Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2024, 12, 3902-3911
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Denoising Data with Random Matrix Theory

Nathan Jiang

Department of Mathematics, Columbia University, New York, NY, USA
Email: nj325@njit.edu

How to cite this paper: Jiang, N. (2024) Denoising Data with Random Matrix Theory. *Journal of Applied Mathematics and Physics*, **12**, 3902-3911.
https://doi.org/10.4236/jamp.2024.1211237

Received: October 9, 2024
Accepted: November 24, 2024
Published: November 27, 2024

Abstract

Properties from random matrix theory allow us to uncover naturally embedded signals from different data sets. While there are many parameters that can be changed, including the probability distribution of the entries, the introduction of noise, and the size of the matrix, the resulting eigenvalue and eigenvector distributions remain relatively unchanged. However, when there are certain anomalous eigenvalues and their corresponding eigenvectors that do not follow the predicted distributions, it could indicate that there's an underlying non-random signal inside the data. As data and matrices become more important in the sciences and computing, so too will the importance of processing them with the principles of random matrix theory.

Keywords

Random Matrix Theory, Universality, Wishart Matrices, Marchenko-Pastur (M-P) Distribution, Noise, Sparsity, Signaling, Linear Sketching

1. Introduction

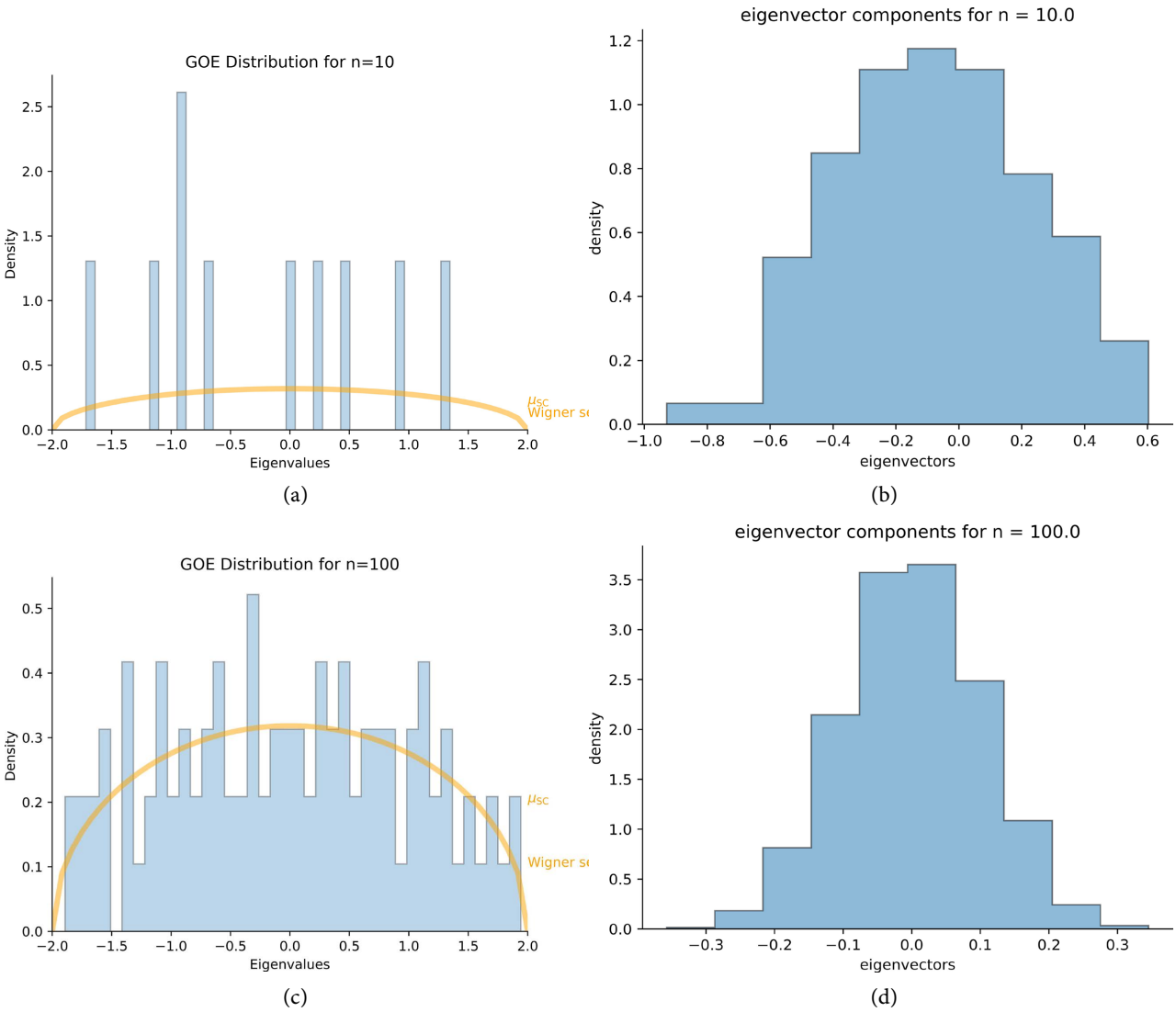
While many random systems exist in the universe, predictions and conclusions can still be drawn from them. Whether it be energy states of Hamiltonian nuclei or biological signals among unicellular processes, the principles of random matrix theory can be used to denoise systems and parse out the most important information by studying the eigenvalues and eigenvectors of a random matrix.

It turns out that random matrix theory imposes more structure, not less. The patterns of universality hold independently of numerous parameters in the sampling of a random matrix, which allows the process described in the article to separate the randomly generated noise from important signals in natural data sets. However, data collection methods are often imperfect; sometimes readings are false, entries are missing, and data sets are too large to compute with a reasonably-powered computer, but the beauty of Random Matrix Theory is that it can use this

well-defined structure, along with some algorithms, to recover the important signals even despite some flaws in the data.

2. Random Matrix Theory Fundamentals

A random matrix is defined as a matrix whose entries are randomly sampled. An ensemble of random matrices is a group of matrices whose entries are sampled in the same manner. Examples of ensembles include the Gaussian Orthogonal Ensemble (GOE), which is sampled as a symmetric, $n \times n$ matrix A where entries above the diagonal are sampled from N(0, 1) and entries on the diagonal are sampled from N(0, 2) with all entries being divided by a factor $\sqrt{2n}$ [1]. The probability density of the eigenvalues of a GOE matrix is defined by the Wigner Semicircle distribution $f(x) = \frac{1}{2\pi}\sqrt{4-x^2}$, and the components of the eigenvectors follow a normal distribution. As n increases, the density of the resulting eigenvalues will converge to the Wigner Semicircle Distribution (**Figure 1**).



(a) (b) (c) (d)

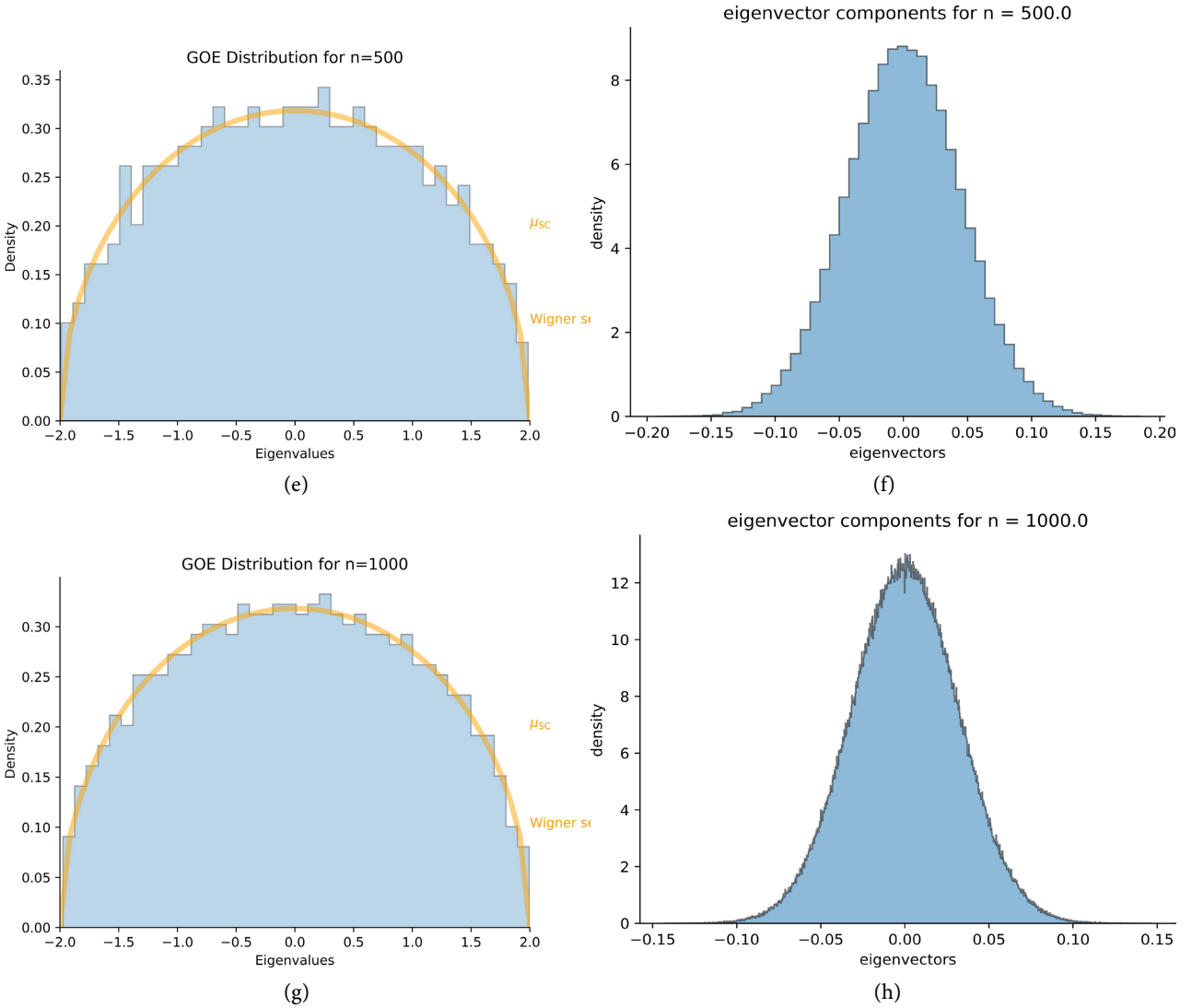


Figure 1. Eigenvalue and eigenvector distribution for GOE matrix of n = 10, 100, 500, 1000 [2].

The Wishart ensemble is defined by an $m \times n$ matrix A containing entries sampled from N(0, 1). The gram matrix $W = \frac{AA^{\mathrm{T}}}{n}$ is then constructed to form the Wishart matrix [3]. The resulting eigenvalue probability density, known as the Marchenko-Pastur (M-P) Distribution, which depends on the parameter r, is defined by:

$$\mu(x) = \begin{cases} \frac{1}{2\pi r x \sigma^2}\sqrt{(\lambda_+ - x)(x - \lambda_-)}, 0 < r \le 1 \\ \left(1 - \frac{1}{r}\right) + \delta(x) + \frac{1}{2\pi r x \sigma^2}\sqrt{(\lambda_+ - x)(x - \lambda_-)}, r > 1 \end{cases},$$

where $r = \frac{m}{n}$, $\lambda_+ = \sigma^2\left(1 + \sqrt{r}\right)^2$, and $\lambda_- = \sigma^2\left(1 - \sqrt{r}\right)^2$ (**Figure 2**) [1].

λ_+ is also known as the Tracy-Widom Critical Eigenvalue since a completely randomly generated Wishart matrix will not have an eigenvalue greater than λ_+.

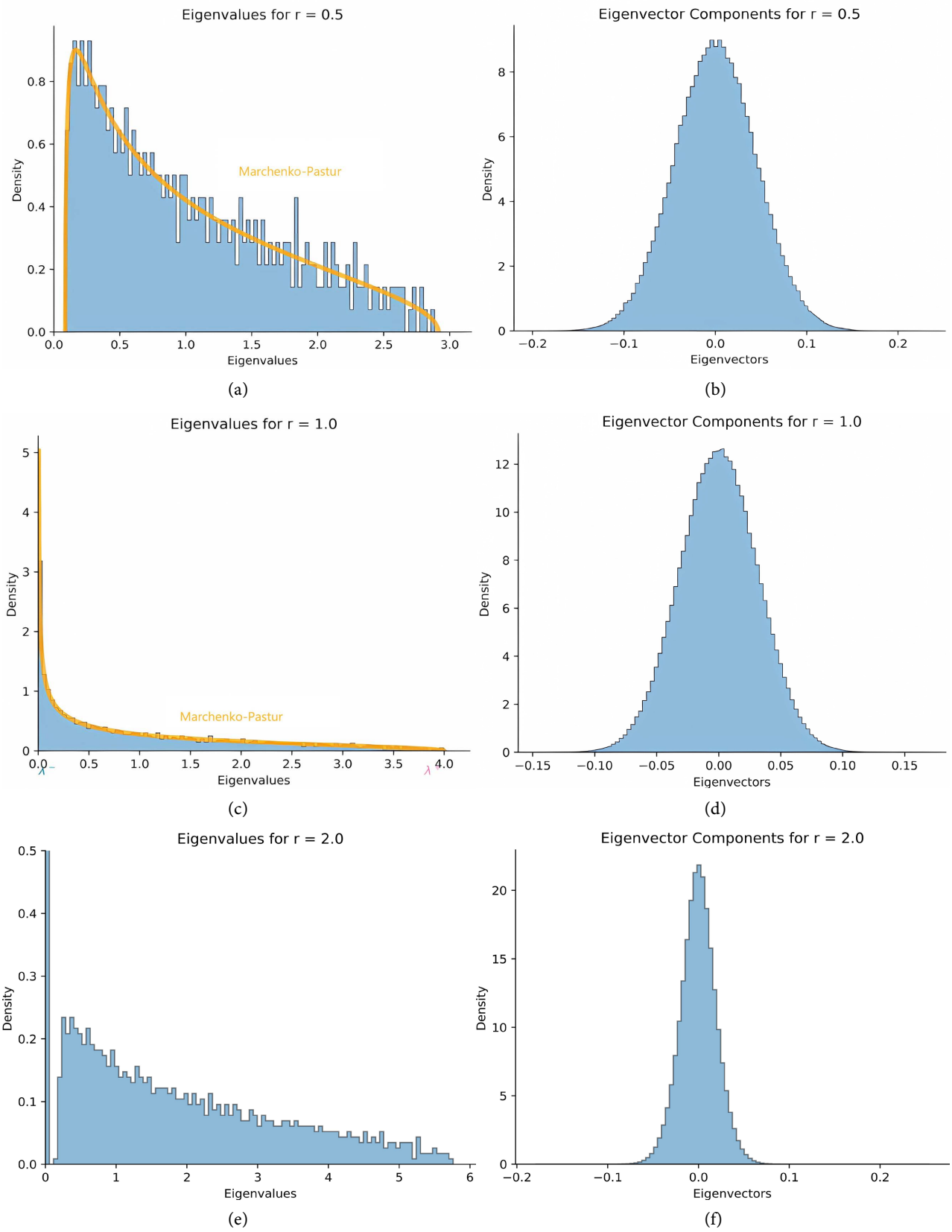
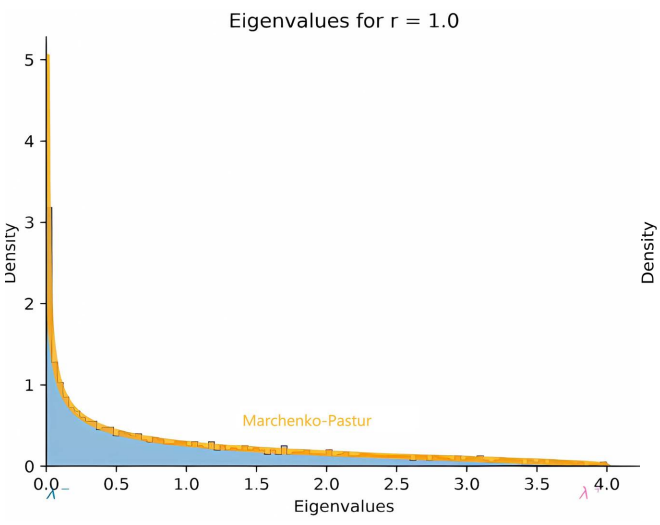


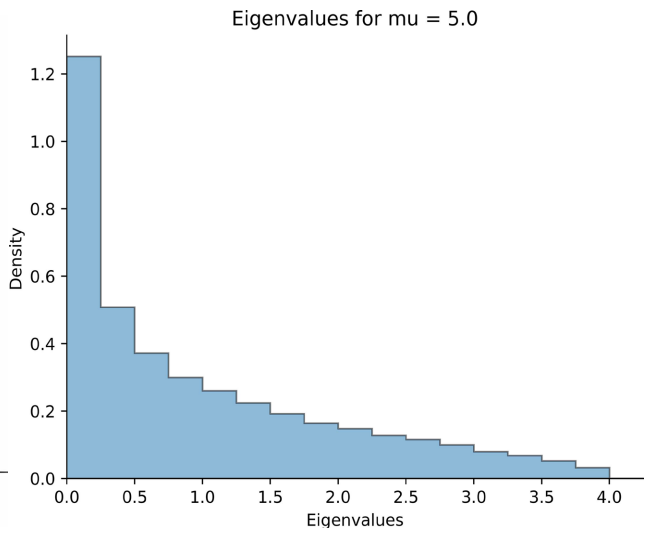
Figure 2. M-P distribution eigenvalues and eigenvectors for matrices with entries sampled from N(0, 1) for different r values and n = 1000 [2].

The eigenvector components follow a normal distribution. It's also worth noting that the process of creating the Wishart ensemble closely resembles the Singular Value Decomposition of A, where $A = U\Sigma V^{\mathrm{T}}$, and the singular values in Σ are proportional to the eigenvalues predicted by the Marchenko-Pastur distribution.

It turns out, however, that many things do not affect the overall structure of random matrices, which is known as universality. As long as the mean μ and variance σ^2 are constant, the distribution of eigenvalues and eigenvectors for a given ensemble does not change based on the overall probability distribution from which the entries are sampled. For example, the Readamacher $P\left(x_{i,j} = -1\right) = P\left(x_{i,j} = 1\right) = \frac{1}{2}$ distribution, a distribution where the probability of sampling a -1 and 1 are equal at $\frac{1}{2}$, can substitute a normal distribution and the eigenvalue distribution of the resulting matrix, whose entries strictly consist of ± 1, remains the same (**Figures 3(a)-(d)**). Likewise, adjusting the mean μ does not change the core of the distribution of the entries, but it introduces distracting outlier eigenvalues outside the bounds of the Marchenko-Pastur distribution, and therefore, it is best for data sets to be z-score normalized. What about σ then? It turns out that adjusting the standard deviation σ scales the length of the distribution by a factor σ^2 and the height by a factor $\frac{1}{\sigma^2}$, maintaining the total area under the probability density curve (**Figures 3(a)-(d)**). The only thing that does affect the patterns is that the entries must be sampled independently, and matrices with heavily dependent columns behave very differently from the predicted distributions (**Figure 3(e)**). Thus, the Random Matrix Theory shows that the eigenvalue distribution for a given ensemble does not change with respect to the distribution of the entries of the matrix, nor the mean nor standard deviation.



(a)



(b)

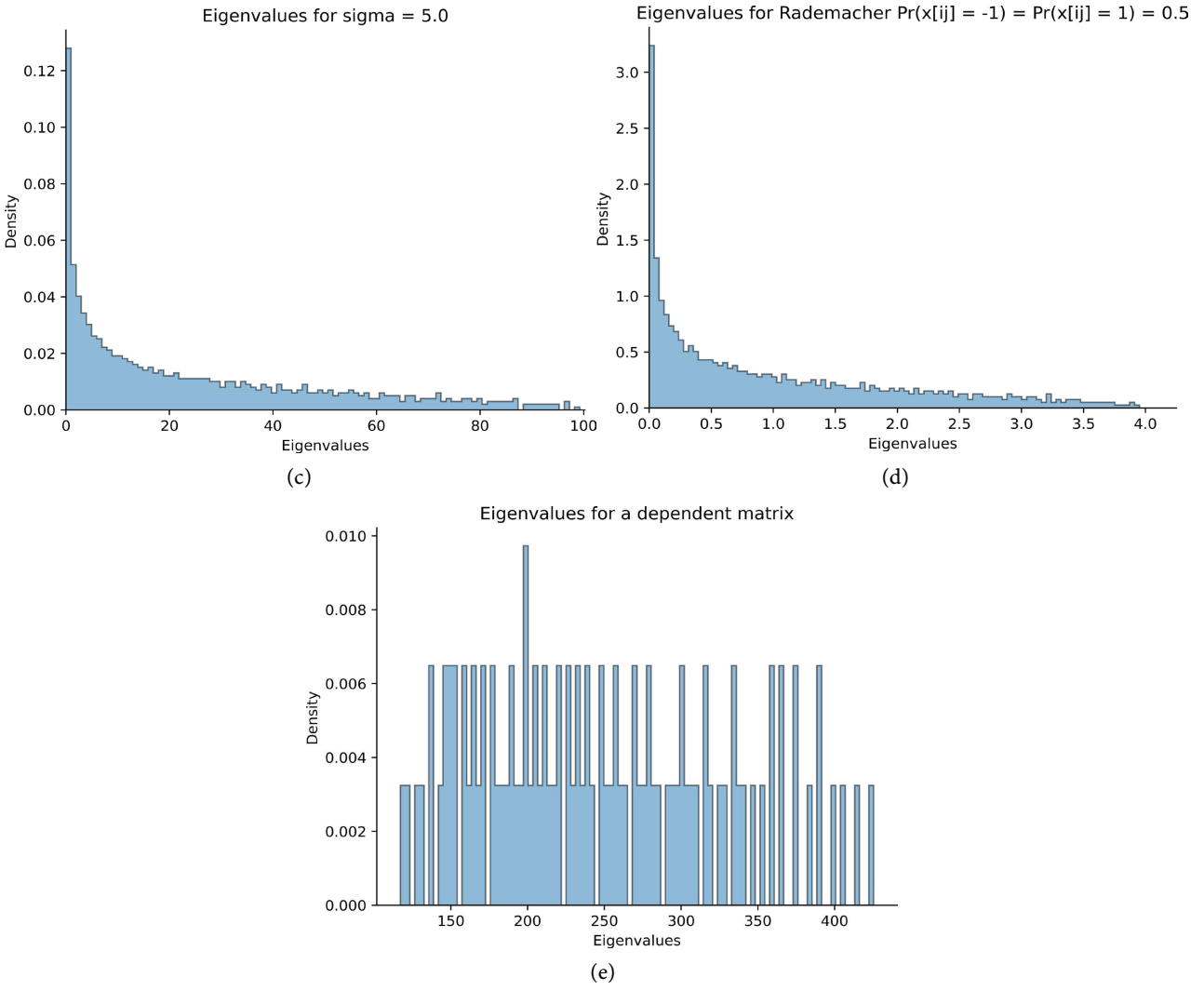


Figure 3. (a)-(d) adjusting the distribution, mean, and standard deviation with respect to a standard normal distribution and r = 1 [2]. (e) eigenvalues for a 1000 × 1000 dependent matrix composed of 10 identical but randomly generated 1000 × 100 blocks [2].

3. Natural Signaling

A common application of random matrix theory is to retrieve natural signals. We embed a simulated natural signal by adding the gram matrix (SS^{T}) of a low-rank randomly generated signal S of dimensions $m \times k$, where $k \ll m$ to random matrix A. Often, biological signals come embedded in this form. Doing so will generate $n-k$ eigenvalues within the M-P distribution and k eigenvalues much larger than the Tracy-Widom Critical Eigenvalue, and their corresponding eigenvector components are not normally distributed like the other eigenvectors [2]. Interestingly, the rank of the signal can be recovered by the eigenvalues alone. The eigenvectors of the k largest eigenvalues should form the span of the natural signal. For the $k=1$ case, the error margin can be easily visualized and measured as the angle between the vectors. The typical error between the eigenvector and signal is around 0.02 - 0.06 for $n=1000$ (**Figure 4**).

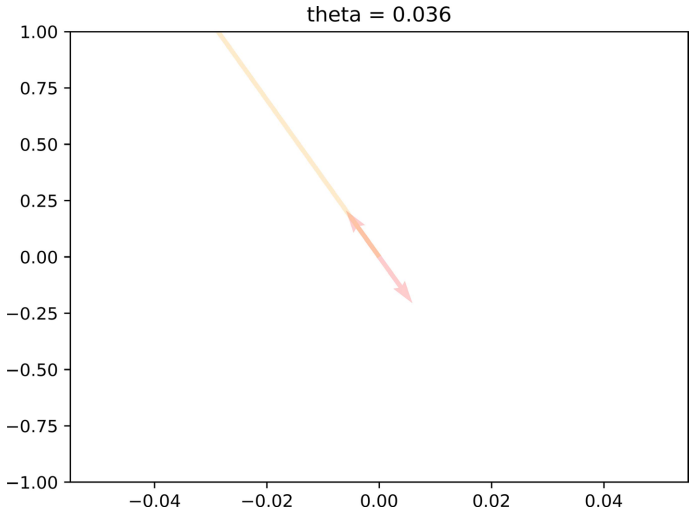


Figure 4. A 2-D projection of the eigenvector (yellow) and signal (red) plotted along with a guess of the signal rank and error margin [2].

4. Noise

A common noise model is known as sparsity, where entries of a matrix are randomly replaced by 0 since experimental data often has large gaps due to imperfect measuring methods. The procedure involved pre-multiplying a diagonal matrix set to 95% 0 s and 5% 1 s to the random $m \times n$ matrix A with the signal already added, which would zero-out 95% of the columns. The following denoising algorithm was then applied to the entries:

$$A_{ij} = \frac{1 \times 10^6}{mn} A_{ij} \tag{1}$$

$$A_{ij} = \log_2 \left(1 + A_{ij}\right) \tag{2}$$

$$A_{ij} = \frac{A_{ij} - \mu}{\sigma} \tag{3}$$

where μ, σ^2 is the mean and variance of all the entries of the previous line [4].

The gram matrix $W = \frac{AA^{\mathrm{T}}}{n}$ is calculated and the eigenvalues and eigenvectors are then computed. The resulting k signal eigenvectors almost form the span of the signal; however, they are instead parallel to DS where D is the diagonal matrix with a 1 for every nonzero column and 0 for every sparse column, and S is the full signal (**Figure 5**).

In the real world, data collected are often sparse, so utilizing the theorems of random matrix theory allows signals to be approximated even when as much as 95% of the entries are 0 due to sparsity. While having more data available would yield a higher accuracy, having the vast majority of entries 0 is still enough to gather information about a potential signal [2].

5. Linear Sketching

Some matrices are too large for a reasonable computer to calculate all the

```
guess rank = 1.0 guess rank = 1.0
True             True
theta (degrees) =theta (degrees) =
[[179.96666656]] [[179.96185846]]
```

(a) (b)

Figure 5. Error-values in the signal for a 95% sparce matrix (left) compared to its non-sparsified counterpart (right).

eigenvalues and eigenvectors. A linear sketch is a projection of a square matrix $\mathbb{R}^n \to \mathbb{R}^p$ with $p \ll n$ [5]. The projection matrix P is set to dimensions $p \times n$ and be 95% 0 s, 5% 1 s, meaning that the columns of a resulting sketch are a linear combination of other vectors [5]. The sketch is then denoised using the same algorithm and its eigenvalues and eigenvectors are calculated. While something like a 10,000 × 10,000 matrix is too large to calculate in a timely manner, it is possible to make multiple 100 × 100 sketches of the large matrix and obtain an accurate approximation of the signal (**Figure 6**). Even though higher-dimensional sketches are more accurate, taking many low-rank ones can dramatically speed up computational efficiency without sacrificing accuracy since eigenvalues are orders of magnitude easier to calculate with a smaller matrix size while still yielding an accurate result.

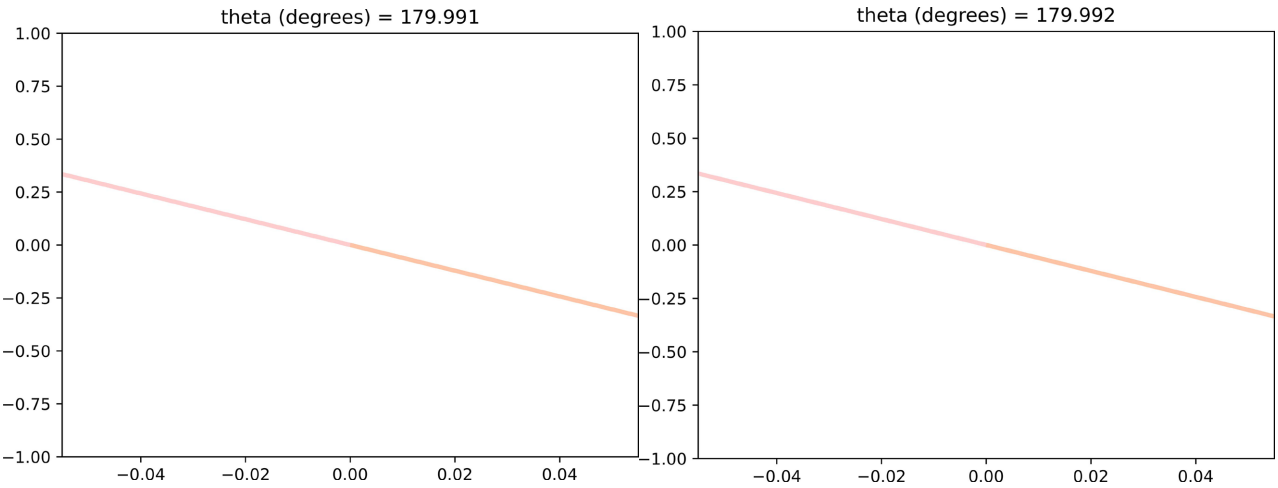


Figure 6. Signal from two 100 × 100 sketches compared to a simulated 10,000 × 1 signal added to a 10,000 × 10,000 matrix [2].

As it can be seen, the linear sketch is able to well-approximate the signal direction of the larger matrix while taking a lot less time to compute all the eigenvalues and eigenvectors. The error θ, measured by

$$\theta = \frac{180}{\pi}\cos^{-1}\left(\frac{v \cdot s}{\|v\|\|s\|}\right),$$

where $v = P^{\mathrm{T}}\left(PP^{\mathrm{T}}\right)^{-1}x$, x is the signal eigenvector, and s is the simulated signal.

Taking multiple sketches and then applying the algorithms to a new matrix T_m, with its i,jth entries being the mean of the i,jth entries across the sketches then decreases the error dramatically (Figure 7).

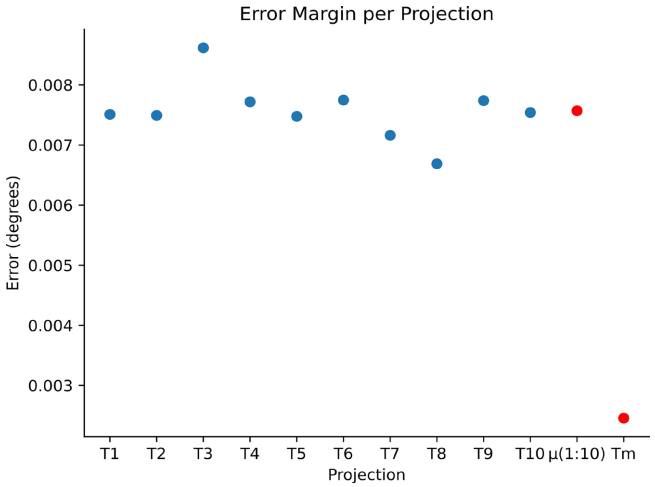


Figure 7. Graph of 10 sketches T1-T10, along with their mean, and the error margin for a matrix Tm [2].

When applying the above sketching process to a public 32,738 × 2700 PBMC data set, representing single-cell mRNA expression vectors from blood cells, by making 10 sketches of the matrix with rank 500, 7 eigenvalues stand out about 1000 times greater than the predicted T-W critical eigenvalue (Figure 8). This, in turn, leads to a rank 7 signal. Higher-dimensional sketches of rank 1000, and rank 2000 were also made, and the result was corroborated; however, taking the entire 32,738-dimensional would have required a much more powerful computer and taken a lot longer.

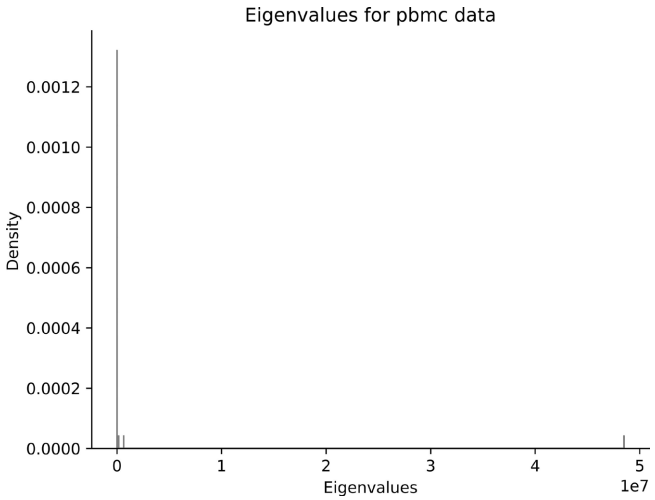


Figure 8. Eigenvalues of a 500 × 500 sketch of the described PBMC data set. While most of the eigenvalues fall within the range of the M-P curve, there is a small but notable spike far out from the rest of the eigenvalues [2].

6. Conclusions

Despite its seemingly random nature, there are many mathematical patterns in the world of random matrix theory. It is, therefore, straightforward to analyze random processes like unicellular data or Hamiltonian nuclei that, even while affected by error in human measurements, still lead to convincing conclusions about the behavior of these systems. In addition, universality with respect to mean, standard deviation, and distribution of the entries of a random matrix further highlights the predictable properties of random matrices.

This approach has been used in many instances to separate noise and improve calculation efficiency while maintaining an accurate depiction of the properties of a matrix that would otherwise be difficult to glean any information from. With the rise in the importance of matrices in emerging fields like biotechnology and artificial intelligence, they will be a very important tool for solving future problems.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Pedregoza, F., Paquette, C., Trogdon, T. and Pennington, J. (n.d.) (2024) Random Matrix Theory and Machine Learning Tutorial [PowerPoint Slides]. https://random-matrix-learning.github.io/#presentation1

[2] Jiang, N. (2024) Jiang Random Matrix Research. https://github.com/nathanjiang100/Jiang-Random-Matrix-Research

[3] Edelman, A. and Rao, N.R. (2005) Random Matrix Theory. *Acta Numerica*, **14**, 233-297. https://doi.org/10.1017/s0962492904000236

[4] Aparicio, L., Bordyuh, M., Blumberg, A.J. and Rabadan, R. (2020) A Random Matrix Theory Approach to Denoise Single-Cell Data. *Patterns*, **1**, Article 100035. https://doi.org/10.1016/j.patter.2020.100035

[5] McGregor, A. (n.d.) (2024) Linear Sketches with Applications to Data Streams [PowerPoint Slides]. University of Massachusetts Amherst. https://people.cs.umass.edu/~mcgregor/stocworkshop/mcgregor.pdf