A Review of Gaussian Random Matrices

Department of Mathematics, Linköping University

Kasper Andersson

LiTH-MAT-EX–2020/08–SE

Credits: **16hp**

Level: **G2**

Supervisor: **Martin Singull**,
Department of Mathematics, Linköping University

Examiner: **Xiangfeng Yang**,
Department of Mathematics, Linköping University

Linköping: **2020**

Abstract

While many university students get introduced to the concept of statistics early in their education, random matrix theory (RMT) usually first arises (if at all) in graduate level classes. This thesis serves as a friendly introduction to RMT, which is the study of matrices with entries following some probability distribution. Fundamental results, such as Gaussian and Wishart ensembles, are introduced and a discussion of how their corresponding eigenvalues are distributed is presented. Two well-studied applications, namely neural networks and PCA, are discussed where we present how RMT can be applied.

Keywords:

Random Matrix Theory, Gaussian Ensembles, Covariance, Wishart Ensembles, PCA, Neural Networks

URL for electronic version:

`http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-171649`

Sammanfattning

Medan många stöter på statistik och sannolikhetslära tidigt under sina universitetsstudier så är det sällan slumpmatristeori (RMT) dyker upp förän på forskarnivå. RMT handlar om att studera matriser där elementen följer någon sannolikhetsfördelning och den här uppsatsen presenterar den mest grundläggande teorin för slumpmatriser. Vi introducerar Gaussian ensembles, Wishart ensembles samt fördelningarna för dem tillhörande egenvärdena. Avslutningsvis så introducerar vi hur slumpmatriser kan användas i neruonnät och i PCA.

Nyckelord:

Slumpmatristeori, Gaussian Ensembles, Kovarians, Wishart Ensembles, PCA, Neuronnät

URL för elektronisk version:

`http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-171649`

Acknowledgements

I would like to thank my supervisor Martin Singull who introduced me to the topics of Random Matrix Theory, which I would never look upon by myself.

I also want to thank my mother, who always supported me throughout my studies.

Contents

Chapter 1

Introduction

A random matrix is a matrix $\boldsymbol{M}$ with its entries m_{ij} being random variables. Two observations of 3×3 random matrices, both with elements being sampled uniformly on the interval [0,1] can be

$$\begin{pmatrix} 0.9575 & 0.9706 & 0.8003 \\ 0.9649 & 0.9572 & 0.1419 \\ 0.1576 & 0.4854 & 0.4218 \end{pmatrix}, \quad \begin{pmatrix} 0.9157 & 0.6557 & 0.9340 \\ 0.7922 & 0.0357 & 0.6787 \\ 0.9595 & 0.8491 & 0.7577 \end{pmatrix}.$$

We can do the same thing for random matrices with its elements being sampled from $\mathcal{N}(0, 1)$

$$\begin{pmatrix} 0.7269 & -0.7873 & -1.0689 \\ -0.3034 & 0.8884 & -0.8095 \\ 0.2939 & -1.1471 & -2.9443 \end{pmatrix}, \quad \begin{pmatrix} 1.4384 & 1.3703 & -0.2414 \\ 0.3252 & -1.7115 & 0.3192 \\ -0.7549 & -0.1022 & 0.3129 \end{pmatrix}.$$

Again, the elements of the two matrices differ from each other, which will be the general case when we are working with random matrices.

When we are studying random matrices, we are not interested in a specific sample, but rather a model of all the instances it can take. These models are known as *matrix ensembles*.

As always when we are working with matrices, eigenvalues are of great interest. Many of the results presented in this thesis will therefore be about how the eigenvalues are distributed for different kind of random matrices. Since the entries of random matrices are random themselves, the corresponding eigenvalues will be random as well. A major result in random matrix theory can be observed

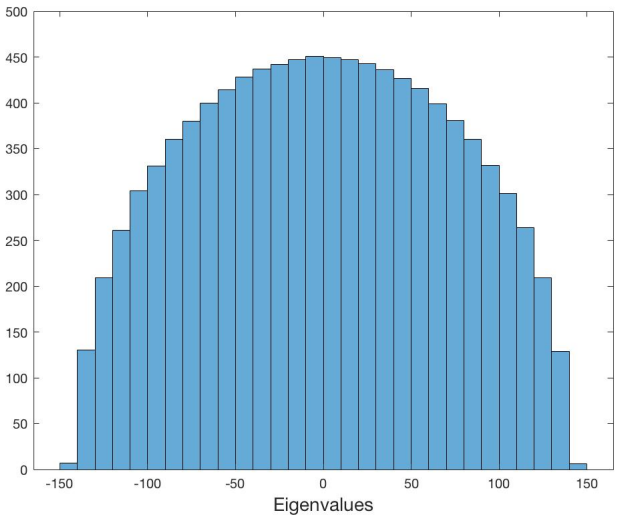

Figure 1.1: Empirical distribution of the eigenvalues for a 10000×10000 random matrix.

by plotting the eigenvalues corresponding to a symmetric matrix with entries being sampled from a standard normal distribution. The occurring elliptical shape, known as *Wigner's semicircle*, is demonstrated in Figure 1.1.

1.1 Aims and Outline

Most of the existing literature associated with random matrices is for graduate studies or research, despite the fact that many results can be well understood and interpreted by students on an undergraduate level. This thesis aims to give the reader a friendly introduction to the topics associated with random matrix theory.

Chapter 1 starts of by introducing the concept of matrix ensembles and briefly generalizes random variables into random matrices. Mathematical results which will be needed to develop the theory of random matrices will be presented here as well.

Chapter 2 collects the core of random matrix theory. The reader is introduced to Gaussian ensembles and to their corresponding eigenvalue distribution. We follow up with the definition of covariance matrices and their properties. Focusing on covariance matrices built upon Gaussian entries, a discussion of how these matrices and their eigenvalues are distributed is represented.

Chapter 3 aims to introduce the reader to how random matrix theory can be applied. First off, the idea of standard neural networks and how they are used is represented. We follow up with the concept of extreme learning machines and summarizes a framework, established by Z. Liao, of how random matrices can be used in neural networks. Lastly, we derive the concept of standard PCA and represent how statistical inference may be worked out when the underlying data follows a normal distribution.

Chapter 4 serves as a final discussion of what we have accomplished in this thesis and how the topics of random matrix could be further developed.

1.2 Preliminaries

Definition 1.1. *A matrix $\boldsymbol{M}$ with entries m_{ij} is **Hermitian** if $m_{ij} = \overline{m}_{ji}$ where $\overline{z}$ is the complex conjugate of z.*

Remark: If all entries in the matrix are real, then the matrix being Hermitian is equivalent to it being symmetric.

Definition 1.2. *The **spectrum** of a matrix $\boldsymbol{M}$ is given by the set $\{\lambda \in \mathbb{C} : \det(\boldsymbol{M} - \lambda \boldsymbol{I}) = 0\}$ where λ is an **eigenvalue** of $\boldsymbol{M}$ and $\boldsymbol{I}$ is the identity matrix.*

Theorem 1.1. *If $X_1, \ldots, X_n$ are i.i.d. random variables with corresponding probability density functions (pdf) $f(x_i)$ then their joint pdf is given by*

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) \, .$$

Remark: We will simply denote the joint pdf for the matrix entries $m_{11}, \ldots, m_{nn}$ of a $n \times n$ matrix $\boldsymbol{M}$ as $f(\boldsymbol{M})$.

Definition 1.3. *The **expected value** (mean) for a continuous random variable X with corresponding pdf $f(x)$ is given by*

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx \, .$$

Remark: For a discrete random variable, replace integral with proper sum and pdf with probability mass function (pmf).

Definition 1.4. *The **variance** for a random variable X is given by*

$$Var[X] = E[(X - E[X])^2] \,.$$

Definition 1.5. *A random variable X is **Gaussian** (normal) **distributed**, with mean μ and variance σ^2 if its pdf is given by*

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in (-\infty, \infty)$$

and we denote this by $X \sim \mathcal{N}(\mu, \sigma^2)$.

Definition 1.6. *The **gamma function** $\Gamma(\cdot)$ is given by*

$$\Gamma(x) = \int_0^\infty y^{x-1} e^{-y} dy \,.$$

Definition 1.7. *We say that X is **chi squared distributed** with n degrees of freedom if its pdf is given by*

$$f(x|n) = \frac{1}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \,, \qquad x \in (0, \infty) \,,$$

which we denote as $X \sim \chi^2(n)$.

Theorem 1.2. *If $X_1, X_2, \ldots, X_n$ are all i.i.d. $\mathcal{N}(0, 1)$ then it follows that*

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n) \,.$$

Definition 1.8. *X is **Rayleigh distributed** with scale parameter $b > 0$ if its pdf is given by*

$$f(x|b) = \frac{x}{b^2} e^{\frac{-x^2}{2b^2}} \,, \qquad x \in (0, \infty) \,,$$

which we denote as $X \sim Rayleigh(b)$.

Lemma 1.1. *If $X \sim \mathcal{N}(0, \sigma^2)$ and $Y \sim \mathcal{N}(0, \sigma^2)$ are independent, then it follows that*

$$\sqrt{X^2 + Y^2} \sim Rayleigh(\sigma).$$

Proof: With Theorem 1.2 we get

$$\begin{aligned} P(\sqrt{X^2+Y^2} \leqslant t) &= P(X^2+Y^2 \leqslant t^2) \\ &= P\left(\frac{X^2}{\sigma^2} + \frac{Y^2}{\sigma^2} \leqslant \frac{t^2}{\sigma^2}\right) \\ &= \int_0^{\frac{t^2}{\sigma^2}} \frac{1}{2} e^{-\frac{x}{2}} dx \\ &= 1 - e^{\frac{-t^2}{2\sigma^2}}. \end{aligned}$$

Taking derivative with the respect to t finally gives us

$$f(t) = \frac{t}{\sigma^2} e^{\frac{-t^2}{2\sigma^2}}$$

and we are done. □

Definition 1.9. *The **trace** for a $n \times n$ matrix $\boldsymbol{M}$ with elements m_{ij} is given by*

$$\mathrm{tr}(\boldsymbol{M}) = \sum_{i=1}^{n} m_{ii}$$

Lemma 1.2. *Given a matrix $\boldsymbol{M}$ with eigenvalues $\lambda_1, \lambda_2 \ldots, \lambda_n$ we have that*

$$\mathrm{tr}(\boldsymbol{M}^n) = \sum_{i=1}^{n} \lambda_i^n \,.$$

Definition 1.10. *A matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ is **positive semi-definite** if $\boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x} \geqslant 0$ for all vectors $\boldsymbol{x} \in \mathbb{R}^{n \times 1}$.*

Lemma 1.3. *A symmetric matrix is positive semi-definite if and only if all of its eigenvalues are non-negative.*

Lemma 1.4. *The eigenvectors corresponding to different eigenvalues of a symmetric matrix are orthogonal to each other.*

Proof: Consider a symmetric matrix $\boldsymbol{X}$ with eigenvalues $\lambda_i \neq \lambda_j$ and corresponding eigenvectors $\boldsymbol{x}_i$, $\boldsymbol{x}_j$. We have that $\boldsymbol{x}_j^T \boldsymbol{X} \boldsymbol{x}_i = \boldsymbol{x}_j^T \lambda_i \boldsymbol{x}_i$ and $\boldsymbol{x}_i^T \boldsymbol{X} \boldsymbol{x}_j = \boldsymbol{x}_i^T \lambda_j \boldsymbol{x}_j$ and since $\boldsymbol{X}$ is symmetric, $\boldsymbol{x}_j^T \boldsymbol{X} \boldsymbol{x}_i = \boldsymbol{x}_i^T \boldsymbol{X} \boldsymbol{x}_j$. Together, this gives us the equality

$$(\lambda_i - \lambda_j) \boldsymbol{x}_i^T \boldsymbol{x}_j = 0 \,,$$

and thus, $\boldsymbol{x}_i^T \boldsymbol{x}_j = 0$. □

Definition 1.11. *A sequence of random variables X_n* ***converges almost surely (a.s.)*** *to the random variable X as $n \to \infty$ if*

$$P\left(\{\omega : X_n(\omega) \to X(\omega) \, as \, n \to \infty\}\right) = 1.$$

Definition 1.12. *A sequence of random variables X_n* ***converges in probability*** *to the random variable X as $n \to \infty$ if $\forall \varepsilon > 0$,*

$$P(|X_n - X| > \varepsilon) \to 0 \quad as \quad n \to \infty.$$

Definition 1.13. *A real valued function f is said to be* ***Lipschitz continuous*** *if there exists a constant C for all $x \neq y$ such that*

$$\left| \frac{f(x) - f(y)}{x - y} \right| \leqslant C.$$

Theorem 1.3. *Let $\mathbb{V}$ be a linear subspace of $\mathbb{R}^n$, $\boldsymbol{u} \in \mathbb{R}^n$ and $\boldsymbol{u}_{||\mathbb{V}}$ the orthogonal projection of $\boldsymbol{u}$ onto $\mathbb{V}$. Then, for all $\boldsymbol{v} \in \mathbb{V}$, we have that*

$$\left\| \boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v} \right\| \leqslant \left\| \boldsymbol{u} - \boldsymbol{v} \right\|.$$

Proof: We have

$$\boldsymbol{u} - \boldsymbol{v} = (\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}}) + (\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v}) \Rightarrow \left\| \boldsymbol{u} - \boldsymbol{v} \right\|^2 = \left\| (\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}}) + (\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v}) \right\|^2$$

expanding the right hand side gives us

$$\begin{aligned}
\left\| (\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}}) + (\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v}) \right\|^2 &= \big((\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}}) + (\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v}) \big)^T \\
&\quad \cdot \big((\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}}) + (\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v}) \big) \\
&= (\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}})^T (\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}}) \\
&\quad + (\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}})^T (\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v}) \\
&\quad + (\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v})^T (\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}}) \\
&\quad + (\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v})^T (\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v}) \\
&= (\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}})^T (\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}}) \\
&\quad + (\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v})^T (\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v}) \\
&= \left\| \boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}} \right\|^2 + \left\| \boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v} \right\|^2,
\end{aligned}$$

as a result of orthogonality between $(\boldsymbol{u} - \boldsymbol{u}_{||\mathbb{V}})$ and $(\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v})$. Finally, due to the non-negative euclidean norm, we obtain

$$\left\| \boldsymbol{u} - \boldsymbol{v} \right\|^2 \geqslant \left\| \boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v} \right\|^2, \tag{1.1}$$

which is equivalent to

$$\|\boldsymbol{u} - \boldsymbol{v}\| \geqslant \left\|\boldsymbol{u}_{||\mathbb{V}} - \boldsymbol{v}\right\|, \tag{1.2}$$

and we are done. □

Definition 1.14. *The* ***Frobenius norm****, denoted* $\|\cdot\|_F$*, of a matrix* $\boldsymbol{A}$ *is given by*

$$\|\boldsymbol{A}\|_F = \sqrt{\operatorname{tr}(\boldsymbol{A}\boldsymbol{A}^T)}.$$

Chapter 2

Random Matrices

2.1 Gaussian Ensembles

A cornerstone in random matrix theory (RMT) is to study the spectrum of random matrices. It's difficult to say anything about the eigenvalues for an arbitrary matrix with random entries, so we will focus on symmetric random matrices.

Definition 2.1. *A $n \times n$ matrix $\boldsymbol{M}$ with entries m_{ij} is a **Wigner matrix** if $m_{ij} = m_{ji}$ and m_{ij} are randomly i.i.d. up to symmetry.*

E. P. Wigner used symmetric random matrices to study nuclear energy levels [23]. Another assumption which gives room for more calculations is to consider the entries being Gaussian.

Definition 2.2. *Let $\boldsymbol{M}$ be a real Wigner matrix with diagonal entries $m_{ii} \sim \mathcal{N}(0,1)$ and $m_{ij} \sim \mathcal{N}(0,1/2)$, $i \neq j$. $\boldsymbol{M}$ is said to belong to the **Gaussian orthogonal ensemble** (GOE).*

One can easily achieve the required properties for a matrix to be considered as a GOE matrix. Consider a random $n \times n$ matrix $\boldsymbol{A}$ with all entries being $\mathcal{N}(0,1)$. Then

$$\boldsymbol{M} = \frac{\boldsymbol{A} + \boldsymbol{A}^T}{2} \tag{2.1}$$

is a GOE matrix.

We can derive a more practical way to describe the GOE.

Theorem 2.1. *The joint pdf for the Gaussian orthogonal ensemble is given by*

$$f(\boldsymbol{M}) = \frac{1}{2^{n/2}} \frac{1}{\pi^{n(n+1)/4}} e^{-\frac{1}{2} tr(\boldsymbol{M}^2)} .$$

Proof: Let m_{ij} be the entries for a $n \times n$ matrix $\boldsymbol{M}$ generated by GOE. Recall, all entries in $\boldsymbol{M}$ are independent so their joint pdf is given by

$$f(\boldsymbol{M}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \prod_{i=j} e^{-\frac{m_{ij}^2}{2}} \left(\frac{1}{\sqrt{\pi}}\right)^{\frac{n(n-1)}{2}} \prod_{1 \leqslant i < j \leqslant n} e^{-m_{ij}^2}. \tag{2.2}$$

Using exponential rules and the fact that $\boldsymbol{M}$ is symmetric, we can rewrite the joint pdf as

$$\begin{aligned} C_n \prod_{i=j} e^{-\frac{m_{ij}^2}{2}} \prod_{1 \leqslant i < j \leqslant n} e^{-m_{ij}^2} &= C_n e^{-\frac{1}{2} \sum\limits_{i=j} m_{ij}^2 - \sum\limits_{1 \leqslant i < j \leqslant n} m_{ij}^2} \\ &= C_n e^{-\frac{1}{2}\left(\sum\limits_{i=j} m_{ij}^2 + 2 \sum\limits_{1 \leqslant i < j \leqslant n} m_{ij}^2\right)} \\ &= C_n e^{-\frac{1}{2} \mathrm{tr}(\boldsymbol{M}^2)} , \end{aligned} \tag{2.3}$$

where

$$\begin{aligned} C_n &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{\pi}}\right)^{\frac{n(n-1)}{2}} = \frac{1}{2^{n/2}} \frac{1}{\pi^{n/2}} \frac{1}{\pi^{n(n-1)/4}} \\ &= \frac{1}{2^{n/2}} \frac{1}{\pi^{n(n+1)/4}} \end{aligned} \tag{2.4}$$

is a normalization constant. □

Remark: The matrices sampled from the GOE are *not* orthogonal. The orthogonal in GOE comes from the fact that GOE matrices are invariant under orthogonal transformations, which means that $\boldsymbol{M}$ and $\boldsymbol{Q}^T \boldsymbol{M} \boldsymbol{Q}$ has the same distribution, where $\boldsymbol{Q}$ is an orthogonal matrix [1].

Matrices sampled from the GOE only contain real entries. Two other common Gaussian ensembles are the *Gaussian unitary ensemble* (GUE) and the *Gaussian symplectic ensemble* (GSE).

A random Hermitian $n \times n$ matrix $\boldsymbol{M}$ is said to belong to the GUE if the diagonal entries $m_{jj} \sim \mathcal{N}(0, 1)$ and the upper triangular entries are given by

$m_{jk} = u_{jk} + iv_{jk}$ where $u_{jk}, v_{jk} \sim \mathcal{N}(0, 1/2)$. The pdf for GUE is given by

$$f(\boldsymbol{M}) = \prod_{j=1}^{n} \frac{1}{\sqrt{\pi}} e^{-m_{jj}^2} \prod_{1 \leqslant j < k \leqslant n} \frac{2}{\pi} e^{-2|m_{jk}^2|} = C_n e^{-\mathrm{tr}(\boldsymbol{M})}\,, \tag{2.5}$$

where

$$C_n = \left(\frac{1}{\sqrt{\pi}}\right)^n \left(\frac{2}{\pi}\right)^n = \frac{2^n}{\pi^{3n/2}} \tag{2.6}$$

is a normalization constant. The U stands for GUE matrices being invariant under unitary transformations [8].

We will not discuss the GSE beside stating that its matrix entries are real quaternions.

2.2 Eigenvalues

We shall now look at the spectrum for random matrices with Gaussian entries. We will mostly limit ourselves to the GOE case but some more general results are presented as well.

Theorem 2.2. (Section 1.2 [8]) *The joint pdf for the eigenvalues of a $n \times n$ GOE matrix is given by*

$$f(\lambda_1, \ldots, \lambda_n) = \frac{1}{C_n} e^{-\frac{1}{2}\sum_{k=1}^{n} \lambda_i^2} \prod_{1 \leqslant i < j \leqslant n} |\lambda_j - \lambda_i|\,,$$

where C_n is a normalization constant.

We shall only show that Theorem 2.2 holds for the simplest case, that is, we want to show that

$$f(\lambda_1, \lambda_2) = k e^{-\frac{1}{2}(\lambda_1^2 + \lambda_2^2)} |\lambda_1 - \lambda_2| \tag{2.7}$$

is the correct eigenvalue distribution for a 2×2 GOE matrix, where k is normalization constant.

Proof (Following from [7]): We will accomplish (2.7) by transforming the GOE pdf in terms of its eigenvalues. Recall that GOE matrices are invariant under orthogonal transformation, so consider $\boldsymbol{M} = \boldsymbol{Q}^T \boldsymbol{M}_\lambda \boldsymbol{Q}$, where $\boldsymbol{M} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}$

is a GOE matrix, $\boldsymbol{Q} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$ is an orthogonal matrix and $\boldsymbol{M}_\lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$. We have that

$$\begin{aligned} \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} &= \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}^T \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta & (\lambda_1 - \lambda_2)\sin\theta\sin\theta \\ (\lambda_1 - \lambda_2)\sin\theta\sin\theta & \lambda_1 \sin^2\theta + \lambda_2 \cos^2\theta \end{pmatrix} \end{aligned} \tag{2.8}$$

that is

$$\begin{cases} m_{11} = \lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta, \\ m_{22} = \lambda_1 \sin^2\theta + \lambda_2 \cos^2\theta, \\ m_{12} = \frac{1}{2}(\lambda_1 - \lambda_2)\sin 2\theta. \end{cases} \tag{2.9}$$

Note that $m_{12} = m_{21}$ due to symmetry, so we only need to consider one of the entries. To ensure that our function in the terms of its new variables (the eigenvalues) is a valid pdf, we need to take the Jacobian determinant into account. The Jacobian is given by

$$\mathcal{J} = \begin{pmatrix} \cos^2\theta & \sin^2\theta & (\lambda_2 - \lambda_1)\sin 2\theta \\ \sin^2\theta & \cos^2\theta & (\lambda_1 - \lambda_2)\sin 2\theta \\ \frac{1}{2}\sin 2\theta & -\frac{1}{2}\sin 2\theta & (\lambda_1 - \lambda_2)\cos 2\theta \end{pmatrix} \tag{2.10}$$

and

$$\begin{aligned} \det(\mathcal{J}) &= \det \begin{pmatrix} \cos^2\theta & \sin^2\theta & (\lambda_2 - \lambda 1)\sin 2\theta \\ \sin^2\theta & \cos^2\theta & (\lambda_1 - \lambda_2)\sin 2\theta \\ \frac{1}{2}\sin 2\theta & -\frac{1}{2}\sin 2\theta & (\lambda_1 - \lambda_2)\cos 2\theta \end{pmatrix} \\ &= \lambda_1 \cos^4\theta \cos 2\theta - \lambda_2 \cos^4\theta \cos 2\theta - \lambda_2 \sin^2 2\theta \\ &\quad + \lambda_1 \sin^2 2\theta + x \sin^2(2\theta) - \lambda_1 \sin^4\theta \cos 2\theta \\ &\quad + \lambda_2 \sin^4\theta \cos 2\theta \\ &= (\lambda_1 - \lambda_2)(\cos^2 2\theta + \sin^2 2\theta) \\ &= \lambda_1 - \lambda_2 \end{aligned} \tag{2.11}$$

Now, the change of variable can be done as $f(m_{11}, m_{22}, m_{12}) \to f(\lambda_1, \lambda_2, \theta)|\det(\mathcal{J})|$ $=: \hat{f}(\lambda_1, \lambda_2, \theta)$, where $f(m_{11}, m_{22}, m_{12})$ is the GOE pdf. We have that

$$\hat{f}(\lambda_1, \lambda_2, \theta) = k' e^{-\frac{1}{2}(\lambda_1^2 + \lambda_2^2)} |\lambda_1 - \lambda_2| \,, \tag{2.12}$$

where k' is a constant. To get rid of θ, all we need to do is integrate with respect to θ which only contributes to the constant value. We have

$$\begin{aligned}\hat{f}(\lambda_1, \lambda_2) &= \int_\theta k' e^{-\frac{1}{2}(\lambda_1^2+\lambda_2^2)} |\lambda_1 - \lambda_2| d\theta \\ &= k e^{-\frac{1}{2}(\lambda_1^2+\lambda_2^2)} |\lambda_1 - \lambda_2|.\end{aligned} \tag{2.13}$$

□

Next thing we are going to look at is the distribution of the distance between two eigenvalues from a 2×2 GOE matrix.

Theorem 2.3. *The pdf for the distance $d = |\lambda_1 - \lambda_2|$ between the two eigenvalues λ_1, λ_2 in a 2×2 GOE matrix is given by*

$$f(d) = \frac{d}{2} e^{-\frac{d^2}{4}} .$$

Proof: Let $\boldsymbol{M} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}$ be GOE matrix with eigenvalues λ_1, λ_2 such that $\lambda_1 > \lambda_2$ and consider $d = \lambda_1 - \lambda_2$. Solving $\det \begin{pmatrix} m_{11} - \lambda & m_{12} \\ m_{21} & m_{22} - \lambda \end{pmatrix} = 0$ gives us

$$\begin{aligned}\lambda_1 &= \frac{m_{11} + m_{22} + \sqrt{(m_{11} - m_{22})^2 + 4 m_{12} m_{21}}}{2} , \\ \lambda_2 &= \frac{m_{11} + m_{22} - \sqrt{(m_{11} - m_{22})^2 + 4 m_{12} m_{21}}}{2}\end{aligned}$$

and therefore

$$d = \sqrt{(m_{11} - m_{22})^2 + 4 m_{12}^2} ,$$

where we used that $m_{12} = m_{21}$ due to symmetry. Now let $X = m_{11} - m_{22}$, $Y = 2m_{12}$. We have that $X \sim \mathcal{N}(0, 2)$, $Y \sim \mathcal{N}(0, 2)$ so

$$d = \sqrt{X^2 + Y^2} \sim Rayleigh(\sqrt{2}) \tag{2.14}$$

according to Lemma 1.1. The pdf for d is therefore given by

$$f(d) = \frac{d}{2} e^{-\frac{d^2}{4}} . \tag{2.15}$$

□

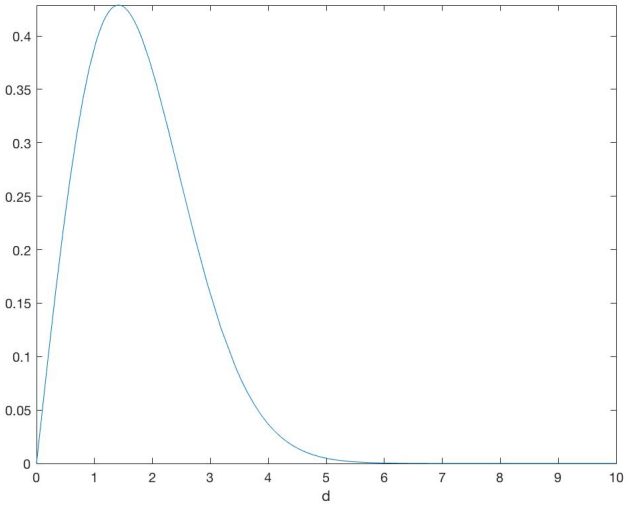

Figure 2.1: Pdf for distance between the two eigenvalues for a 2×2 GOE matrix.

What Theorem 2.3 actually says is that the eigenvalues do not want to be too close to each other, neither do they want to be too far apart. For d close to 0, the linear factor will dominate while for large d, the exponential factor will take over, see Figure 2.1.

One may rescale (2.15) to

$$f(s) = \frac{\pi s}{2} e^{-\frac{\pi s^2}{4}} , \tag{2.16}$$

which is known as *Wigner's surmise* [18].

Wigner suggested (2.16) to calculate the energy level spacings in the atomic nucleus where $s = \frac{\lambda_{i+1} - \lambda_i}{D}$ and D equals the mean distance for the energy levels. While (2.15) do gives us an correct pdf for the spacings between the eigenvalues, Wigner's surmise only works as an approximation for true value of the energy level spacings. The error may be up to 2% [9].

As we mentioned earlier there exist more Gaussian ensembles than the orthogonal one. Each of these ensembles have their own eigenvalue distribution as

well.

Theorem 2.4. (Section 1.3 [8]) *The joint pdf for the eigenvalues of the GOE ($\beta = 1$), the GUE ($\beta = 2$) and the GSE ($\beta = 4$) is given by*

$$f(\lambda_1, \ldots, \lambda_n) = \frac{1}{G_{n,\beta}} e^{-\frac{\beta}{2}\sum_{k=1}^{n}\lambda_i^2} \prod_{1 \leqslant i < j \leqslant n} |\lambda_j - \lambda_i|^\beta ,$$

where β is known as the Dyson index and

$$G_{n,\beta} = \beta^{-n/2 - n\beta(n-1)/4}(2\pi)^{n/2} \prod_{j=0}^{n-1} \frac{\Gamma(1 + (j+1)\beta/2)}{\Gamma(1 + \beta/2)}$$

is a normalization constant.

So how do we actually calculate the probability that an eigenvalue ends up within a given interval $\mathcal{I}$? Normally, this would be done as

$$P(\lambda \in \mathcal{I}) = \int_{\mathcal{I}} f(\lambda) d\lambda. \tag{2.17}$$

But how do we get the pdf for the eigenvalue? A good guess could be to integrate away the other variables in the joint pdf given in Theorem 2.4, that is

$$f(\lambda) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(\lambda, \lambda_2, \ldots, \lambda_n) d\lambda_2 \ldots d\lambda_n. \tag{2.18}$$

However, computing this multiple integral is far from trivial, even if we limit ourselves to the GOE. As we saw earlier in Figure 1.1, it seems like the eigenvalue distribution takes on a epileptic shape.

Theorem 2.5. (Section 1.4 [8]) *As the size n of a matrix $\boldsymbol{M}$ from the Gaussian ensemble approaches infinity, the eigenvalue distribution for the scaled matrix $\frac{\boldsymbol{M}}{\sqrt{2n}}$ has a limit*

$$f(\lambda) = \frac{2}{\pi}\sqrt{1 - \lambda^2}\,, \qquad \lambda \in [-1, 1]\,.$$

Theorem 2.5 is known as *Wigner's semicircle law*. Wigner was first to prove it with the help of combinatorics in the 1950's. Later on, M. L. Mehta and M. Gaudin [18] proved it by computing the integral given in (2.18) and then taking the limit $n \to \infty$.

The curve generated for the semicircle law won't be a half-circle, but rather a half-ellipse, see Figure 2.2. Depending on how you choose to scale your matrix you will get different intervals for your elliptic curve.

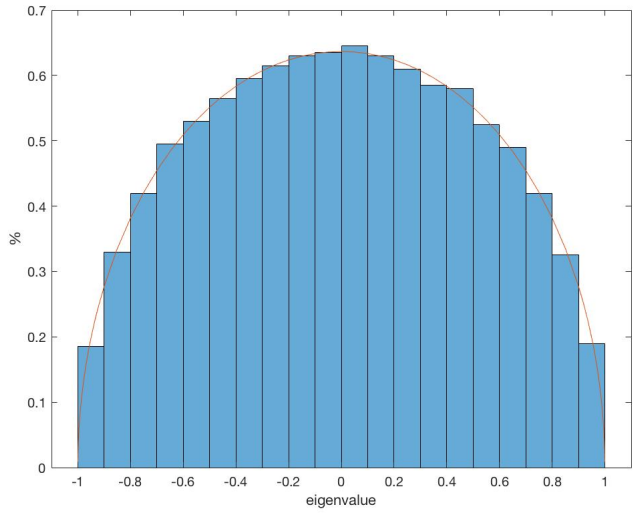

Figure 2.2: Eigenvalue distribution of 1000×1000 GOE matrix together with semicircle law.

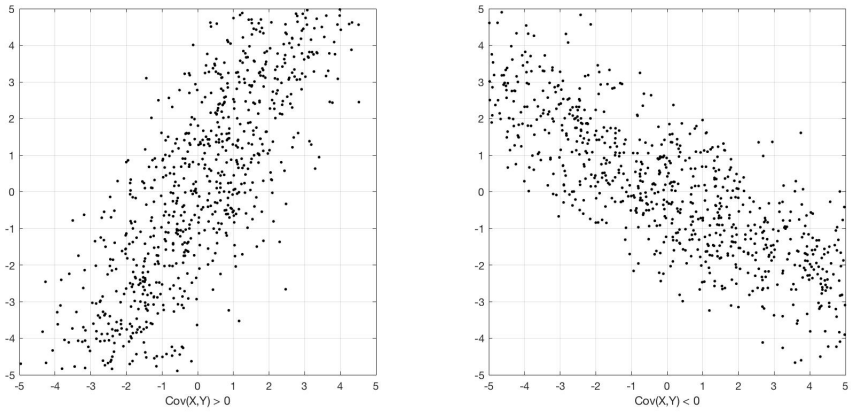

Figure 2.3: Plots of two data sets with positive and negative covariance respectively.

2.3 Covariance and Wishart Ensembles

So far have we only discussed theory regarding Wigner matrices and their natural symmetric counterpart $(\boldsymbol{A} + \boldsymbol{A}^T)/2$, which mostly appear when modelling, for example, physical systems [23]. However, when we are studying applications of multivariate statistics, such as variance analysis, regression analysis, dimension reduction etc. we are often interested in variation and dependency in our input data. A more occurring matrix in these topics is the *covariance matrix*, which will be the focus of this section.

Definition 2.3. *The **covariance** between two random variables X and Y, where $E[X] = \mu_X, E(Y) = \mu_Y$, is given by*

$$Cov[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] \,.$$

If X and Y are independent then $Cov(X, Y) = 0$. However, $Cov(X, Y) = 0$ does **not** necessary mean that X and Y are independent. The covariance acts as a metric of how two random variables vary compared to each other. Positive/negative covariance means that X and Y tend to be large/small simultaneously. This is illustrated in Figure 2.3.

Definition 2.4. *The **covariance matrix** of a random vector $X \in \mathbb{R}^{p\times 1}$ is*

given by

$$\boldsymbol{C} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T],$$

where $\boldsymbol{\mu} = E[\boldsymbol{X}] = (E[X_1], E[X_2], \ldots, E[X_p])^T$ *is the mean vector of* $\boldsymbol{X}$.

Each element c_{ij} in the covariance matrix $\boldsymbol{C}$ denotes the covariance between X_i and X_j. Since $c_{ii} = Cov(X_i, X_i) = E[(X_i - \mu_{X_i})(X_i - \mu_{X_i})] = \sigma_i^2$ we have that the variance of X_i can be found on the diagonal of $\boldsymbol{C}$, as

$$\boldsymbol{C} = \begin{pmatrix} \sigma_1^2 & c_{12} & \cdots \\ c_{21} & \sigma_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}. \tag{2.19}$$

Note that $Cov(X, Y) = Cov(Y, X)$ so $\boldsymbol{C}$ is always symmetric. Also, for $\boldsymbol{v} \in \mathbb{R}^{n \times 1}$ we have that

$$\begin{aligned} \boldsymbol{v}^T \boldsymbol{C} \boldsymbol{v} &= \boldsymbol{v}^T E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T] \boldsymbol{v} \\ &= E[\boldsymbol{v}^T (\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{v}] \\ &= E[((\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{v})^T (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{v}] \\ &= E[\left\| (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{v} \right\|^2] \geqslant 0 \end{aligned} \tag{2.20}$$

so the covariance is positive semi-definite and therefore, due to Lemma 1.3, the eigenvalues of $\boldsymbol{C}$ are non-negative. What (2.20) actually tells us is that the variance (see Definition 1.4) for a linear combination $\boldsymbol{a}^T \boldsymbol{X}$ of random variables is given by $Var[\boldsymbol{a}^T \boldsymbol{X}] = \boldsymbol{a}^T \boldsymbol{C} \boldsymbol{a}$.

With the covariance matrix defined we may expand the single variable Gaussian distribution into its multivariate case.

Definition 2.5. *A random vector* $\boldsymbol{X} \in \mathbb{R}^{p \times 1}$ *has a* ***multivariate normal distribution*** *if its pdf is given by*

$$f(\boldsymbol{X} | \boldsymbol{\mu}, \boldsymbol{C}) = \frac{1}{(2\pi)^{p/2} \det(\boldsymbol{C})} e^{-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}) \boldsymbol{C}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})^T},$$

where $\boldsymbol{\mu}$ *is the mean vector and* $\boldsymbol{C}$ *is the covariance matrix. This will be denoted as* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$.

In most real-world scenarios, we don't know the true values of our parameters so we have to estimate them. For example, in the univariate case we, we estimate the mean μ with the sample mean $\bar{x} = \frac{1}{n} \sum x_i$ and variance is mostly estimated with the sample variance $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$. What we often desire from our estimators is that they are *unbiased*, which means that their expected value is the true value.

Definition 2.6. *Suppose we have n independent samples* (x_i, y_i). *The* ***sample covariance*** *is given by*

$$\hat{c} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

For the sample covariance, we have that

$$\begin{aligned}
\hat{c} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y}\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i + \sum_{i\neq j} x_i y_j\right)\right)
\end{aligned} \tag{2.21}$$

and calculating the corresponding expected value of $\hat{C}$ gives us

$$\begin{aligned}
E[\hat{C}] &= \frac{1}{n-1}(E[\sum_{i=1}^{n} X_i Y_i] - \frac{1}{n}(E[\sum_{i=1}^{n} X_i Y_i] + E[\sum_{i\neq j} X_i Y_i])) \\
&= \frac{1}{n-1}(nE[XY] - \frac{1}{n}(nE[XY] + n(n-1)E[X]E[Y])) \\
&= E[XY] - E[X]E[Y] \\
&= Cov(X, Y)
\end{aligned} \tag{2.22}$$

and thus, the sample covariance is an unbiased estimator of the true covariance. The last step can be derived by expanding the covariance formula in Definition 2.3.

Definition 2.7. *Suppose we have n independent samples* $\boldsymbol{x}_i \in \mathbb{R}^{p\times 1}$ $i = 1, \ldots, n$, *then the* ***sample*** *(empirical)* ***covariance matrix*** *is given by*

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T .$$

By evaluating the sum for the sample covariance matrix $\boldsymbol{S}$ as we did in (2.21), we get the more practical matrix notation

$$\boldsymbol{S} = \frac{1}{n-1}(\boldsymbol{X}\boldsymbol{X}^T - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T), \tag{2.23}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$ is a $p \times n$ data matrix with each column being a sample and $\bar{\boldsymbol{x}}$ is the sample mean.

We shall now consider the distribution for matrices on the form $\boldsymbol{X}\boldsymbol{X}^T$ where the entries $\boldsymbol{x}_i$ of $\boldsymbol{X}$ are normal distributed. In general, when $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$, we have the *non-central Wishart distribution* [13]. For the sample covariance matrix $\boldsymbol{S}$, we indirectly assume $\boldsymbol{\mu} = \boldsymbol{0}$ as a result of subtracting the sample mean from each $\boldsymbol{x}_i$. Assuming zero mean explicitly is rather common in RMT and statistical literature, since most interesting results will be the same as with $\boldsymbol{\mu} \neq \boldsymbol{0}$ [20].

Theorem 2.6. (Section 7.2 [2]) *Let $\boldsymbol{W} = n\boldsymbol{S}$ where $\boldsymbol{S}$ is the sample covariance matrix built upon $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{p \times n}$, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$. $\boldsymbol{W}$ is then called a **Wishart matrix** and has a (central) **Wishart Distribution**, denoted $W_p(n, \boldsymbol{C})$, with the pdf*

$$f(\boldsymbol{W}) = \frac{\det(\boldsymbol{W})^{\frac{n-p-1}{2}} \, e^{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1}\boldsymbol{W}\right)}}{2^{\frac{np}{2}} \, \pi^{\frac{p(p-1)}{4}} \det(\boldsymbol{C})^{\frac{n}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)}$$

Remark: To be consistent with the statistical literature we are referring to, we define the sample covariance matrix with a factor n^{-1} instead of $(n-1)^{-1}$. This change will be of minor importance for the theory we are representing and the empirical results will be sufficiently equal as n grows large.

Theorem 2.7. [6] *Let $\boldsymbol{X}$ be a $p \times n$ matrix ($n \geqslant p$) with i.i.d entries x_{ij} being standard normal distributed. Then the matrix $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{X}^T$ belongs to the **Wishart Ensemble** with the pdf*

$$f_\beta(\boldsymbol{W}) = \frac{1}{C_{p,n,\beta}} \det(\boldsymbol{W})^{\beta(p-n+1)/2-1} e^{-\frac{1}{2}Tr(\boldsymbol{W})} \,,$$

where $C_{p,n,\beta}$ is a normalization constant and $\beta = 1, 2, 4$ corresponds to the orthogonal (real), unitary (complex) and symplectic (quaternion) ensembles, respectively.

Remark: We have only defined covariance for real matrices in this chapter but one may easily get the complex case by changing the real transpose operator $(\cdot)^T$ to the complex one $(\cdot)^*$. Note that for $\beta = 1$ we have the Wishart distribution $W_p(n, \boldsymbol{I})$.

Like the GOE case, matrices from the Wishart orthogonal ensemble are not orthogonal, they are invariant under an orthogonal transformation. Also, since the

entries of a Wishart matrix are not independent in general, it is not as trivial to prove Theorem 2.7 than, for example, Theorem 2.1. A proof by A. Edelman for the real and complex case, with help of matrix factorization, can be found in [5].

Just as eigenvalues tells us how eigenvectors are stretched when multiplied with their corresponding matrix, eigenvalues of covariance matrices do inform us how the data is spread (more about this in section 3.2). Like the Gaussian ensembles, we have a joint eigenvalue distribution for each of the Wishart ensembles.

Theorem 2.8. *[6] The joint eigenvalue pdf for a matrix generated by the Wishart ensemble is given by*

$$f(\lambda_1, \ldots, \lambda_p) = c_{\beta,p} \prod_{i<j} |\lambda_i - \lambda_j|^\beta \prod_{i=1}^{n} \lambda_i^{\frac{\beta}{n}(p-n+1)-1} e^{-\frac{1}{2}\sum_{i=1}^{n}\lambda_i} ,$$

where $c_{\beta,p}$ is a normalization constant.

Since we only defined the Wishart ensemble for standard Gaussian entries, we do expect our covariance matrix to take the form of an identity matrix. One may mistakenly conclude assume that the eigenvalues will concentrate around 1. However, it turns out that $\lambda_1, \ldots, \lambda_p \approx 1$ only holds when the sample size is much larger than the dimension of our data.

Theorem 2.9. *[20] Consider a $p \times n$ matrix $\boldsymbol{X}$ ($n \geqslant p$) with i.i.d. entries x_{ij} being standard normal distributed and let $\gamma = \frac{p}{n} \in (0, 1]$. As the sample size n approaches infinity, the eigenvalue pdf for $\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T$ has a limit*

$$f(\lambda) = \frac{1}{2\pi\gamma\lambda}\sqrt{(b_+ - \lambda)(\lambda - b_-)}\,, \qquad \lambda \in [b_-, b_+]\,, \tag{2.24}$$

where $b_\pm = (1 \pm \sqrt{\gamma})^2$.

Theorem 2.9 is known as the *Marčenko–Pastur law*. We observe in Figure 2.4 that our prediction about the eigenvalues being close to 1 gets less accurate when the dimension p approaches the sample size n.

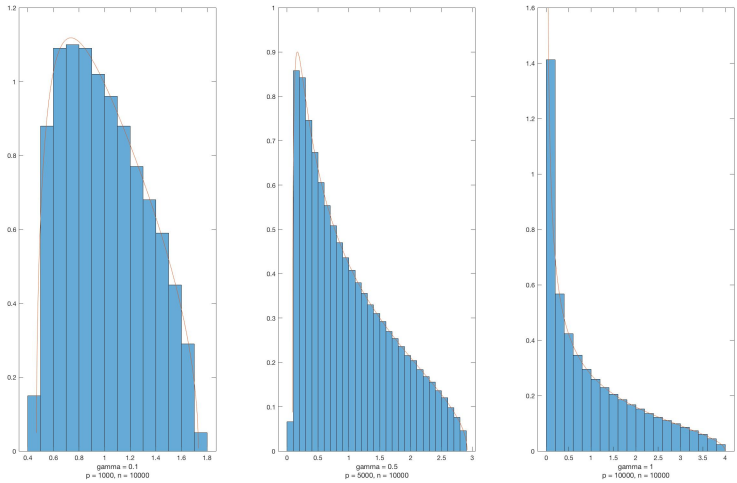

Figure 2.4: Empirical eigenvalue distribution of covariance matrices together with Marčenko–Pastur law for different $\gamma = \frac{p}{n}$ values.

Chapter 3

Applications

3.1 Random Matrices in Neural Networks

3.1.1 Neural Networks in General

Neural networks (NNs) are trained to predict/classify new data. The most basic approach is to train NNs with training samples $\{\boldsymbol{x}_i,y_i\}$ $i = 1,\ldots,n$ where $\boldsymbol{x}_i \in \mathbb{R}^p$ is a feature vector and y_i is the correct class label. The output is a function $f(\boldsymbol{x};\boldsymbol{w})$ which is a prediction of the true value y, where $\boldsymbol{w}$ is a weight vector that determines how important each of the components in $\boldsymbol{x}$ is.

The key idea is to find a function f and to adjust the parameters w such that the NN predicts class labels y with a high accuracy. We measure the accuracy of our NN with a *loss function* L, which we wish to minimize. The most elementary NNs uses $f(\boldsymbol{x};\boldsymbol{w}) = \boldsymbol{w}^T\boldsymbol{x}$, which results in the loss function

$$L = \sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2. \tag{3.1}$$

This is a least-square problem with an optimal solution $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}$ where $\boldsymbol{X} = [\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]$ is our data matrix and $\boldsymbol{y}$ contains all our class labels y_i. An illustration of a very simple NN is given in Figure 3.1.

This NN structure can only solve linearly-separable classification problems which puts heavy limitations on the practical uses of it and a classic example of its limitations is the XOR problem [4]. A solution to this, which is based on Cover's

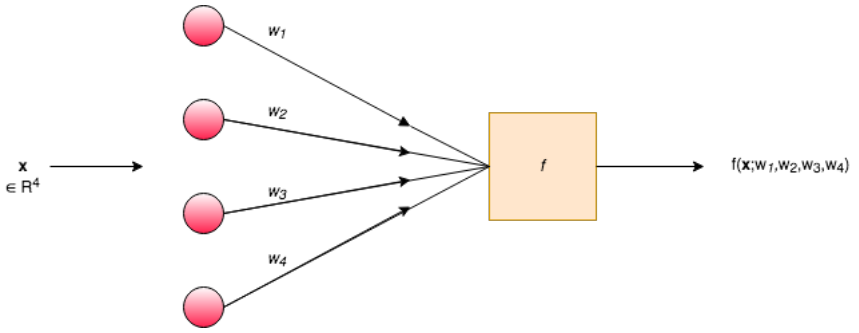

Figure 3.1: A very basic neural network with 4 dimensional feature vectors.

theorem [10], is to apply entry-wise a non-linear activation function $\sigma : \mathbb{R} \mapsto \mathbb{R}$ on our data. Our new loss function is then given by

$$L = \sum_{i=1}^{n} (y_i - \sigma(\boldsymbol{w}^T \boldsymbol{x}_i))^2 \,. \tag{3.2}$$

Due to the non-linear mapping, our new optimization problem is non-linear and our least-square solution will not work. In general, a NN with a non-linear activation function will result in a non-convex optimization problem which is considered hard to solve. NNs are in general built of hidden-layer(s) which are used to apply our activation function and the training time for these networks can be up to days [14]. This is due to the massive amounts of parameters in large NNs, which in general results in many local minima while seeking an optimal solution. See Figure 3.2 for an illustration of this[1].

A way to tackle this problem is to consider networks with a single hidden-layer where the input weights are randomly selected, so called *Extreme Learning Machines* (ELMs) [12]. ELMs has shown to perform really good for large non-complex data sets compared to other popular classification algorithms (such as support vector machines)[11]. The fast performance comes from the fact that we do not train the first layer in ELMs, so our non-linear optimization problem turns into a regression problem with a closed form solution.

Due to the random weights we may use results from RMT to model our ELMs. A framework for this will be represented in the next subsection, solely based on

[1]Figure is taken from https://github.com/tomgoldstein/loss-landscape

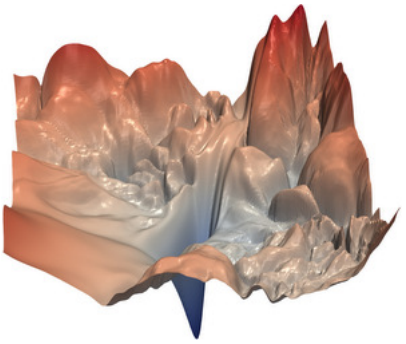

Figure 3.2: Illustration of a loss surface with many local minima .

the work by Z. Liao, C. Louart and R. Couillet [24] [19].

3.1.2 RMT Model for Neural Networks

Consider a data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{p \times n}$ with a target matrix[2] $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] \in \mathbb{R}^{d \times n}$, a random weight matrix $\boldsymbol{H} \in \mathbb{R}^{N \times p}$ with i.i.d. Gaussian standard entries and a weight matrix $\boldsymbol{\beta} \in \mathbb{R}^{N \times d}$ for the second layer. We shall denote the output from the hidden layer as $\boldsymbol{\Sigma} = \sigma(\boldsymbol{H}\boldsymbol{X}) \in \mathbb{R}^{N \times n}$, where $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is our activation function applied entry-wise, see figure Figure 3.3. As mentioned earlier, only the output layer $\boldsymbol{\beta}$ is to be trained, while $\boldsymbol{H}$ is static but randomly chosen. We seek weights $\boldsymbol{\beta}$ which minimizes the loss function

$$L(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{y_i} - \boldsymbol{\beta}^T \sigma(\boldsymbol{H}\boldsymbol{x_i}) \right\|^2 + \lambda \left\| \boldsymbol{\beta} \right\|_F^2 ,$$

for some regularization parameter $\lambda \geqslant 0$ to ensure our model is not overfitting. An explicit solution for this given by

$$\boldsymbol{\beta} = \frac{1}{n} \left(\frac{1}{n} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T + \lambda \boldsymbol{I}_N \right)^{-1} \boldsymbol{\Sigma}\boldsymbol{Y}^{\boldsymbol{T}} = \frac{1}{n} \boldsymbol{\Sigma} \left(\frac{1}{n} \boldsymbol{\Sigma}^T \boldsymbol{\Sigma} + \lambda \boldsymbol{I}_n \right)^{-1} \boldsymbol{Y}^T$$

which, per sample, results in a mean square training error

$$E_{train} = MSE_{train}(\lambda) = \frac{1}{n} \left\| \boldsymbol{\beta}^T \boldsymbol{\Sigma} - \boldsymbol{Y} \right\|_F^2 = \frac{\lambda^2}{n} \mathrm{tr} \left(\boldsymbol{Y}\boldsymbol{Q}^2\boldsymbol{Y}^T \right) , \tag{3.3}$$

[2]This is a more general notation of the one we used in (3.1), which allows us to have multiple output nodes. However, when we represent the results in the end of this chapter, we will use unidimensional output columns ($d = 1$) as in (3.1).

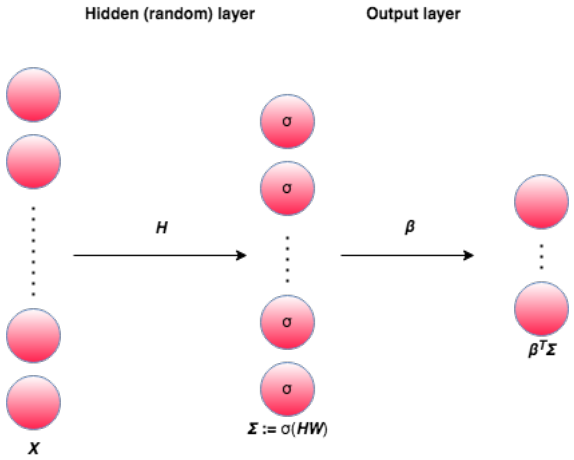

Figure 3.3: Sketch of our neural network (ELM) model with random weights.

where $\boldsymbol{Q} = \boldsymbol{Q}(\lambda) = \left(\frac{1}{n}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda \boldsymbol{I}_n\right)^{-1} \in \mathbb{R}^{n\times n}$ is known as the *resolvent* of the matrix $\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$. Once training is accomplished, we test our network by its capability to generalize. This is again measured in terms of mean squared error but on a previously unseen data set $\widehat{\boldsymbol{X}}$ of size $\hat{n}$ with corresponding outputs $\widehat{\boldsymbol{Y}}$, as

$$E_{test} = MSE_{test}(\lambda) = \frac{1}{\hat{n}} \left\| \boldsymbol{\beta}^{\boldsymbol{T}} \widehat{\boldsymbol{\Sigma}} - \boldsymbol{Y} \right\|_F^2 . \tag{3.4}$$

where $\widehat{\boldsymbol{\Sigma}} = \sigma(\boldsymbol{H}\widehat{\boldsymbol{X}})$.

Our goal is to predict how this network will perform. Due to the random weights, we estimate the performance by studying a *deterministic equivalent* $\bar{\boldsymbol{Q}}$ of $\boldsymbol{Q}$ (where all the randomness is captured). This means that we are looking for $\bar{\boldsymbol{Q}}$ such that

$$\frac{1}{n}\mathrm{tr}(\boldsymbol{A})(\boldsymbol{Q} - \bar{\boldsymbol{Q}}) \to 0, \quad \boldsymbol{a}^T(\boldsymbol{Q} - \bar{\boldsymbol{Q}})\boldsymbol{b} \to 0,$$

holds for deterministic (non-random) matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and unit vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, where convergence is understood a.s. or in probability. To accomplish this, we shall assume that our activation function σ is Lipschitz continuous. We shall also assume a bounded growth rate for our network, that is, as $n \to \infty$, we have

1) $\frac{p}{n} \to c_1$ and $\frac{N}{n} \to c_2$ where $c_1, c_2 \in (0, \infty)$, 2) Input matrix $\boldsymbol{X}$ has a bounded spectral norm, and 3) Output matrix $\boldsymbol{Y}$ has bounded entries.

Determining a deterministic equivalent to $\boldsymbol{Q}$ often boils down to calculating $E[\boldsymbol{Q}]$ which unfolds from the fact that $\left\|E[\boldsymbol{Q}] - \bar{\boldsymbol{Q}}\right\| \to 0$. Given the model we represented, Z. Liao shows that

$$\bar{\boldsymbol{Q}} = \bar{\boldsymbol{Q}}(\lambda) = (\check{\boldsymbol{K}} - \lambda \boldsymbol{I}_n)^{-1}, \quad \check{\boldsymbol{K}} = \frac{N}{n(1+\delta)} \boldsymbol{K},$$

where $\boldsymbol{K} = E_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^T \boldsymbol{X})^T \sigma(\boldsymbol{w}^T \boldsymbol{X})]$ is an *equivalent kernel matrix* for $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ and δ implicitly defined as the solution to $\delta = \frac{1}{n}\mathrm{tr}\left(\boldsymbol{K}\bar{\boldsymbol{Q}}\right)$ do result in

$$\left\|E[\boldsymbol{Q}] - \bar{\boldsymbol{Q}}\right\| \to 0,$$

as n, p, $N \to \infty$. This, together with (3.3) and (3.4), gives us the deterministic training and test mean-squared errors

$$\begin{aligned}
\bar{E}_{train} &= \frac{\lambda^2}{n}\mathrm{tr}\left(\boldsymbol{Y}\bar{\boldsymbol{Q}}\left(\frac{\frac{1}{N}\mathrm{tr}\left(\bar{\boldsymbol{Q}}\check{\boldsymbol{K}}\bar{\boldsymbol{Q}}\right)}{1 - \frac{1}{N}\mathrm{tr}\left(\check{\boldsymbol{K}}\bar{\boldsymbol{Q}}\check{\boldsymbol{K}}\bar{\boldsymbol{Q}}\right)}\check{\boldsymbol{K}} + \boldsymbol{I}_n\right)\bar{\boldsymbol{Q}}\boldsymbol{Y}^T\right), \\
\bar{E}_{test} &= \frac{1}{\hat{n}} = \left\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}\bar{\boldsymbol{Q}}\check{\boldsymbol{K}}_{\boldsymbol{X}\hat{\boldsymbol{X}}}\right\|_F^2 \\
&\quad + \frac{\frac{1}{N}\mathrm{tr}\left(\boldsymbol{Y}\bar{\boldsymbol{Q}}\check{\boldsymbol{K}}\bar{\boldsymbol{Q}}\boldsymbol{Y}^T\right)}{1 - \frac{1}{N}\mathrm{tr}\left(\check{\boldsymbol{K}}\bar{\boldsymbol{Q}}\check{\boldsymbol{K}}\bar{\boldsymbol{Q}}\right)}\left(\frac{1}{\hat{n}}\mathrm{tr}\left(\check{\boldsymbol{K}}_{\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}}\right) - \frac{1}{\hat{n}}\mathrm{tr}\left(\boldsymbol{I}_n + \frac{N}{n}\lambda\bar{\boldsymbol{Q}}\right)\check{\boldsymbol{K}}_{\boldsymbol{X}\hat{\boldsymbol{X}}}\check{\boldsymbol{K}}_{\hat{\boldsymbol{X}}\boldsymbol{X}}\bar{\boldsymbol{Q}}\right),
\end{aligned}$$

which results in the desired convergence

$$E_{train} - \bar{E}_{train} \to 0, \quad E_{test} - \bar{E}_{test} \to 0,$$

understood almost surely, where we used the notations

$$\boldsymbol{K}_{\boldsymbol{AB}} = E_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^T\boldsymbol{A})^T\sigma(\boldsymbol{w}^T\boldsymbol{B})],\ \check{\boldsymbol{K}}_{\boldsymbol{AB}} = \frac{N}{n}\frac{\boldsymbol{K}_{\boldsymbol{AB}}}{1+\delta},\ \boldsymbol{K} = \boldsymbol{K}_{\boldsymbol{XX}},\ \check{\boldsymbol{K}} = \check{\boldsymbol{K}}_{\boldsymbol{XX}}.$$

We represent the result[3] on a classification task of a 2-class MNIST[4] dataset (7 and 9) for two different choices of activation function σ in Figure 3.4. We can observe an almost perfect match between theory and simulation for relatively small $n, \hat{n}, p$ and N.

[3]The results are represented with a modified version of Z. Liaos code. Z. Liaos original code can be found on https://github.com/Zhenyu-LIAO/RMT4ELM

[4]http://yann.lecun.com/exdb/mnist/

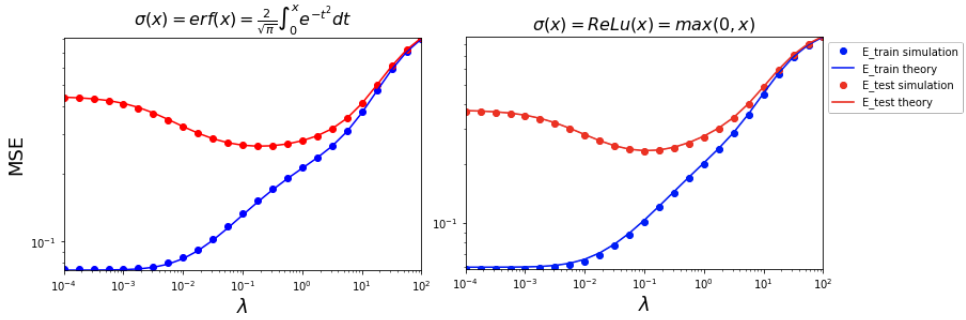

Figure 3.4: Comparison between predicted and simulated MSE (log-log scale) for sample size $n = \hat{n} = 1024$, variables $p = 784$ and hidden neurons $N = 512$.

3.2 Dimensionality Reduction - PCA

In modern technology it is very common that data sets are very large and problems may arise when we try to draw statistical conclusions [3, 22]. A common strategy to tackle this is to transform the data such that our transformed data contains fewer variables but most of the relevant information remains. A very popular method to accomplish this is the *Principal Component Analysis* (PCA), which may traced back to 1901 [21].

3.2.1 Minimizing Mean Squared Error of Projection

Consider a sampled data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \in \mathbb{R}^{p \times n}$ that we want to project onto a lower dimensional subspace spanned by the orthogonal unit vectors $[\boldsymbol{v}_1, \dots, \boldsymbol{v}_n] = \boldsymbol{V} \in \mathbb{R}^{d \times n}$, where $d < p$. It is safe to assume that the sample mean $\bar{x}_j$ for each column in $\boldsymbol{X}$ equals to zero (on a real data set, this can achieved by replacing our variables x_{ij} with $\hat{x}_{ij} = x_{ij} - \bar{x}_j$). This projection will result in loss of information and our goal is to minimize it. Due to Theorem 1.3, we know that the optimal projection (in terms of least squares) is given by the orthogonal projection. This gives us the mean squared error (MSE)

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{x}_i \right\|_2^2 &= \frac{1}{n}\sum_{i=1}^{n} \left(\|\boldsymbol{x}_i\|_2^2 - \left\|\boldsymbol{V}^T\boldsymbol{x}_i\right\|_2^2 \right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|_2^2 - \frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{V}^T\boldsymbol{x}_i\right\|_2^2 . \end{aligned} \tag{3.5}$$

As we are minimizing the MSE in terms of $\boldsymbol{V}$, we can simply ignore the first term and focusing on maximizing the second term. With help of the trace identities $\boldsymbol{a}^T\boldsymbol{a} = \operatorname{tr}(\boldsymbol{a}\boldsymbol{a}^T)$ and $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{C}\boldsymbol{A})$ we get

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{V}^T\boldsymbol{x}_i\right\|_2^2 &= \frac{1}{n}\sum_{i=1}^{n} \operatorname{tr}\left(\boldsymbol{V}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{V}\right) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{V}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{V}\right) \\ &= \operatorname{tr}\left(\boldsymbol{V}^T\boldsymbol{S}\boldsymbol{V}\right) = \operatorname{tr}\left(\boldsymbol{S}\boldsymbol{V}\boldsymbol{V}^T\right) \\ &= \sum_{i=1}^{n} \operatorname{tr}\left(\boldsymbol{S}\boldsymbol{v}_i\boldsymbol{v}_i^T\right) = \sum_{i=1}^{n} \left(\boldsymbol{S}\boldsymbol{v}_i\right)^T \boldsymbol{v}_i \\ &= \sum_{i=1}^{n} \boldsymbol{v}_i^T\boldsymbol{S}\boldsymbol{v}_i. \end{aligned} \tag{3.6}$$

Now let $y_i = \boldsymbol{v}^T \boldsymbol{x_i}$, we have that

$$\begin{aligned} \boldsymbol{v}^T \boldsymbol{S} \boldsymbol{v} &= \frac{1}{n-1} \left(\sum_{i=1}^{n} \boldsymbol{v}^T \boldsymbol{x_i} \boldsymbol{x_i}^T \boldsymbol{v} - n \boldsymbol{v}^T \bar{\boldsymbol{x}} \bar{\boldsymbol{x}}^T \boldsymbol{v} \right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{n} \left(\boldsymbol{v}^T \boldsymbol{x}_i \right)^2 - n \left(\boldsymbol{v}^T \bar{\boldsymbol{x}} \right)^2 \right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{n} \left(\boldsymbol{v}^T \boldsymbol{x}_i \right)^2 - \frac{1}{n} \left(\sum_{i=1}^{n} \boldsymbol{v}^T \boldsymbol{x}_i \right)^2 \right) \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 = s_y^2 \end{aligned} \tag{3.7}$$

which tells us that $\boldsymbol{v}^T \boldsymbol{S} \boldsymbol{v}$ is the sample variance of $y_1, \ldots, y_n$. This means that minimizing the MSE of the orthogonal projection is equivalent to maximizing the variance of our data. We shall exploit this in the next subsection to acquire a rather simple model for reducing the dimensionality while preserving as much information as possible.

3.2.2 Maximizing the Variance

We have that the key idea behind PCA is to maximize the variance of our data which means that we seek a linear combination $\boldsymbol{v}^T \boldsymbol{x} = \sum_{i=1}^{p} v_i x_i$ so that the variance is maximized. The sample variance for such a linear combination is given by $\boldsymbol{v}^T \boldsymbol{S} \boldsymbol{v}$ where $\boldsymbol{S}$ is the sample covariance matrix. To proceed with a non-trivial optimization problem, we shall assume that the vectors spanning the subspace $\boldsymbol{V}$ are of unit norm and orthogonal to each other, that is, $\|\boldsymbol{v}\|_2^2 = \boldsymbol{v}_i^T \boldsymbol{v}_j = 1$ for $i = j$ and $\boldsymbol{v}_i^T \boldsymbol{v}_j = 0$ otherwise. This results in the following optimization problem

$$\begin{aligned} \max_{\boldsymbol{v} \in \mathbb{R}^{p \times 1}} \quad & \boldsymbol{v}^T \boldsymbol{S} \boldsymbol{v} \\ \text{s.t.} \quad & \boldsymbol{v}^T \boldsymbol{v} = 1 \end{aligned} \tag{3.8}$$

Note that (3.8) is not a convex optimization problem. However, we can still solve it by looking at the spectral decomposition of $\boldsymbol{S}$

$$\boldsymbol{S} = \boldsymbol{E} \boldsymbol{D} \boldsymbol{E}^T = \begin{pmatrix} | & & | \\ \boldsymbol{e}_1 & \cdots & \boldsymbol{e}_p \\ | & & | \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} \begin{pmatrix} — & \boldsymbol{e}_1 & — \\ & \vdots & \\ — & \boldsymbol{e}_p & — \end{pmatrix},$$

where, without loss of generality, we can assume that the eigenvalues are ordered as $\lambda_1 \geqslant \ldots \geqslant \lambda_p$. Due to the symmetry of $\boldsymbol{S}$, we also know that the eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots$ will be orthogonal to each other (see Lemma 1.4). Multiplying $\boldsymbol{S}$ from both sides with $\boldsymbol{v}$ gives us

$$\boldsymbol{v}^T \boldsymbol{S} \boldsymbol{v} = \boldsymbol{v}^T \boldsymbol{E} \boldsymbol{D} \boldsymbol{E}^T \boldsymbol{v} = \left(\boldsymbol{E}^T \boldsymbol{v}\right)^T \boldsymbol{D} \boldsymbol{E}^T \boldsymbol{v} = \boldsymbol{z}^T \boldsymbol{D} \boldsymbol{z} = \sum_{i=1}^{p} \lambda_i z_i^2 \leqslant \lambda_1 \sum_{i=1}^{p} z_i^2, \quad (3.9)$$

where we used the notation $\boldsymbol{z} = \boldsymbol{E}^T \boldsymbol{v}$. Note that

$$\sum_{i=1}^{p} z_i^2 = \|\boldsymbol{z}\|_2^2 = \left(\boldsymbol{E}^T \boldsymbol{v}\right)^T \boldsymbol{E}^T \boldsymbol{v} = \boldsymbol{v}^T \boldsymbol{E} \boldsymbol{E}^T \boldsymbol{v} = \boldsymbol{v}^T \boldsymbol{v} = 1, \quad (3.10)$$

which tells us that $\boldsymbol{v}^T \boldsymbol{S} \boldsymbol{v} \leqslant \lambda_1$. Choosing $\boldsymbol{v}$ to be the eigenvector $\boldsymbol{e}_1$ corresponding to λ_1 results in

$$\boldsymbol{v}^T \boldsymbol{S} \boldsymbol{v} = \boldsymbol{e}_1^T \boldsymbol{E} \boldsymbol{D} \boldsymbol{E}^T \boldsymbol{e}_1 = \left(\boldsymbol{E}^T \boldsymbol{e}_1\right)^T \boldsymbol{D} \boldsymbol{E}^T \boldsymbol{e}_1 = \lambda_1, \quad (3.11)$$

since

$$\boldsymbol{E}^T \boldsymbol{e}_1 = \begin{pmatrix} — & \boldsymbol{e}_1 & — \\ & \vdots & \\ — & \boldsymbol{e}_p & — \end{pmatrix} \begin{pmatrix} | \\ \boldsymbol{e}_1 \\ | \end{pmatrix} = \begin{pmatrix} \boldsymbol{e}_1^T \boldsymbol{e}_1 \\ \boldsymbol{e}_2^T \boldsymbol{e}_1 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \end{pmatrix}. \quad (3.12)$$

This can be done for all the eigenvalues of $\boldsymbol{S}$, that is,

$$\lambda_1 = \boldsymbol{e}_1^T \boldsymbol{S} \boldsymbol{e}_1 = \text{sample variance for linear combinations } \boldsymbol{e}_1^T \boldsymbol{x}_i \; \forall i = 1, \ldots, n,$$

$$\vdots$$

$$\lambda_p = \boldsymbol{e}_p^T \boldsymbol{S} \boldsymbol{e}_p = \text{sample variance for linear combinations } \boldsymbol{e}_p^T \boldsymbol{x}_i \; \forall i = 1, \ldots, n,$$

which tells us that the largest eigenvalue and it's corresponding eigenvector matches the direction in which the data varies most. This means that reducing the dimensionality of a dataset while preserving maximum variance is obtained by projecting the data onto the eigenvector(s) corresponding to the largest eigenvalue(s).

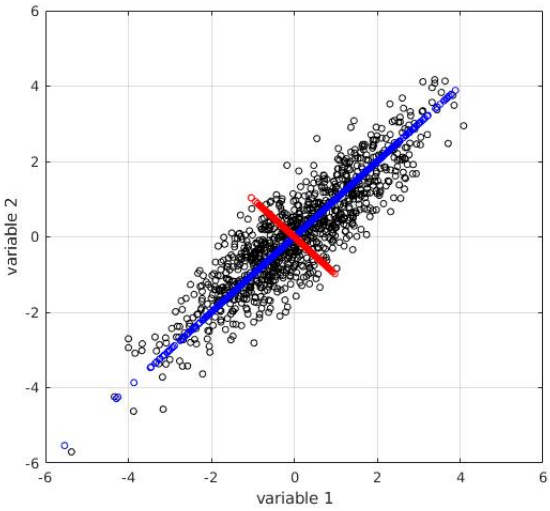

Figure 3.5: Blue/red dots represent the 2-dimensional data projected onto the first/second eigenvector.

We can measure how much of the variance is accounted for in a direction by calculating

$$\frac{\lambda_k}{\mathrm{tr}(\boldsymbol{S})}\,, \tag{3.13}$$

where λ_k is the eigenvalue corresponding to the eigenvector $\boldsymbol{e}_k$ we project our data onto. As an illustrating example, consider the 2-dimensional dataset in Figure 3.5. Projecting the data onto the first eigenvector (blue dots) will take about 94% of the the variance into account.

In general, one direction isn't enough to describe larger datasets sufficiently good. We can take more eigenvalues into consideration in our calculation and evaluate the accounted variances as

$$\Lambda_{\mathcal{S}} = \frac{\sum_{i\in\mathcal{S}} \lambda_i}{\mathrm{tr}(\boldsymbol{S})}\,, \tag{3.14}$$

where $\mathcal{S}$ is the set of eigenvalues we want to consider in our analysis. Nevertheless, correlation often do exist between sampled variables and then it's usually

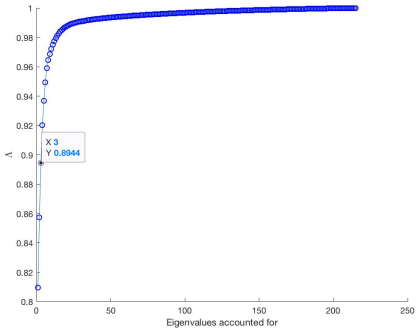

Figure 3.6: Cumulative sum of the eigenvalues corresponding to the sample covariance matrix of the data set 'ovariancancer'.

enough with a small set of eigenvalues to get a satisfying $\Lambda_{\mathcal{S}}$. To illustrate this, we perform PCA on the dataset 'ovariancancer'[5] which contains $n = 216$ samples[6] and $p = 4000$ variables that has been accounted for. Observing the result in Figure 3.6, we can see that we only need 3 eigenvalues (out of 216 unique ones) to explain about 90% of the variance. Another interesting outcome is that most of the eigenvalues tend to be very small and lump together while only representing a small portion of the variance. Plotting the data projected onto the first three eigenvectors gives us a noticeable clustering of the data (see Figure 3.7), which illustrates how PCA can be used to represent high dimensional data.

[5]Can be loaded in MATLAB with the command "load ovariancancer"
[6]121 with ovarian cancer and 95 without.

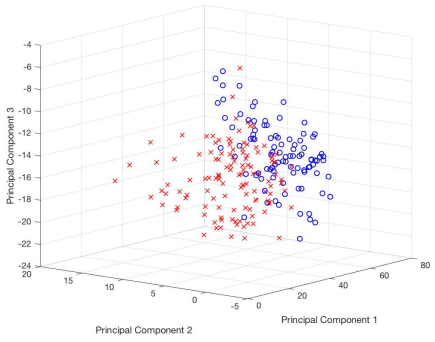

Figure 3.7: Red/Blue markers representing cancer/healthy patients in 3 dimension.

3.2.3 Statistical inference of PCA

So far, we treated PCA as a linear algebra problem without any assumptions about the underlying distribution of the data, and for many applications, this is good enough. For example, PCA may work as first step for other machine learning algorithms so intractable problems can be solved in a satisfactory time, or as we demonstrated in Figure 3.7, to visualize high-dimensional data. However, if we want to perform some kind of statistical inference, we need to consider some distribution for our data samples in $\boldsymbol{X}$.

Consider $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$ $i = 1, \dots n$ and let λ_j $j = 1, \dots, p$ be unique distinct eigenvalues for $\boldsymbol{C}$ and let $\hat{\lambda}_j$ be the eigenvalues for a sample covariance matrix built upon $\boldsymbol{x}_i$. D. N. Lawley [17] proves that the mean and variance for the random variable $\hat{\lambda}_j$ is given by

$$\begin{aligned} E[\hat{\lambda}_j] &= \lambda_j + \frac{\lambda_j}{n} \sum_{i=1, i\neq j}^{p} \left(\frac{\lambda_i}{\lambda_j - \lambda_i} \right) + \mathcal{O}\left(\frac{1}{n^2} \right), \\ Var[\hat{\lambda}_j] &= \frac{2\lambda_j^2}{n} \left(1 - \frac{1}{n} \sum_{i=1, i\neq j}^{p} \left(\frac{\lambda_i}{\lambda_j - \lambda_i} \right)^2 \right) + \mathcal{O}\left(\frac{1}{n^3} \right). \end{aligned} \tag{3.15}$$

T. W. Anderson [2] shows that the eigenvalues $\hat{\lambda}_j$ in fact are asymptotically

normal distributed. This means that as n grows large, we have

$$\hat{\lambda}_j \sim \mathcal{N}\left(\lambda_j, \frac{2\lambda_j^2}{n}\right) \tag{3.16}$$

approximately. Normalizing (3.16) results in

$$\frac{\hat{\lambda}_j - \lambda_j}{\lambda_j\sqrt{2/n}} \sim \mathcal{N}(0,1), \tag{3.17}$$

which can be used for classic statistical inference such as hypothesis testing and interval estimation. For example, a confidence interval (CI) for λ_j with a confidence level $1-\alpha$ may be constructed as

$$\begin{gathered} -z_{\alpha/2} \leqslant \frac{\hat{\lambda}_j - \lambda_j}{\lambda_j\sqrt{2/n}} \leqslant z_{\alpha/2} \\ \Leftrightarrow \\ -z_{\alpha/2}\lambda_j\sqrt{2/n} + \lambda_j \leqslant \hat{\lambda}_j \leqslant z_{\alpha/2}\sqrt{2/n} + \lambda_j \\ \Leftrightarrow \\ \frac{\hat{\lambda}_j}{1 + z_{\alpha/2}\sqrt{2/n}} \leqslant \lambda_j \leqslant \frac{\hat{\lambda}_j}{1 - z_{\alpha/2}\sqrt{2/n}} \end{gathered} \tag{3.18}$$

and thus, we have a CI

$$I_{\lambda_j} = \left(\frac{\hat{\lambda}_j}{1 + z_{\alpha/2}\sqrt{2/n}}, \frac{\hat{\lambda}_j}{1 - z_{\alpha/2}\sqrt{2/n}}\right), \tag{3.19}$$

where $z_{\alpha/2}$ is the z-score which can be found in a normal table. Recall that we are only interested in I_{λ_j} for large n, so it is safe to assume that $z_{\alpha/2}\sqrt{2/n} < 1$.

For hypothesis testing, we may want to test

$$H_0 : \lambda_j = \lambda_{j0} \quad \text{vs} \quad H_1 : \lambda_j \neq \lambda_{j0} \tag{3.20}$$

where we reject H_0 on a significance level α if

$$\left|\frac{\hat{\lambda}_j - \lambda_{j0}}{\lambda_j\sqrt{2/n}}\right| \geqslant z_{\alpha/2}. \tag{3.21}$$

A perhaps more interesting test would be to check if the smallest $(p-k)$ eigenvalues are all the same, that is

$$\begin{aligned} H_0 &: \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p \\ &\text{vs} \\ H_1 &: \text{At least two of the smallest } (p-k) \text{ eigenvalues are not equal.} \end{aligned} \tag{3.22}$$

Accepting H_0 as in (3.22) justifies that the k largest PCs are measuring some significant variation in our input data while the remaining $(p-k)$ PCs effectively measure noise. To test (3.22), one may use

$$Q = \left(\frac{\prod_{j=k+1}^{p} \hat{\lambda}_j}{\left(\sum_{j=k+1}^{p} \frac{\hat{\lambda}_j}{p-k} \right)^{p-k}} \right)^{n/2} \tag{3.23}$$

as a test statistic, where $-2\ln(Q) \sim \chi^2(f)$ approximately under H_0 and $f = \frac{1}{2}(p-k+2)(p-k-1)$ [15]. This lets us reject the null hypothesis at a significance level α if

$$-2\ln(Q) \geqslant \chi^2_{\alpha/2}(f), \tag{3.24}$$

where $\chi^2_{\alpha/2}(f)$ can be found in a χ^2 table.

Chapter 4

Ending discussion

We have introduced, discussed and applied the results from random matrix theory. The focus has solely been on matrices with entries being normal distributed, or even stricter, $\mathcal{N}(0,1)$ distributed. Nevertheless, many interesting results, such as *Wigner's semicircle law* and the *Marčenko–Pastur law*, still emerges.

As we mentioned in the beginning, RMT mostly appears on research level. The aim of this thesis has been to introduce and perhaps simplify some results of this rather advanced topic, so that someone with basic probability and linear algebra skills can get a solid grasp of the properties related to random matrices.

For further research, it would be interesting to generalize some of the results. We have mostly focused on the spectrum of large matrices with an underlying $\mathcal{N}(0,1)$ distribution. Having explicit non-asymptotically eigenvalue distributions for matrices with other distributions than standard normal would perhaps shed some light on the topics of RMT, such that it would be introduced much earlier in a mathematical oriented education.

Finally, as for applying the theory, we have only introduced RMT in already well established applications. Starting from the other side, that is, given an arbitrary matrix ensemble and trying to answer 'what does this model?' would perhaps result in new applications beside those we already have today.

Bibliography

[1] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2009.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, 3rd edition, 2003.

[3] S. Ayesha, M. K. Hanif, and R. Talib. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, 59:44–58, 2020.

[4] A. Cyr, F. Thériault, and S. Chartier. Revisiting the xor problem: a neurorobotic implementation. *Neural Computing and Applications*, 32(14):9965–9973, 2020.

[5] A. Edelman. *Eigenvalues and Condition Numbers of Random Matrices*. PhD thesis, Massachusetts Institute of Technology, 1989.

[6] A. Edelman and N. Raj Rao. Random matrix theory. *Acta Numerica*, 14:233–297, 2005.

[7] G. Ergün. *Random Matrix Theory*, pages 2549–2563. Springer New York, New York, NY, 2012.

[8] P. J. Forrester. *Log-gases and random matrices*. Princeton University Press, Princeton, 2010.

[9] P. J. Forrester and N. S. Witte. Exact Wigner surmise type evaluation of the spacing distribution in the bulk of the scaled random matrix ensembles. *arXiv e-prints*, pages math–ph/0009023, September 2000.

[10] M. Hoogendoorn and B. Funk. *Machine Learning for the Quantified Self: On the Art of Learning from Sensory Data*. Cognitive Systems Monographs. Springer International Publishing, 2017.

[11] G. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):513–529, 2012.

[12] G. Huang, Q-Y. Zhu, and C-K. Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1):489 – 501, 2006. Neural Networks.

[13] A. T. James and R. A. Fisher. The non-central wishart distribution. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 229(1178):364–366, 1955.

[14] A. H. Jiang, D. L. K. Wong, G. Zhou, D. G. Andersen, J. Dean, G. R. Ganger, G. Joshi, M. Kaminksy, M. Kozuch, Z. C. Lipton, and P. Pillai. Accelerating deep learning by focusing on the biggest losers. *arXiv e-prints*, page arXiv:1910.00762, October 2019.

[15] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.

[16] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, April 2016.

[17] D. N. Lawley. Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43:128, June 1956.

[18] G. Livan, M. Novaes, and P. Vivo. Introduction to random matrices. *SpringerBriefs in Mathematical Physics*, 2018.

[19] C. Louart, Z. Liao, and R. Couillet. A Random Matrix Approach to Neural Networks. *arXiv e-prints*, page arXiv:1702.05419, February 2017.

[20] D. Paul. and A. Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

[21] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[22] S. Velliangiri, S. Alagumuthukrishnan, and S Iwin Thankumar joseph. A review of dimensionality reduction techniques for efficient computation. *Procedia Computer Science*, 165:104–111, 2019.

[23] E. P. Wigner. Random matrices in physics. *SIAM Review*, 9(1):1–23, 1967.

[24] L. Zhenyu. *A random matrix framework for large dimensional machine learning and neural networks.* Theses, Université Paris-Saclay, September 2019.