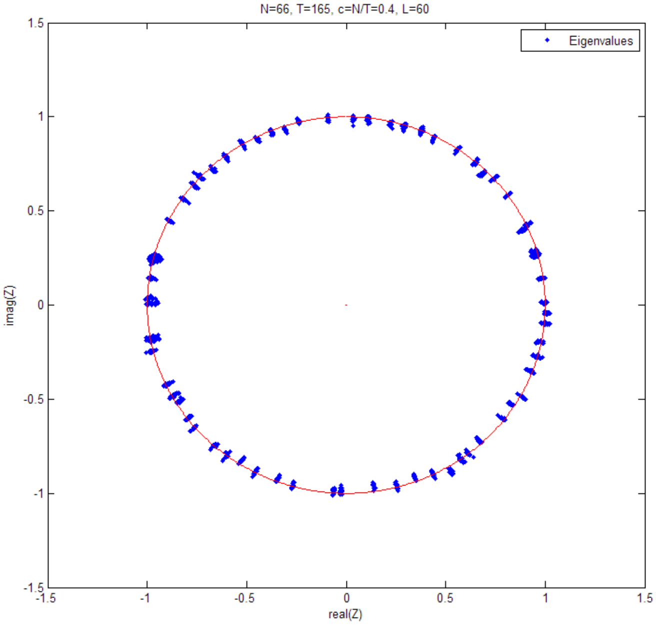

Fig. 13: The eigenvalue distribution for geometric mean of non-Hermitian random matrix, signal present, all snapshots, L=60.

D. Product of Random Ginibre Matrices

We study the product of k independent random square Ginibre matrices, $\mathrm{Z} = \prod\limits_{1}^{k} \mathrm{G}_i$. When the random Ginibre martices, G_i, are square, the eigenvalues of ZZ^H have asymptotic distribution $\rho^{(k)}\left(x\right)$ in the large matrix limit. In terms of free probability theory, it is the free multiplicative convolution product of k copies of the Marchenko-Pastur distribution. In this model, we applied the space distributed approach to for the random matrix.

For $k = 2$, the spectral density is explicitly given by

$$\rho^{(2)}\left(x\right) = \frac{2^{1/3}\sqrt{3}}{12\pi} \frac{\left[2^{1/3}\left(27 + 3\sqrt{81 - 12x}\right)^{2/3} - 6x^{1/3}\right]}{x^{2/3}\left(27 + 3\sqrt{81 - 12x}\right)^{1/3}} \tag{19}$$

where $x \in [0, 27/4]$. For general k, the explicit form of the distribution is a superposition of hyper-geometric function of the

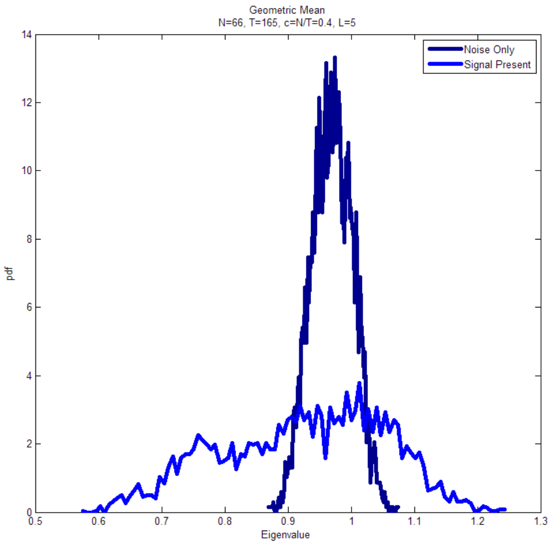

Fig. 14: Probability of eigenvalue for geometric mean of the non-Hermitian random matrix, both cases, with $L = 5$.

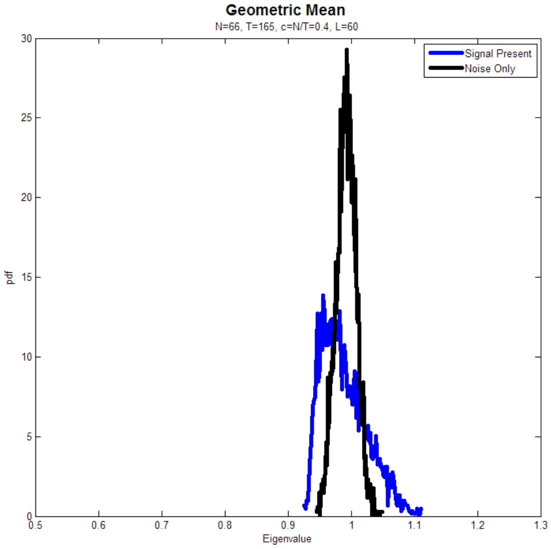

Fig. 15: Probability of eigenvalue for geometric mean of the non-Hermitian random matrix, both cases, with $L = 60$.

type ${}_kF_{k-1}$

$$\rho^{(k)}\left(x\right) = \sum_{i=1}^{k} \Lambda_{i,k} x^{\frac{i}{k+1}-1}.$$

$${}_kF_{k-1}\left(\left[\left\{a_j\right\}_{j=1}^{k}\right];\left[\left\{b_j\right\}_{j=1}^{i-1},\left\{b_j\right\}_{j=i+1}^{k}\right];\frac{k^k}{\left(k+1\right)^{k+1}}x\right) \tag{20}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2015.2433920, IEEE Access

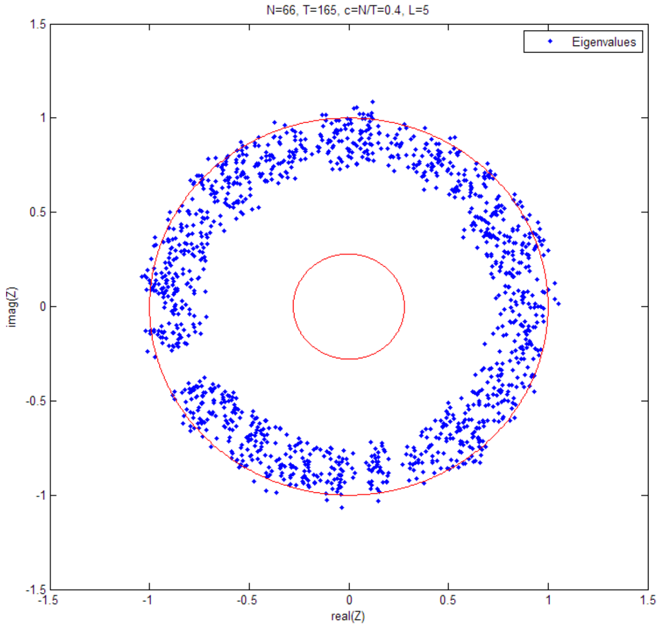

Fig. 16: The eigenvalue distribution for arithmetic mean of non-Hermitian random matrix, noise only, L=5.

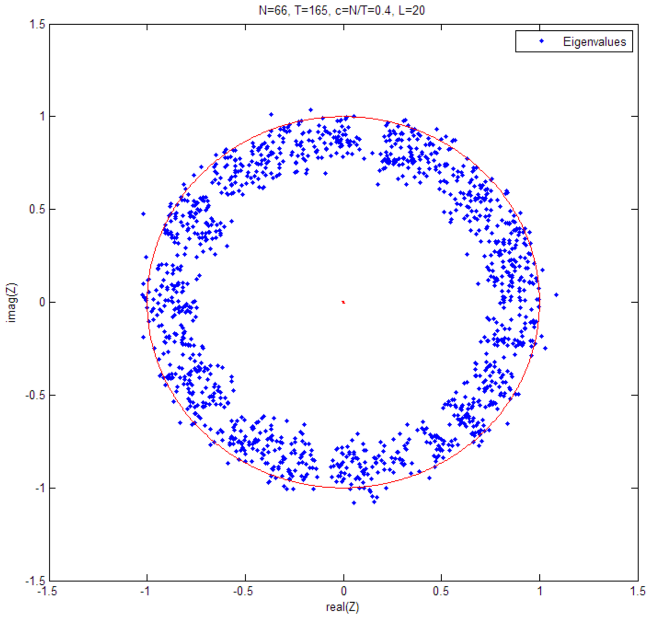

Fig. 17: The eigenvalue distribution for arithmetic mean of non-Hermitian random matrix, noise only, L=20.

where $a_j = 1 - \frac{1+j}{k} + \frac{i}{k+1}$, $b_j = 1 + \frac{i-j}{k+1}$, and

$$\Lambda_{i,k} = \frac{1}{k^{3/2}}\sqrt{\frac{k+1}{2\pi}}\left(\frac{k^{k/(k+1)}}{k+1}\right)^i \cdot$$

$$\frac{\left[\prod_{j=1}^{i-1}\Gamma\left(\frac{j-i}{k+1}\right)\right]\left[\prod_{j=k+1}^{k}\Gamma\left(\frac{j-i}{k+1}\right)\right]}{\prod_{j=1}^{k}\Gamma\left(\frac{j+1}{k}-\frac{i}{k+1}\right)} \tag{21}$$

where ${}_pF_q\left(\left[\{a_j\}_{j=1}^{p}\right];\left[\{b_j\}_{j=1}^{q}\right];x\right)$ stands for the hypergeometric function of the type ${}_pF_q$.

From the noise data captured by k USRP sensors, we obtained the histogram for the spectral density of the product of the

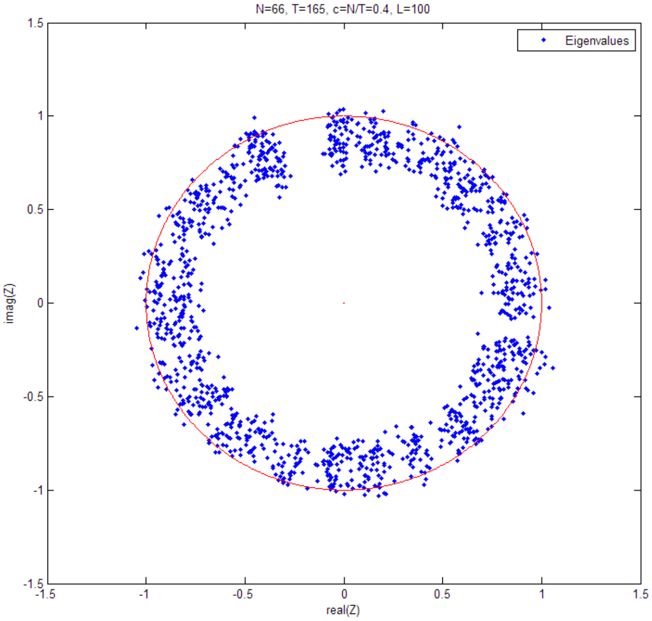

Fig. 18: The eigenvalue distribution for arithmetic mean of non-Hermitian random matrix, noise only, L=100.

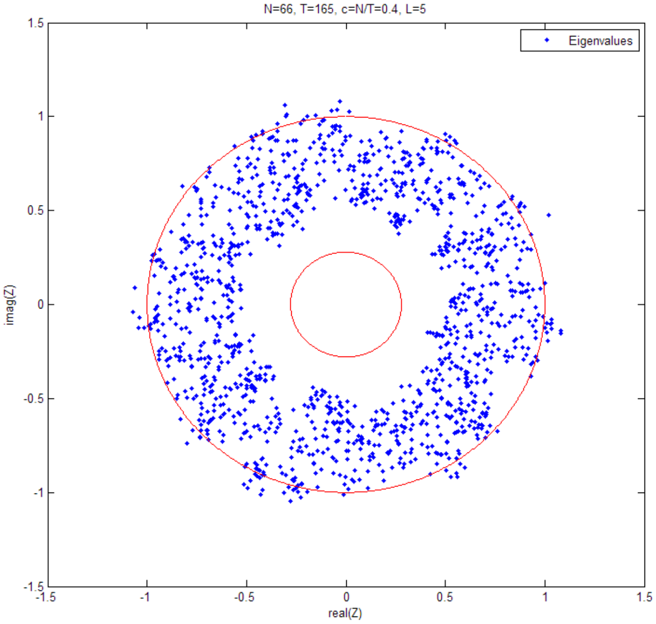

Fig. 19: The eigenvalue distribution for arithmetic mean of non-Hermitian random matrix, signal present, L=5.

Ginibre random matrices. Fig. 24 to Fig. 26 show that the histograms match the theoretical pdf well, for different k.

E. Summary of Theoretical Validation by Experimental Data

We applied variant data models on the massive data collected by our massive MIMO testbed. Firstly, we found that the theoretical eigenvalue distribution (if exists) can be validated by the experimental data for noise-only case. The random matrix based big data analytic model is successfully connected to the experiment. Secondly, the signal-present case can be differentiated from the noise-only case by applying the same data model. This result reveals the potential usage of the random-matrix based data model in signal detection, although the future work on the performance analysis is needed.

VI. Initial Applications of Massive MIMO Testbed as Big Data System

Besides signal detection, we demonstrated two applications based on the massive data analytic through the random-matrix method. The theoretical model in section V-A is used, i.e., we mainly apply the product of non-Hermitian random matrices on

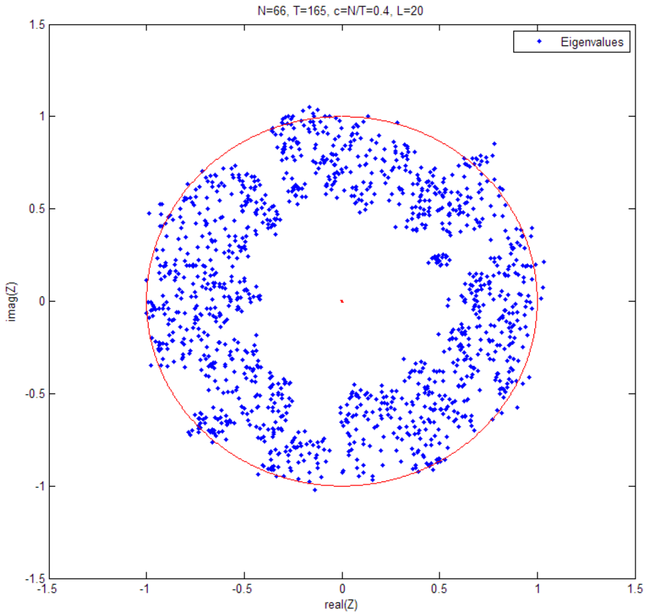

Fig. 20: The eigenvalue distribution for arithmetic mean of non-Hermitian random matrix, signal present, L=20.

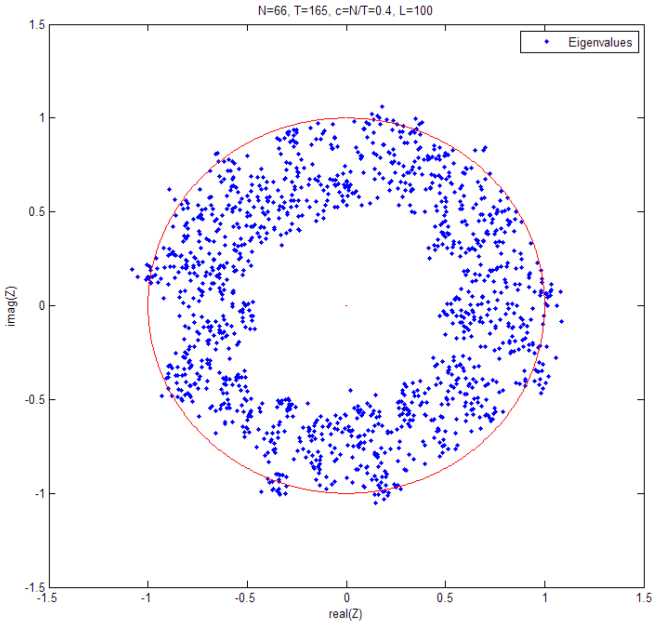

Fig. 21: The eigenvalue distribution for arithmetic mean of non-Hermitian random matrix, signal present, L=100.

the collected mobile data to investigate the corresponding eigenvalue distribution. Our aim is to make sense of massive data to find the hidden correlation between the random-matrix-based statistics and the information. Once correlations between causes and effects are empirically established, one can start devising theoretical models to understand the mechanisms underlying such correlations, and use these models for prediction purposes [23].

A. *Mobile User Mobility Data*

In a typical scenario where the mobile user is communicating with the massive MIMO base station while moving, the uplink waveform data received at each receiving antenna are collected. We applied the product of Hermitian random matrices to the data to observe the relationship between the eigenvalue distribution and the behavior of the moving mobile user. We are using the data from 10 antennas associated with 10 USRP receivers. Another USRP placed on a cart acts as the mobile user, which moves on the hallway of the 4th floor of the Clement Hall at Tennessee Technological University. The base station with up to 70 USRPs is on the same floor. The experiment results show that the moving speed of the mobile user is directly associated with the inner circle of the eigenvlaue distribution for the product of the Hermitian random matrices.

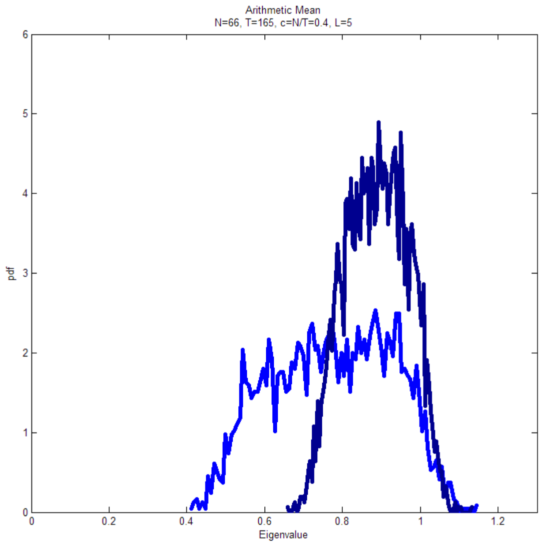

Fig. 22: Probability of eigenvalue for arithmetic mean of the non-Hermitian random matrix, both cases, with $L = 5$.

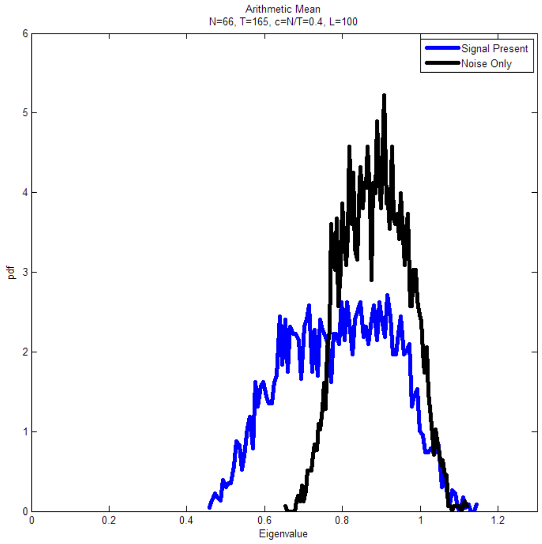

Fig. 23: Probability of eigenvalue for arithmetic mean of the non-Hermitian random matrix, both cases, with $L = 100$.

The experiments include five cases with different the moving speeds.

a) Case 1: **The Mobile User Stands in a Certain Place without Moving**

In this case, the mobile user has zero speed. What we observed in Figure 27 is that the inner radius of the circle is almost not changing. The average inner radius is a little less than 0.05 for the whole procedure.

b) Case 2: **The Mobile User Moves at a Nearly Constant Walking Speed**

In this case, the mobile user moves along a straight line at a nearly constant walking speed from a distant point to a point near the base station. Figure 28 shows the change of the inner radius of the circle law with time. The moving mobile user is actually on a cart pushed by a man. We see the inner radius is much bigger at the beginning when the cart is accelerating from almost motionless to a walking speed than the rest of the time. During the moving stage, the inner radius is much smaller and very stable at around 0.005.

c) Case 3: **The Mobile User Moves at a Very slow speed**

In this case, we move the mobile user at a very slow speed that is much smaller than walking speed. We see in Figure 29 that the inner radius is mostly vacillating between 0.02 and 0.05. This value is much smaller than that of the stationary case, but bigger than the walking-speed case.

d) Case 4: **The Mobile User Moves at Varying speed: Half the Time walking, Half the Time at a Very Slow Speed.**

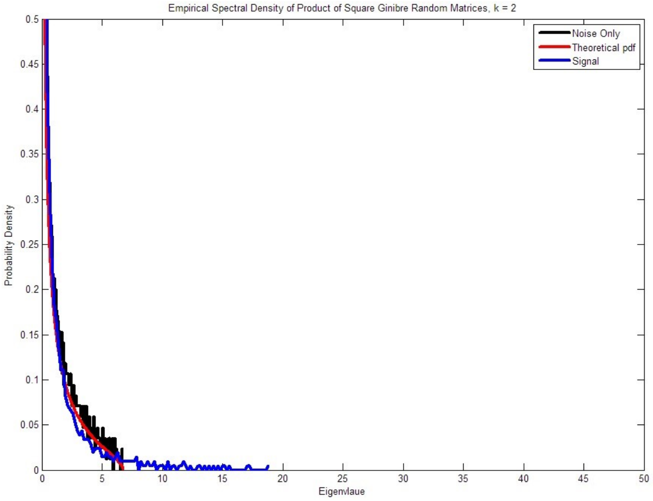

Fig. 24: Spectral density of eigenvalues for product of square Random Ginibre Matrices, k=2

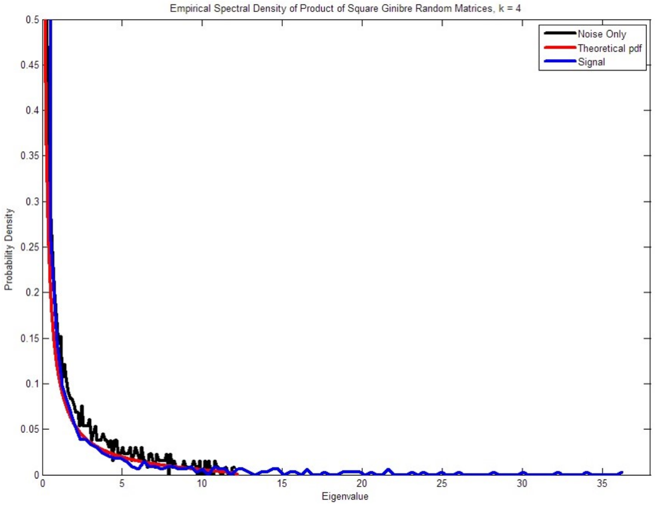

Fig. 25: Spectral density of eigenvalues for product of square Random Ginibre Matrices, k=4

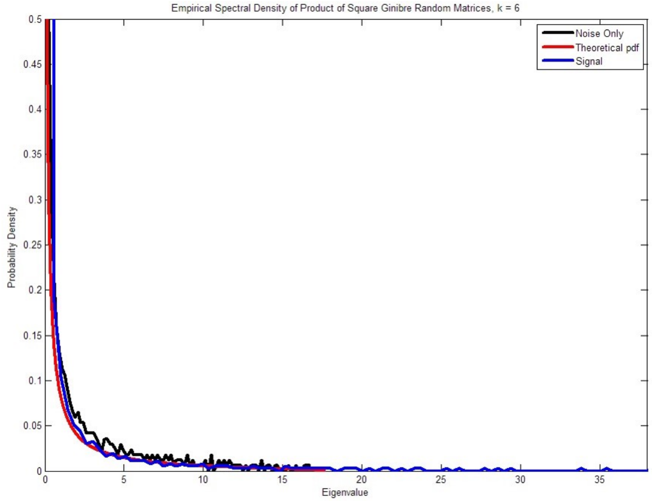

Fig. 26: Spectral density of eigenvalues for product of square Random Ginibre Matrices, k=6

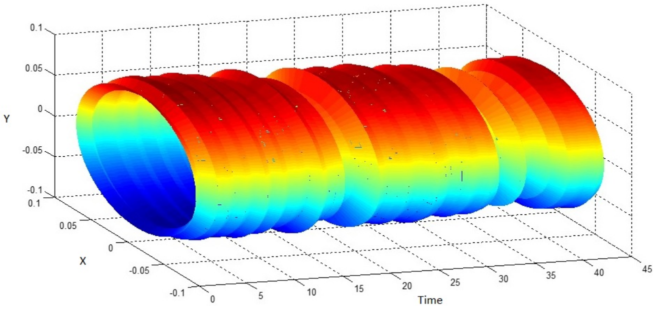

Fig. 27: Ring law inner radius changing with time for moving mobile user, case 1.

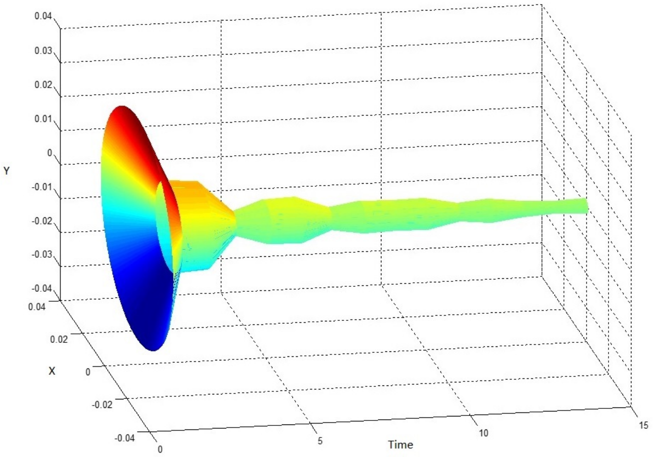

Fig. 28: Ring law inner radius changing with time for moving mobile user, case 2.

In this case, we try to observe the difference for the impacts from different moving speeds on the inner radius in one figure. Figure 30 shows that the radius in the first half is much smaller than that in the second half. Correspondingly, the moving speed in the first half is much higher than the latter half.

e) Case 5: **The Mobile User Moves at Varying Speed: Half the Distance Walking, Half the Distance at a Very Slow Speed.**

Similar to case 4, the impacts from different speeds are observed. A higher moving speed brings a smaller inner radius of the eigenvalue distribution. Because the walking speed part has equal distance with the slow speed part, the occupied time of the former is smaller than the later part, just as shown in Figure 31.

All the above cases reveal a common observation that the faster the mobile user moves, the smaller the inner radius of ring law is. From the big data point of view, we can get insight that a massive MIMO based station can use the inner radius of the ring law to estimate the moving status of the mobile user. As we know, basically more correlation in the signal brings a smaller inner radius of the ring law. Thus, this result is reasonable, as the faster speed of the mobile user causes more Doppler effect to the random signal received in the massive MIMO base station, i.e., more correlation detected by the product of the Hermitain random matrices.

B. Correlation Residing in Source Signal

Besides the correlation introduced by the moving environment, as in the above experiment, the correlation residing in the transmitting signal also has a significant impact on the eigenvalue distribution of the random matrix. Actually, in the section on theoretical model validation, we only compared the cases of noise-only and signal-present. The correlation within the signal creates the derivation of the eigenvalue distribution. In this section, we intentionally adjust the auto-correlation level of the generated signal that is transmitted by the mobile user. The corresponding effect on the inner radius of the ring law is also investigated by analyzing the collected data from antennas at the massive MIMO base station.

We generate the output signal following Eq. 22:

$$y\left(n\right) = \left(1 + r\right)x\left(n\right) + ry\left(n - 1\right) \tag{22}$$

which can also be represented by Figure 32. In the experiment, $x\left(n\right)$ is set as Gaussian white noise.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2015.2433920, IEEE Access

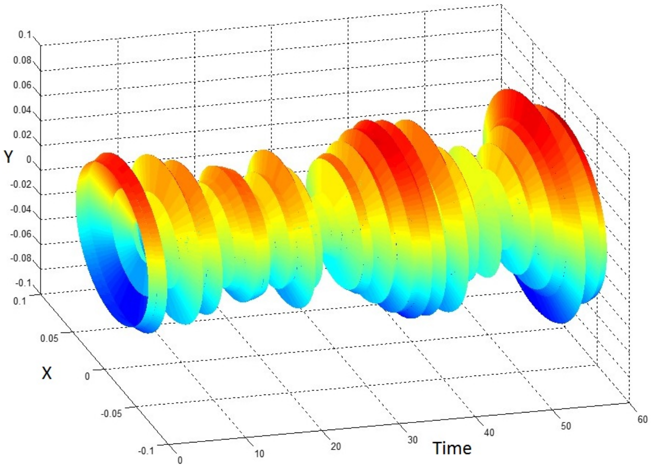

Fig. 29: Ring law inner radius changing with time for moving mobile user, case 3.

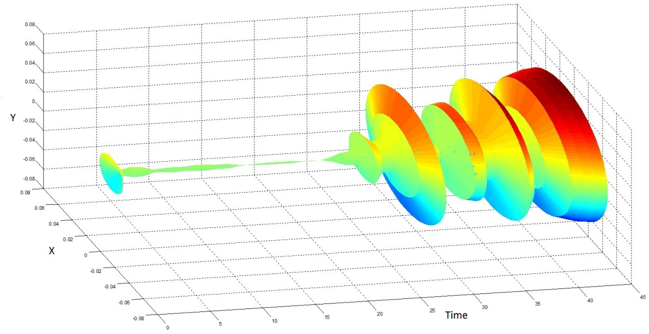

Fig. 30: Ring law inner radius changing with time for moving mobile user, case 4.

Essential to this signal generator is an auto-regression filter in which the parameter r is used to control the frequency response as shown in Figure 33 A bigger r leads a sharper frequency response that introduces more correlation within the transmitted signal. Thus, we can see that the inner radius of the ring law observed at the massive MIMO base station is as in Figure 34.

C. Insights from Applications

Both the applications bring us insights that the correlation residing in the signal can be matched to certain events in the network. In the network under our monitoring, such correlations can be detected and measured by our random-matrix-based data analysis method and finally be used to visualize the real event, such as the mobile user moving, fluctuation of the source signal correlation. This is a typical big data approach. The massive MIMO system is not only a communications system but also an expanded data science platform. We make sense of data by storing and processing the massive waveform data. Information will not be discarded, thus the energy of every bit/sample can be utilized as possible as we can. To our best knowledge, it is the first time, by concrete experiments, to reveal the value of the 5G massive MIMO as a big data system. We believe that more applications emerge in the future.

VII. Conclusion

The paper gives a first account for the 70-node testbed that takes TTU four years to develop. Rather than focusing on the details of the testbed hardware development, we use the testbed as a platform to collect massive datasets. The motivated application of this paper is massive MIMO. First, by using our initial experimental data, we find that large random matrices are natural models for the data arising from this tested. Second, the recently developed non-Hermitian free probability theory makes the theoretical predictions very accurately, compared with our experimental results. This observation may be central to our paper. Third, the visualization of the datasets are provided by the empirical eigenvalue distributions on the complex plane. Anomaly detection can be obtained through visualization. Fourth, when no visualization is required, we can formulate spectrum sensing or network monitoring in terms of matrix hypothesis testing. This formulation is relatively new in our problem at hand for massive MIMO. To our best knowledge, our work may be the first time. A new algorithm is proposed for distributed data across a number of servers.

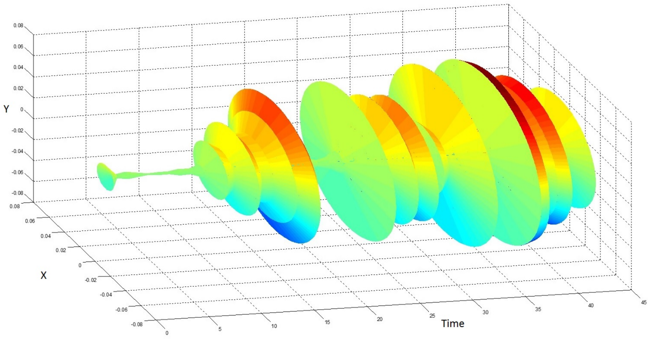

Fig. 31: Ring law inner radius changing with time for moving mobile user, case 5.

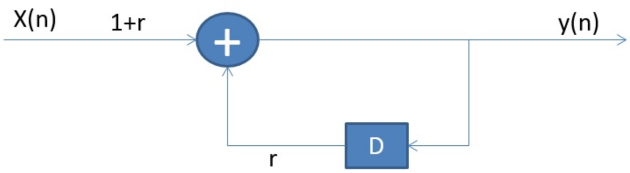

Fig. 32: Auto-regression filter used to generate the signal with adjustable autocorelation.

At this moment of writing [10], we feel that both theoretical understanding and experimental work allows for extension to other applications. First, thousands of vehicles need be connected. Due to mobility, streaming data that are spatially distributed across $N = 1,000$ becomes essential. We have dealt with hypothesis testing problem. How do we reduce the data size while retaining the statistical information in the data? Sketching [24] is essential [10]. Second, the testbed allows for the study of data analytical tools that will find applications in large-scale power grid, or Smart Grid [9]. For example, the empirical eigenvalue distribution of large random matrices is used for power grid in [25].

Acknowledgment

This work is funded by the National Science Foundation through three grants (ECCS-0901420, ECCS-0821658, and CNS-1247778), and the Office of Naval Research through two grants (N00010-10-1-0810 and N00014-11-1-0006).

References

[1] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, “Scaling up mimo: Opportunities and challenges with very large arrays,” *Signal Processing Magazine, IEEE*, vol. 30, no. 1, pp. 40–60, 2013.

[2] F. Boccardi, R. W. Heath Jr, A. Lozano, T. L. Marzetta, and P. Popovski, “Five disruptive technology directions for 5g,” *arXiv preprint arXiv:1312.0229*, 2013.

[3] N. R. Council, “Frontiers in massive data analysis.” The National Academies Press, 2013.

[4] O. Edfors and F. Tufvesson, “Massive mimo for next generation wireless systems,” *IEEE Communications Magazine*, p. 187, 2014.

[5] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Energy and spectral efficiency of very large multiuser mimo systems,” *Communications, IEEE Transactions on*, vol. 61, no. 4, pp. 1436–1449, 2013.

[6] C. Zhang and R. C. Qiu, “Data modeling with large random matrices in a cognitive radio network testbed: initial experimental demonstrations with 70 nodes,” *arXiv preprint arXiv:1404.3788*, 2014.

[7] R. C. Qiu, Z. Hu, H. Li, and M. C. Wicks, *Cognitive radio communication and networking: Principles and practice*. John Wiley & Sons, 2012.

[8] R. Qiu and M. Wicks, *Cognitive Networked Sensing and Big Data*. Springer Verlag, 2014.

[9] R. Qiu and P. Antonik, *Big Data and Smart Grid*. John Wiley and Sons, May 2015.

[10] R. Qiu, *Principles of Massive Data Analysis: The Random Matrix Approach*. Manuscript Draft.

[11] T. Anderson, *An Introduction to Multivariate Statistical Analysis*. Hoboken, NJ: Wiley-Interscience [John Wiley & Sons],, 2003. 3rd Edition, Wiley Series in Probability and Statistics,.

[12] J. A. Tropp, “An introduction to matrix concentration inequalities,” *arXiv preprint arXiv:1501.01571*, 2015.

[13] J. W. Mauchly, “Significance test for sphericity of a normal n-variate distribution,” *The Annals of Mathematical Statistics*, vol. 11, no. 2, pp. 204–209, 1940.

[14] Z. Bai, D. Jiang, J.-F. Yao, and S. Zheng, “Corrections to lrt on large-dimensional covariance matrix by rmt,” *The Annals of Statistics*, pp. 3822–3840, 2009.

[15] Z. Bai and J. Silverstein, “Clt for linear spectral statistics of large-dimensional sample covariance matrices,” *The Annals of Probability*, vol. 32, no. 1A, pp. 553–605, 2004.

[16] Q. Wang, J. W. Silverstein, and J.-f. Yao, “A note on the clt of the lss for sample covariance matrix from a spiked population model,” *Journal of Multivariate Analysis*, vol. 130, pp. 194–207, 2014.

[17] A. Onatski, M. J. Moreira, and M. Hallin, “Asymptotic power of sphericity tests for high-dimensional data,” *The Annals of Statistics*, vol. 41, no. 3, pp. 1204–1231, 2013.

[18] A. Onatski, M. J. Moreira, M. Hallin, *et al.*, “Signal detection in high dimension: The multispiked case,” *The Annals of Statistics*, vol. 42, no. 1, pp. 225–254, 2014.

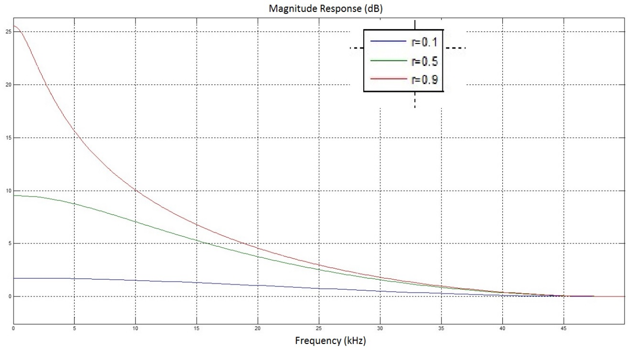

Fig. 33: Bigger r leads to sharper frequency response of the AR filter for signal generator.

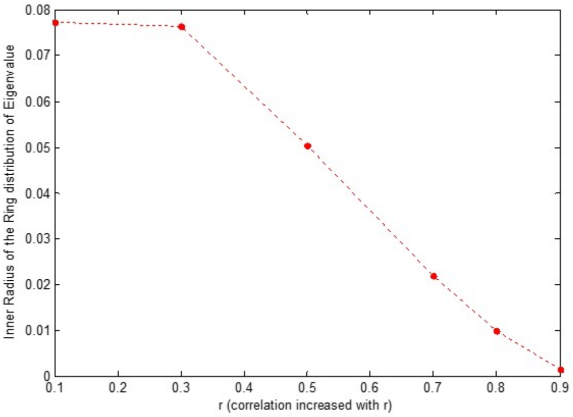

Fig. 34: Inner radius of ring law changes with the r, bigger r leads smaller inner radius.

[19] A. Van der Vaart, *Asymptotic Statistics*. Cambridge Univ Press, 1998.

[20] D. Passemier, M. R. McKay, and Y. Chen, “Hypergeometric functions of matrix arguments and linear statistics of multi-spiked hermitian matrix models,” *arXiv preprint arXiv:1406.0791*, 2014.

[21] D. Passemier, M. R. Mckay, and Y. Chen, “Asymptotic linear spectral statistics for spiked hermitian random matrix models,” *arXiv preprint arXiv:1402.6419*, 2014.

[22] C. Zhang and R. C. Qiu, “Massive mimo testbed-implementation and initial results in system model validation,” *arXiv preprint arXiv:1501.00035*, 2014.

[23] R. C. Qiu, *Smart Grid and Big Data: Theory and Practice*. John Wiley, 2014.

[24] D. P. Woodruff, “Sketching as a tool for numerical linear algebra,” *arXiv preprint arXiv:1411.4357*, 2014.

[25] X. He, Q. Ai, C. Qiu, W. Huang, and L. Piao, “A big data architecture design for smart grids based on random matrix theory,” *arXiv preprint arXiv:1501.07329*, 2015. submitted to IEEEE Trans. Smat Grid, in revision.