Modeling Massive Amount of Experimental Data with Large Random Matrices in a Real-Time UWB-MIMO System

Xia Li, Student Member, IEEE, Feng Lin, Student Member, IEEE, Robert C. Qiu, Senior Member, IEEE

Abstract—The aim of this paper is to study data modeling for massive datasets. Large random matrices are used to model the massive amount of data collected from our experimental testbed. This testbed was developed for a real-time ultra-wideband, multiple input multiple output (UWB-MIMO) system. Empirical spectral density is the relevant information we seek for. After we treat this UWB-MIMO system as a black box, we aim to model the output of the black box as a large statistical system, whose outputs can be described by (large) random matrices. This model is extremely general to allow for the study of non-linear and non-Gaussian phenomenon. The good agreements between the empirical results and the theoretical predictions validate the correctness of the our suggested data model.

Index Terms—Data modeling, time-lagged correlation matrix, kernel density estimation, Marcenko-Pastur law, data collection testbed, large random matrices.

I. INTRODUCTION

The purpose of statistical methods is to reduce a large quantity of data to a few that are capable of containing as much as possible of the relevant information in the original data. Because the data will generally supply a large number of “facts”, many more than are sought, much information in the data is irrelevant. In this paper, our relevant information is the empirical spectral density of the (random) data matrix.

In [1], the idea of modeling Big Data with (large) random matrices was proposed for the first time, where the theory for large random matrices may be, arguably, the foundation for communications, networking and sensing. In recent works [2], [3], this idea has been carried out from distributed sensing to the Smart Grid. It is argued in [3] that the theme of the 5G wireless network may be unified in the framework of Big Data.

These series works [1]–[3] clearly articulated the ideas and treated the necessary analytical tools. The concrete results, on the other hand, are still few,—maybe none, to our best knowledge—especially using experimental data. This short paper is aimed at filling this gap. In particular, we deal with the pressing issue daily encountered in our Lab: to make sense of the massive amount data that are accumulated at an accelerating speed. The ultra-wideband, multiple input multiple output (UWB-MIMO) system [4] is used as an experimental testbed. After we treat this UWB-MIMO system as a black box,—the assumption of linear time-varying system can be discarded—our goal is to model the output of the black box as a large statistical system, whose outputs can be described by (large) random matrices. This model is extremely general to allow for the study of non-linear and non-Gaussian phenomenon.

The structure of this paper is as follows. First, the theoretical models are introduced. Then, the empirical spectral results are compared with the theoretical predictions. The good agreements validate the correctness of the our suggested data model.

II. EMPIRICAL SPECTRAL DENSITY OF RANDOM MATRIX

Suppose that \( X_{ij} \) are independent and identically distributed (i.i.d.) real random variables. Let \( \mathbf{X}_n = (X_{ij})_{p \times n} \) and \( \mathbf{T}_n \) be a \( p \times p \) nonrandom Hermitian nonnegative definite matrix. Consider the random matrices [5]

\[
\mathbf{A}_n = \frac{1}{n} \mathbf{T}_n^{1/2} \mathbf{X}_n \mathbf{X}_n^T \mathbf{T}_n^{-1/2}.
\]

When \( E(X_{ij}) = 0 \) and \( E(X_{ij}^2) = 1 \), \( \mathbf{A}_n \) can be viewed as a sample covariance matrix drawn from the population with covariance matrix \( \mathbf{T}_n \). Moreover, if \( \mathbf{T}_n \) is another sample covariance matrix, independent of \( \mathbf{X}_n \), then \( \mathbf{A}_n \) is a Wishart matrix.

The basic limit theorem regarding \( \mathbf{A}_n \) concerns its empirical spectral distribution \( F_{\mathbf{A}_n} \), which is given by

\[
F_{\mathbf{A}_n}(x) = \frac{1}{p} \sum_{k=1}^{p} I(\lambda_k \leq x),
\]

where \( \lambda_k, k = 1, \ldots, p \) denote the eigenvalues of \( \mathbf{A}_n \).

Suppose the ratio of the dimension to the sample size \( c_n = p/n \) tends to \( c \) as \( n \to \infty \). When \( \mathbf{T}_n \) becomes the identity matrix, \( F_{\mathbf{A}_n} \) tends to the well-known Marcenko and Pastur law with the density function

\[
f_c(x) = \begin{cases} (2\pi cx)^{-1} \sqrt{(b-x)(x-a)}, & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}
\]

It has point mass \( 1-c^{-1} \) at the origin if \( c > 1 \), where \( a = (1 - \sqrt{c})^2 \) and \( b = (1 + \sqrt{c})^2 \).

Suppose \( f(x) \) is an unknown density function of \( X_{ij} \), and \( F_n(x) \) is the empirical distribution function determined by the sample. A popular nonparametric estimate of \( f(x) \) is then

\[
\hat{f}_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K \left( \frac{x - X_{ij}}{h} \right)
\]

\[
= \frac{1}{h} \int K \left( \frac{x - y}{h} \right) dF_n(y)
\]
where the function $K(\cdot)$ is a Borel function and $h = h(n)$ is the bandwidth which tends to 0 as $n \to \infty$.

Under some regularity conditions on the kernel function $K(\cdot)$, it is well known that $f_n(x) \to f(x)$ in some sense. The kernel density estimation would be

$$
  f_n(x) = \frac{1}{ph} \sum_{i=1}^{p} K \left( \frac{x - \mu_i}{h} \right)
  = \frac{1}{h} \int K \left( \frac{x^2}{h^2} \right) dF \mathcal{A}_n(y)
$$

where $\mu_i, i = 1, \cdots, p$, are eigenvalues of $\mathcal{A}_n$.

### III. Eigenvalue Spectra of Time-Lagged Correlation Matrices

The time-lagged analog to the sample covariance matrix is defined as $C^\tau = \sum r_i^T r_t^\tau$, where one time series is shifted by $\tau$ timesteps with respect to the other. Sample covariance matrix can be viewed as a special case of time-lagged covariance matrix with $\tau = 0$, also known as equal-time correlations. For $\tau \neq 0$, the lagged correlation matrix $C_\tau$ is non-symmetric\footnote{Thus its eigenvalues are complex numbers, in contrast with the real values of symmetric sample covariance matrix.} and contains the lagged auto-correlations in the diagonal. It can be written as

$$
  C_\tau = \frac{1}{T} X D_\tau X^\tau
$$

where $D_\tau \equiv \delta_{i,t+i} \tau$ and where $X$ is the $N \times T$ normalized time-series data.

The projection of eigenvalue density $\rho(\tau)$ onto the $x$-axis is denoted by $\rho_x(\tau)$, and the projection onto the $y$-axis is denoted by $\rho_y(\tau)$. They are nothing but the rescaled spectra of the solution to the symmetric, $\rho^S(\lambda)$, and to the anti-symmetric problem, $\rho^A(\lambda)$. To be more explicit, we have [6]

$$
\begin{align*}
  \rho_x(\lambda) &\equiv \rho \left( \text{Re} \left( \lambda \right) \right) = \int S \rho(r) dy = \rho^S(\sqrt{2}x) \\
  \rho_y(\lambda) &\equiv \rho \left( \text{Im} \left( \lambda \right) \right) = \int S \rho(r) dx = \rho^A(\sqrt{2}y),
\end{align*}
$$

where the integration extends over the support $S$ in the complex plane.

The eigenvalue density of the symmetric problem can be obtained from the well-known relation

$$
  \rho^S(x) = \sum_{n} \delta(x - x_n) = \frac{1}{\pi} \lim_{\epsilon \to 0^+} \left[ \text{Im} \left( G^S(x - i\epsilon) \right) \right].
$$

For a radial symmetric problem, $\rho^S \sim \rho^A$. The $G(z)$ in [5] is the Green's function given by

$$
\begin{align*}
\frac{1}{Q^2} z^2 G^4(z) &- 2 \frac{1}{Q^2} \left( \frac{1}{Q} - 1 \right) z G^3(z) \\
- \frac{1}{Q} \left( z^2 - \left( \frac{1}{Q} - 1 \right)^2 \right) G^2(z) \\
+ 2 \left( \frac{1}{Q} - 1 \right) z G(z) &+ 2 - \frac{1}{Q} = 0,
\end{align*}
$$

with $Q \equiv T/N$ playing the role of a information-to-noise ratio. This equation is independent of a specific value for $\tau$ and is valid for any value of it.

\footnote{Thus its eigenvalues are complex numbers, in contrast with the real values of symmetric sample covariance matrix.}
**Case 3**: Wide band NC-OFDM signal is used as transmission signal with BPSK baseband modulation. NC-OFDM offers more flexibility compared with OFDM signal due to the capability of configuring its occupancy of sub-carriers (also called tones). The total number of sub-carriers of NC-OFDM signal in our experiment is 1024, in which only 426 sub-carriers are occupied. The positions of the valid sub-carriers are chosen arbitrarily. In this case, they are allocated in the positions with the indexes of \([10 : 50, 80 : 100, 140 : 200, 300 : 400, 600 : 700, 800 : 900]\). The received time domain signal is first converted to the frequency domain by fast Fourier transform (FFT). Then the frequency domain signal is used to form a (large) random matrix \(X\), to obtain its eigenvalues. The frequency spectrum of the received NC-OFDM signal is shown in Fig. 6. The perfect agreement between the empirical spectral density and its theoretical prediction (Marcenko-Pastur law) is illustrated in Fig. 7. This is, remarkably, similar to Fig. 3—the white Gaussian noise only. The reason behind the remarkable result is that the carriers of the frequency domain NC-OFDM signal can be modeled as i.i.d. random variables. It is natural to expect that a result similar to Fig. 7 will be obtained, if all the 1024 tones are used, instead of 426 tones. This is indeed the case, validated using our experimental data.
V. EXPERIMENTAL RESULTS OF TIME-LAGGED CORRELATION MATRICES

Any deviation from the benchmark Marcenko-Pastur law—valid for a random sample covariance matrix obtained using i.i.d. random variables—indicates the presence of some useful correlation—called “signal”. In this section, we study the benchmark for the time-lagged correlation matrix, obtained by solving (9).

The comparisons with $Q = 0.5$ and $Q = 10$ are shown in Fig. 8, Fig. 9 and Fig. 10, Fig. 11 respectively.

VI. CONCLUSION

Treating a ultra-wideband, multiple input multiple output (UWB-MIMO) system testbed as a black box, we have modeled the output of the system testbed as a large statistical system, whose outputs can be described by (large) random matrices. This model is extremely general to allow for the study of non-linear and non-Gaussian phenomenon. The good agreements between the theoretical models and the empirical results validate the correctness of the our suggested data model. This is the initial experimental demonstration of the framework of modeling massive datasets using large random matrices, which, to our best knowledge, is the first time in the field of wireless communications and radar [1]–[3].

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