UNIVERSITY OF CALIFORNIA,
IRVINE

Relaxing Independence in the Marchenko-Pastur Law for Random Matrices and the Application to Approximate Embeddings

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

Jennifer Bryson

Dissertation Committee:
Professor Roman Vershynin, Chair
Chancellor's Professor Hongkai Zhao, Chair
Professor Michael Cranston

2019

TABLE OF CONTENTS

Page

LIST OF FIGURES **iv**

ACKNOWLEDGMENTS **v**

CURRICULUM VITAE **vi**

ABSTRACT OF THE DISSERTATION **viii**

1 Introduction **1**
- 1.1 Dissertation organization . . . 1
- 1.2 Notation convention . . . 1
- 1.3 Background . . . 2
  - 1.3.1 Random variables and random matrices . . . 2
  - 1.3.2 Sample covariance matrix . . . 6
  - 1.3.3 The Marchenko-Pastur law . . . 9
  - 1.3.4 Related works . . . 22
  - 1.3.5 Vector and matrix norms . . . 24
  - 1.3.6 Asymptotic notation . . . 26
  - 1.3.7 Binomial coefficients . . . 27

2 Weakening Independence in the Marchenko-Pastur Law **30**
- 2.1 Block version and vectorized tensor version for the Marchenko-Pastur law . 30
  - 2.1.1 Idea of the Proof . . . 33
- 2.2 Proof of Theorem 2.2 . . . 35
  - 2.2.1 Diagonal contribution . . . 35
  - 2.2.2 Off-diagonal contribution: setting up partitions . . . 36
  - 2.2.3 Partition $(2, 2)$ . . . 37
  - 2.2.4 Partition $(1, 1, 1, 1)$ . . . 38
  - 2.2.5 Partition $(2, 1, 1)$ . . . 39
  - 2.2.6 Exchangeable distributions . . . 40
- 2.3 Proof of Theorem 2.3 . . . 43
  - 2.3.1 Diagonal contribution . . . 44
  - 2.3.2 Off-diagonal contribution . . . 46
- 2.4 Numerical experiments . . . 55

3 Approximate Embeddings and the Marchenko-Pastur Law **62**
3.1 General lower bound on the least dimension required for ε-embedding vectors 63
3.2 Asymptotic formulas for the least dimension required for ε-embedding i.i.d. random vectors 65
3.3 ε-embedding random vectors with a known covariance 68
3.4 Numerical experiments 70

Bibliography **74**

A Matlab Code for Numerical Simulations **79**
A.1 Function for plotting the the Marchenko-Pastur density function 79
A.2 Plotting the empirical spectral density with the Marchenko-Pastur density 81
A.2.1 i.i.d. entries 81
A.2.2 Block uncorrelated entries 81
A.2.3 Vectorized tensor entries 83
A.3 Plotting the theoretically calculated N^ϵ vs. the actual N^ϵ 84

LIST OF FIGURES

Page

1.1 Example of the Marchenko-Pastur law. 11
1.2 Visualizing the Marchenko-Pastur density for various ratios, $\lambda = \frac{\#rows}{\#columns}$. . . 12
1.3 Visualizing the Marchenko-Pastur density for various entry variances, σ^2. . . 12

2.1 Pictorial view of the variables used for counting the off-diagonal terms 51
2.2 Examples of block version of the Marchenko-Pastur law 58
2.3 Examples of vectorized 2-tensors version of the Marchenko-Pastur law 59
2.4 Examples of the vectorized 3-tensors version of the Marchenko-Pastur law for $n = 45$ 60
2.5 Vectorized 3-tensor tending to the Marchenko-Pastur density as n increases . 61

3.1 Comparing the asymptotic formulas for N^ε and R^ε to their numerically computed values for random vectors with i.i.d. entries 71
3.2 N^ε for random vectors with a given covariance 72
3.3 Comparing N^ε for random vectors with specific covariance and N^ε for i.i.d. random vectors 73

ACKNOWLEDGMENTS

I would like to thank my co-advisors, Roman Vershynin and Hongkai Zhao, for all their excellent help, ideas, and editing. I would also like to thank my parents for supporting me and encouraging me to always do my best. Last, but not least, John Hofrichter for always being there for me and working weekends in coffee shops together.

The research presented in this dissertation was partially supported by NSF Graduate Research Fellowship Program DGE-1321846 and ARCS Foundation.

Chapter 3 was published with the permission of the Society for Industrial and Applied Mathematics. The text of this chapter is mostly a reprint of the material as it appears in Section 4 of Intrinsic Complexity and Scaling Laws: From Random Fields to Random Vectors in the Multiscale Modeling and Simulation journal.

CURRICULUM VITAE

Jennifer Bryson

EDUCATION

Doctor of Philosophy in Mathematics **2019**
University of California, Irvine *Irvine, CA*

Masters in Mathematics **2016**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Mathematics **2013**
Texas A&M University *College Station, TX*

RESEARCH EXPERIENCE

Graduate Research Assistant **2016–2019**
University of California, Irvine *Irvine, California*

Research Experience for Undergraduates **2012**
Emory University *Atlanta, GA*

TEACHING EXPERIENCE

Teaching Assistant **2014–2019**
University of California, Irvine *Irvine, CA*
Courses: Math 2A, 2B, 2D, 2E, 3A, 9, 105A, 105B, 120A, 130B

Teaching Assistant for Undergraduate Summer School **2016**
Park City Math Institute *Park City, UT*
Course: Data Analysis with Wavelets and Other Mathematical Tools

Instructor **2014**
Duke Talent Identification Program (TIP) *Davidson, NC*
Course: Cryptography and the Mathematics of Spying

REFEREED JOURNAL PUBLICATIONS

Intrinsic Complexity and Scaling Laws: From Random Fields to Random Vectors **2019**
Multiscale Modeling and Simulation

Unimodal Sequences and Quantum and Mock Modular Forms **2012**
Proceedings of the National Academy of Sciences

UNDER REVIEW AND WORKING PAPERS

Relaxing Independence in the Marchenko-Pastur Law for Random Matrices

HONORS AND AWARDS

Connelly Award for top graduate student(s) in both research and teaching **2019**
University of California, Irvine, Department of Mathematics

Achievement Rewards for College Scientists (ARCS) Scholar **2017-2019**
University of California, Irvine

National Science Foundation Graduate Research Fellowship **2013-2018**
National Science Foundation

Outstanding Contributions to the Department Award **2016** & **2017**
University of California, Irvine, Department of Mathematics

Outstanding Mathematics Teaching Assistant Honorable Mention **2015**
University of California, Irvine, Department of Mathematics

ABSTRACT OF THE DISSERTATION

Relaxing Independence in the Marchenko-Pastur Law for Random Matrices and the Application to Approximate Embeddings

By

Jennifer Bryson

Doctor of Philosophy in Mathematics

University of California, Irvine, 2019

Professor Roman Vershynin, Chair
Chancellor's Professor Hongkai Zhao, Chair

This dissertation adds to the collection of works studying the Marchenko-Pastur law in two ways. First it considers two new models of random column vectors that have weaker independence hypotheses than the well-known i.i.d. hypothesis and shows that random matrices, formed by concatenating random column vectors of a model, still follow the Marchenko-Pastur law. The two models of random column vectors are block columns and vectorized tensor columns. The block column vectors will be made up of n blocks each of length d, with $d = o(n)$. If the entries are mean zero, variance one, have uniformly bounded fourth moments, entries within a block are uncorrelated, and entries in different blocks are independent, then the Marchenko-Pastur theorem holds as $n \to \infty$. Furthermore, if additionally an exchangeability criteria is satisfied, then the theorem holds without requiring $d = o(n)$. The vectorized tensor columns will be made up of a vectorized t-tensor of an i.i.d. vector of length n which has entries that are mean zero, variance one, uniformly bounded fourth moments, and $t^3 = o(n)$, and again the Marchenko-Pastur theorem holds as $n \to \infty$.

The second contribution of this dissertation is in studying a particular type of an approximate embedding of vectors. For any collection of vectors, a general lower bound for the least dimension required for an approximate embedding is given. For vectors which are column

vectors of a matrix that follows the Marchenko-Pastur law, an asymptotic formula for the exact value of the least dimension required is given. Numerical results show this asymptotic formula holds quite well, even for relatively small dimensions. Because this works so well for small dimensions, this gives an easy numerical test that provides evidence for answering the question, "Does a specific covariance structure have a limiting spectral distribution or not?" We consider a particular covariance structure which relates to the number of terms needed in the Karhunen-Loéve expansion to approximate a random field within a specified tolerance.

Chapter 1

Introduction

1.1 Dissertation organization

This dissertation has three chapters. This first chapter provides some background material which will aid in the understanding of this work. The second chapter states and proves an extension of the Marchenko-Pastur law toward matrices with weaker independence criteria. Specifically we give a block version of the Marchenko-Pastur law, and a vectorized tensor version as well. In the third chapter we give a specific notion of an approximate embedding, and we show how the Marchenko-Pastur law can determine the smallest possible dimension needed for such an embedding. Lastly, Appendix A gives some Matlab code for the key components needed for producing the figures included in the dissertation.

1.2 Notation convention

A main focus of this work is studying the limiting spectral distribution of the sample covariance matrix (XX^T) for certain random matrices, X, as their dimensions grow to infinity

with a fixed ratio of number of rows to number of columns. Throughout this dissertation we will try to stick to the notation of $X \in \mathbb{R}^{p\times m}$. We will consider a sequence of matrices which will be indexed by p, meaning $\{X\}_{p=1}^{\infty}$ with $\frac{p}{m} \to \lambda \in (0, +\infty)$. Therefore we think of $m = m(p)$ as a function of p. In Chapter 2, p is going to be a function of another variable n which will tend towards infinity. In Chapter 3 it is convenient to consider matrices of the form $V^T V$, so we can think of $X = V^T \in \mathbb{R}^{p\times m}$ in this chapter.

1.3 Background

1.3.1 Random variables and random matrices

This section gives a brief review of the main definitions needed from probability theory.

Definition 1.1. *A* **probability space** *or* **probability triple**, $(\Omega, \mathcal{F}, P)$, *consists of*

1. *a sample space,* Ω*, which is a set of possible outcomes*
2. *a* σ*-algebra,* $\mathcal{F}$*, which is a set of events, where each event is a set containing zero or more outcomes, and*
3. *a probability measure,* P*, which is a measure* $P : \mathcal{F} \to [0, 1]$ *that assigns probabilities to events with* $P(\Omega) = 1$*.*

Definition 1.2. *A* **σ-algebra**, $\mathcal{F}$*, on the set* Ω *is a nonempty collection of subsets of* Ω *that satisfy (i) if* $A \in \mathcal{F}$*, then* $A^c \in \mathcal{F}$*, and (ii) if* $A_i \in \mathcal{F}$ *is a countable sequence of sets, then* $\bigcup_i A_i \in \mathcal{F}$*.*

Definition 1.3. *A* **measure**, μ*, on a measurable space* $(\Omega, \mathcal{F})$*, is a function* $\mu : \mathcal{F} \to \mathbb{R}$ *such that (i)* $\mu(A) \geq \mu(\varnothing) = 0$*, i.e.* μ *is nonnegative, and (ii) if* $A_i \in \mathcal{F}$ *is a countable sequence of disjoint sets, then* $\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)$*, i.e.* μ *is countably additive.*

Example 1.1. *Consider a single roll of a fair 6-sided die. The possible outcomes are* $\Omega = \{1,2,3,4,5,6\}$. *One possible* σ*-algebra is to consider all subsets of* Ω *as events, so* $\mathcal{F}$ *is the power set of* Ω. *Thus* $\mathcal{F}$ *contains* $\varnothing, \Omega, \{1,3,5\} =$ *the event of rolling an odd number,* $\{2,4,6\} =$ *the event of rolling an even number,* $\{1\} =$ *the event of rolling a 1, etc. The probability measure of an event* A *is* $P(A) = \frac{|A|}{6}$. *For example,* $P(\{1,3,5\}) = \frac{1}{2}$, *meaning the probability of rolling an odd number is* $\frac{1}{2}$ *as expected.*

Definition 1.4. *A function* $X : \Omega \to \mathbb{R}$ *is a* **random variable** *if for every Borel set* $B \subset \mathbb{R}$ *we have* $X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{F}$.

Having a probability triple $(\Omega, \mathcal{F}, P)$ and a random variable $X : \Omega \to \mathbb{R}$, there is an induced probability measure μ on $\mathbb{R}$, called its **distribution**, given by $\mu(A) = P(X^{-1}(A)) = P(X \in A)$ for Borel sets A. The distribution is most commonly described by its distribution function, defined below.

Definition 1.5. *Given a probability triple* $(\Omega, \mathcal{F}, P)$ *and a random variable* $X : \Omega \to \mathbb{R}$, *the* **distribution function**, $F : \mathbb{R} \to \mathbb{R}$, *is given by* $F(x) = P(X \leq x)$.

Definition 1.6. *When the distribution function* $F(x) = P(X \leq x)$ *has the form*

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

we say that X *has* **density function** *or* **probability density function (pdf)** f.

Definition 1.7. *The* **expected value** *or* **mean** *of a nonnegative random variable* X *on* $(\Omega, \mathcal{F}, P)$ *is defined to be*

$$\mathbb{E}[X] = \int X dP$$

. *For a general random variable, let* $x^+ = \max\{x, 0\}$ *be the positive part and* $x^- = \max\{-x, 0\}$ *be the negative part. We say that the expected value of* X, $\mathbb{E}[X]$, *exists and is equal to* $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$ *whenever the subtraction makes sense.*

For a random variable X with density function $f(x)$, $\mathbb{E}[X] = \int_{\mathbb{R}} xf(x)dx$.

Definition 1.8. *The* **variance** *of a random variable X is defined as*

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

For a random variable X with mean zero and density function $f(x)$, $Var(X) = \int_{\mathbb{R}} x^2 f(x)dx$.

We will frequently use three types of random variables in our numerical examples. We define those random variables now by defining their distributions in the next three examples.

Example 1.2. *Let X be a discrete random variable that takes the value 1 with probability p and the value 0 with probability $1 - p$. X is called a* **Bernoulli random variable**. *The Bernoulli distribution is given by*

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

In our numerical examples, we will use the shifted and scaled Bernoulli distribution where X takes the value 1 with probability $\frac{1}{2}$ and the value -1 with probability $\frac{1}{2}$, because now the random variable has mean 0 and variance 1.

Example 1.3. *Let X be the random variable that takes the values between 0 and 1 with equal probabilities. X is called a* **Uniform random variable on (0,1)**. *The Uniform(0,1)*

distribution is given by

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

And the density function is $f(x) = 1$ *for* $x \in (0,1)$ *and 0 otherwise. We write* $X \sim Uniform(0,1)$.

In our numerical examples, we will use the shifted and scaled Uniform distribution where X takes the values between $-\sqrt{3}$ and $\sqrt{3}$, i.e. $X \sim Uniform(-\sqrt{3}, \sqrt{3})$, because now the random variable has mean 0 and variance 1.

Example 1.4. *Let* X *be the random variable that has a density function given by*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for fixed $\mu, \sigma^2 \in \mathbb{R}$. X *is said to be a* **normal random variable** *or* **Gaussian random variable**. *We write* $X \sim N(\mu, \sigma^2)$. *When* $\mu = 0$ *and* $\sigma^2 = 1$, *we say* X *is a* **standard normal** *or* **standard Gaussian** *random variable.*

Definition 1.9. Random matrices *are matrices whose entries are random variables.*

Definition 1.10. *Two random variables* X *and* Y *are* **independent** *if for all* A, B *in the Borel set on* $\mathbb{R}$, $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$. *Similarly, a collection of random variables* $X_1, ..., X_n$ *are* **independent** *if for all* $\{A_i\}_{i=1}^n$ *in the Borel set on* $\mathbb{R}$,

$$P\left(\bigcap_{i=1}^{n} \{X_i \in A_i\}\right) = \prod_{i=1}^{n} P(X_i \in B_i).$$

Definition 1.11. *When a collection of random variables* $X_1, X_2, ...$ *are independent and each*

have the same distribution, we say they are **independent and identically distributed** *or* **i.i.d.** *for short.*

Definition 1.12. *A collection of random variables X_i, $i \in I$ with $\mathbb{E}[X_i^2] < \infty$ for all i is* **uncorrelated** *if*

$$\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j] \text{ for all } i \neq j.$$

Independence implies uncorrelated, meaning if X and Y are independent random variables, then X and Y are uncorrelated. Our work in Chapter 2 generalizes a result that required i.i.d. random variables by weakening the hypothesis to allow some of the random variables to just be uncorrelated, but not independent.

For an example of random variables that are uncorrelated but not independent, consider $X \sim N(0,1)$ and $Y = X^2$. X and Y are certainly not independent because $P(X \in (1,2), Y \in (9,10)) = 0 \neq P(X \in (1,2))P(Y \in (9,10))$. However, X and Y are uncorrelated because $\mathbb{E}[XY] = \mathbb{E}[X^3] = 0$, since the density function for X is an even function.

1.3.2 Sample covariance matrix

The correlation coefficient between two random variables is a number in $[-1,1]$ which roughly tells you how linearly related the two random variables are. For example if we have random variables X and Y, where $Y = \alpha X + \beta$, then the correlation coefficient will be 1 if $\alpha > 0$ and -1 if $\alpha < 0$. When two random variables are uncorrelated, then the correlation coefficient is 0.

Definition 1.13. *The* **correlation coefficient**, $\rho_{X,Y}$, *between two random variables X and Y with expected values μ_X and μ_Y and variances σ_X^2 and σ_Y^2 is defined as*

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}.$$

The numerator of the correlation coefficient is the covariance between two random variables X and Y. The covariance measures the joint variability of the two random variables, but the magnitude of the covariance is less intuitive without the the normalizing factor in the denominator.

Definition 1.14. *The* **covariance**, $cov(X, Y)$, *between two random variables* X *and* Y *with expected values* μ_X *and* μ_Y *is defined as*

$$cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Note that the covariance between X and itself is the variance of X, i.e. $cov(X, X) = Var(X)$.

For a random vector $\mathbf{x} = [x_1, x_2, ..., x_p]^T$ with expected value $\mu_{\mathbf{x}} = [\mu_{x_1}, \mu_{x_2}, ..., \mu_{x_p}]^T$, the covariance matrix, $cov[\mathbf{x}, \mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})^T]$, is a symmetric matrix where the $cov[\mathbf{x}, \mathbf{x}]_{i,j} = cov(x_i, x_j)$. In particular, the diagonal entries are the variances, $cov[\mathbf{x}, \mathbf{x}]_{i,i} = var(x_i)$.

The covariance matrix is a generalization of variance to multiple dimensions and is important to understanding data that have more than one predictor or feature. Before we get into that we need the notion of the sample covariance matrix because in application we generally do not know the joint probability distribution.

Definition 1.15. *Suppose that we have* m *draws of a random vector* $\mathbf{x} = [x_1, x_2, ..., x_p]^T$ *from some unknown joint probability distribution. Arrange the observation vectors as the columns of a matrix* $X \in \mathbb{R}^{p \times m}$*, where* $X_{i,j} = x_i$ *from the* j^{th} *draw. Let* $\mu_{\mathbf{x}} = [\mu_{x_1}, \mu_{x_2}, ..., \mu_{x_p}]^T$ *be the sample mean vector, i.e.* $\mu_{x_i} = \frac{1}{m} \sum_{j=1}^{m} X_{i,j}$*. The* **sample covariance matrix** *defined as*

$$S = \frac{1}{m-1} \sum_{k=1}^{m} (\mathbf{x}_{\cdot,\mathbf{k}} - \mu_{\mathbf{x}})(\mathbf{x}_{\cdot,\mathbf{k}} - \mu_{\mathbf{x}})^* = \frac{1}{m-1} (X - \mu_{\mathbf{X}} \mathbf{1}_m^T)(X - \mu_{\mathbf{X}} \mathbf{1}_m^T)^*.$$

This formula makes sense since the i, j^{th} entry is $\frac{m}{m-1}$ times the sample average value of $(x_i - \mu_{x_i})(x_j - \mu_{x_j})$. The $\frac{m}{m-1}$ factor is called Bessel's correction, and it is needed since the true mean was not used but instead the sample mean was used so the sample covariance is biased by the factor $\frac{m-1}{m}$, meaning $\mathbb{E}[\text{the sample average value of } (x_i - \mu_{x_i})(x_j - \mu_{x_j})] = \frac{m-1}{m} cov(x_i, x_j)$. The intuitive reasoning for having $\frac{1}{m-1}$ instead of $\frac{1}{m}$ out front is because we used the m vectors to estimate the mean vector, so we effectively only have $m-1$ degrees of freedom left to estimate the covariance.

Now because $\mu_{\mathbf{x}} {\mu_{\mathbf{x}}}^*$ is a rank 1 matrix, Theorem A.44 of [7] shows the removal of $\mu_{\mathbf{x}}$ does not effect the limiting spectral distribution. Thus for the sake of finding the limiting spectral distribution we can consider the sample covariance matrix to be

$$S = \frac{1}{m} \sum_{k=1}^{m} x_{\cdot,k} x_{\cdot,k}^* = \frac{1}{m} X X^*.$$

To understand the importance of the sample covariance matrix, let's consider the example of trying to understand the dollar value of a car based on its milage and age of the car in days. Let's say we have 100 cars where we know their milage, age, and dollar value. We can visualize the data as plotting 100 points in $\mathbb{R}^3$. The eigenvector of the sample covariance matrix corresponding to the largest eigenvalue will give you the best linear representation of the data points, where by "best" here we mean it minimizes the sum of squared errors. This will give you the best 1-dimensional representation of the data. Similarly, the eigenvector of the sample covariance matrix corresponding to the second largest eigenvalue will give you the next most valuable direction that is perpendicular to the first eigenvector. Thus the span of these two eigenvectors will give you the best 2-dimensional representation of the data. The proof of this fact comes from understanding singular values and the singular value decomposition. A nice constructive proof is given in Chapter 3 of [10]. This concept is called principal component analysis or PCA for short. Furthermore, the square roots

of the eigenvalues are the singular values, i.e. $\sqrt{\lambda_i} = \sigma_i$. Thinking of the columns of the matrix, X, as data points, the largest eigenvalue is the sum of the squared lengths of the projections of the points onto the line determined by the corresponding eigenvector. Thus the larger the eigenvalue is, the more information is captured in by its eigenvalue. Furthermore letting X_k be the projections of the columns of X onto the k top singular vectors (i.e. the eigenvectors of XX^T corresponding to the k largest eigenvalues $\lambda_1 \geq \lambda_2 \geq ...$), we have that $\|X - X_k\|_2^2 = \sigma_{k+1}^2 = \lambda_{k+1}$. This is one of the reasons we care about the eigenvalues of the sample covariance matrix. Incredibly, the Marchenko-Pastur law completely determines the limiting distribution of eigenvalues of the sample covariance matrix for certain random matrices. We introduce the Marchenko-Pastur law in the next section.

1.3.3 The Marchenko-Pastur law

The Marchenko-Pastur law or Marchenko-Pastur distribution was discovered by Ukrainian mathematicians Vladimir Marchenko and Leonid Pastur in 1967 [28]. In that cornerstone paper, they determine the limiting spectral distribution (i.e. limiting distribution of eigenvalues) of the scaled sample covariance matrix, $\frac{1}{m}XTX^*$, where X is a $p \times m$ random matrix with independent entries and T is a $m \times m$ diagonal matrix with a limiting spectral distribution. Before giving the general statement of the Marchenko-Pastur law, we first give a simplified version when $T = I$.

Definition 1.16. *Let A_p be a Hermitian $p \times p$ matrix with eigenvalues $\lambda_j, j = 1, ..., p$. We define the* **empirical spectral distribution** *(ESD) of the matrix A_p as*

$$F^{A_p}(x) = \frac{1}{p}\#\{j \leq p : \lambda_j \leq x\}.$$

Definition 1.17. *For a given sequence of random matrices, $\{A_p\}$, $\lim_{p\to\infty} F^{A_p}$ sometimes converges to a (possibly defective) limit distribution F called the* **limiting spectral distri-**

bution *(LSD) of* $\{A_p\}$.

Definition 1.18. *A* **defective** *distribution function is one where the total mass is less than 1 due to some eigenvalues tending to* $\pm\infty$.

Theorem 1.1 (Marchenko-Pastur [28])**.** *Let* X *be an* $p \times m$ *random matrix whose entries are i.i.d. random variables with mean 0 and variance* $\sigma^2 < \infty$*. Let* $S_p = \frac{1}{m} XX^T$ *and let* $\lambda_1 \geq ... \geq \lambda_p$ *be the eigenvalues of* S_p*. Finally, consider the random measure* $\mu_p(B) = \frac{1}{p}\#\{\lambda_j \in B\}$ *for* $B \subset \mathbb{R}$*, which has distribution* $F^{S_p}(x) = \frac{1}{p}\#\{\lambda_j \leq x\}$*. Assume that* $p, m \to \infty$ *so that the ratio* $p/m \to \lambda \in (0, \infty)$*. Then almost surely* $F^{S_p} \to F$ *vaguely as* $p \to \infty$*, where* F *has density function*

$$f_\lambda(x) = \begin{cases} \frac{1}{2\pi\sigma^2\lambda x}\sqrt{(\lambda_+ - x)(x - \lambda_-)}, & \text{if } \lambda_- \leq x \leq \lambda_+ \\ 0, & \text{otherwise} \end{cases}$$

and has a point mass 1-$\frac{1}{\lambda}$ *at the origin if* $\lambda > 1$*, where* $\lambda_\pm = \sigma^2(1 \pm \sqrt{\lambda})^2$.

For example, if we fill in a matrix $X \in \mathbb{R}^{1500 \times 3000}$ with numbers drawn independently from a standard Gaussian and create a histogram of the eigenvalues of $\frac{1}{3000} XX^T$, this this histogram will look very similar to the density function, ν, from the Marchenko-Pastur law where $\lambda = \frac{1}{2}$ and $\sigma^2 = 1$. This is shown in Figure 1.1. Letting the size of the matrix tend to infinity while keeping the ratio of number of rows to columns, which in this case is $\frac{1}{2}$, the histogram will match up perfectly with the red curve which is the Marchenko-Pastur density.

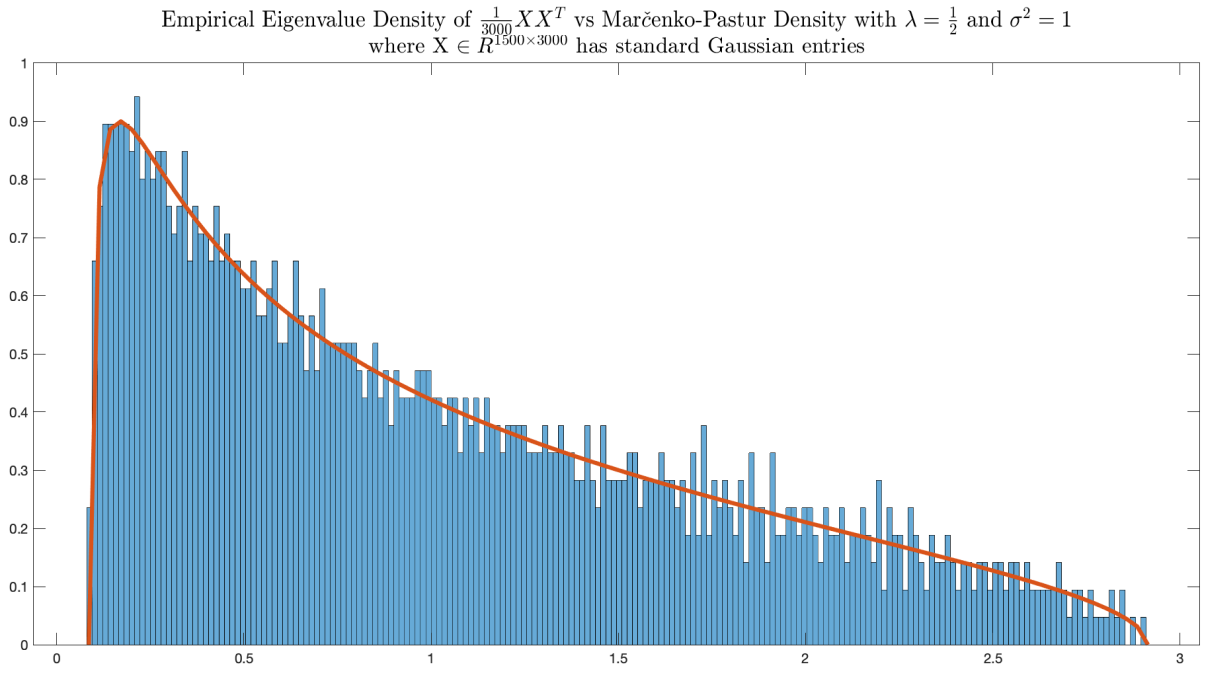


Figure 1.1: Example of the Marchenko-Pastur law.

Looking at the Marchenko-Pastur density function for various parameters values, λ and σ^2, is interesting. If we think of the columns of the matrix X as being samples of a vector drawn from a joint distribution, where the entries of the vector are i.i.d. mean zero and variance one, then as the number of samples goes to infinity (so the ratio $\lambda = \frac{\#rows}{\#columns} \to \infty$) the sample covariance matrix will approach the identity matrix, and thus the eigenvalues all tend to 1. This is depicted in Figure 1.2. On the other hand, fixing the ratio λ and varying the variance of the entries, σ^2, increasing σ^2 shifts the distribution to the right, as depicted in Figure 1.3. This is to be expected because again if we had infinitely many samples and variance σ^2, then the sample covariance matrix will approach $\sigma^2 I$, and thus the eigenvalues all tend to σ^2.

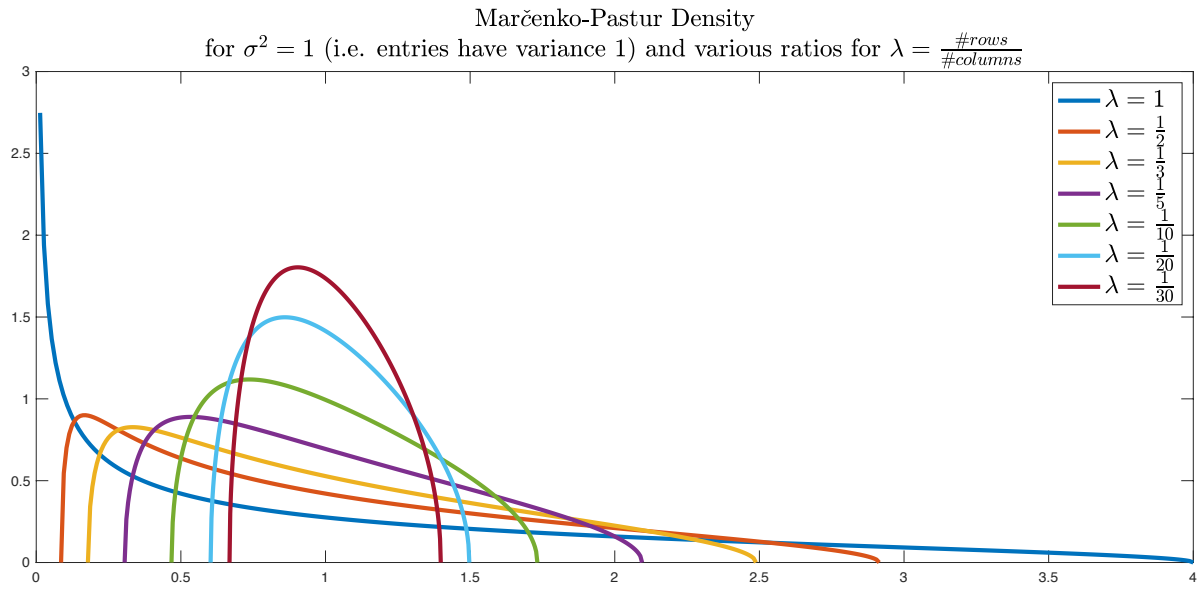


Figure 1.2: Visualizing the Marchenko-Pastur density for various ratios, $\lambda = \frac{\#rows}{\#columns}$.

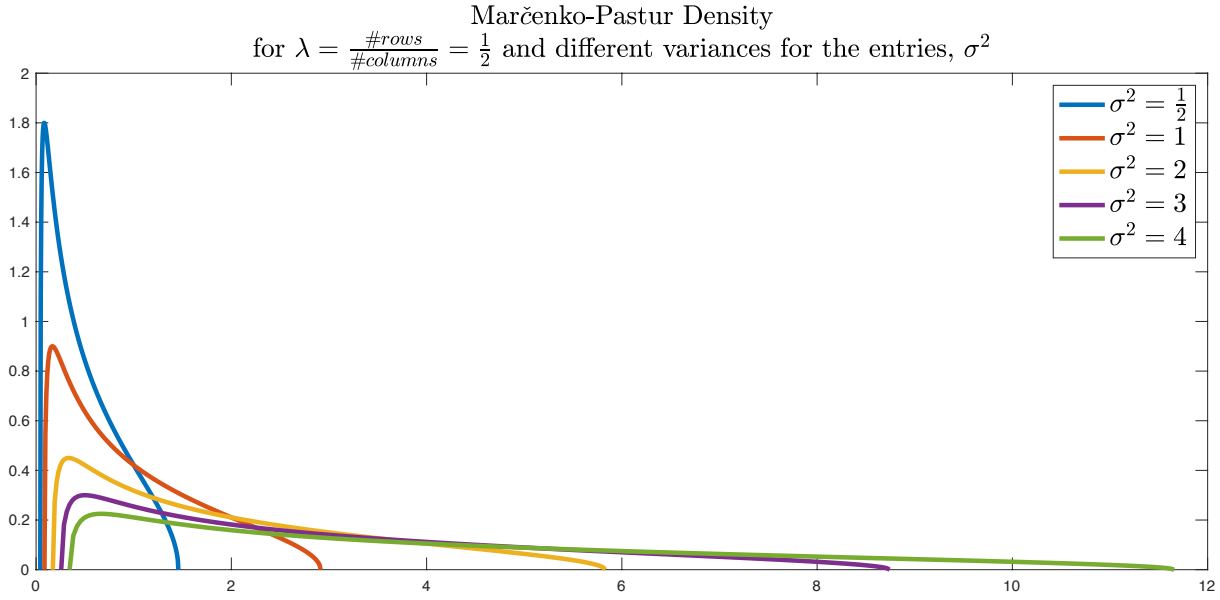


Figure 1.3: Visualizing the Marchenko-Pastur density for various entry variances, σ^2.

Before stating the general version of the Marchenko-Pastur law, we need to introduce the Stieltjes transform, which we motivate by discussing the two different methods for proving the Marchenko-Pastur law. The two methods are the moment method and using the Stieltjes transform.

The moment method is a technique that was introduced by Pafnuty Chebyshev for proving convergence in distribution.

Lemma 1.2 (Unique limit - Lemma B.1 in [7]). . *Let $\{F_p\}$ be a sequence of distribution functions. Let the k^{th} moment of the distribution F_p be denoted by*

$$\beta_{p.k} = \beta_k(F_p) := \int x^k dF_p(x).$$

The sequence of distribution functions $\{F_p\}$ converges weakly to a limit if the following conditions are satisfied:

1. *Each F_p has finite moments of all orders.*
2. *For each fixed integer $k \geq 0$, $\beta_{p,k}$ converges to a finite limit β_k as $p \to \infty$.*
3. *If two right-continuous nondecreasing functions F and G have the same moment sequence $\{\beta_k\}$, then $F = G + const.$*

Works by M. Riesz [Lemma B.2 in [7]] and Carleman [Lemma B.3 in [7]] give ways of showing condition (3) of the lemma. A proof of the simplified version of the Marchenko-Pastur law stated above is done via the moment method in Section 3.1 of [7].

An alternative way to proving a sequence of distribution functions converge in distribution is with the Stieltjes transform.

Definition 1.19. *If $G(x)$ is a function of bounded variation on the real line, then its* **Stieltjes transform** *is defined by*

$$s_G(z) = \int \frac{1}{\lambda - z} dG(\lambda)$$

where $z \in \mathbb{C}^+ \equiv \{z \in \mathbb{C} : \Im z > 0\}$.

When $G = F^{A_p}$ where F^{A_p} is the empirical spectral distribution of a $p \times p$ Hermitian matrix,

A_p, with eigenvalues $\lambda_1 \geq ... \geq \lambda_p$, then

$$s_{F^{A_p}}(z) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i - z} = \frac{1}{p}\operatorname{tr}(A_p - zI)^{-1}.$$

To understand how Stieltjes transforms help prove convergence in distribution we have the following definitions and theorem.

Definition 1.20. *A measure μ on $(\mathbb{R}^1, \mathcal{B}^1)$ with $\mu(\mathbb{R}^1) \leq 1$ is called a* **subprobability measure**.

Definition 1.21 ([36])**.** *Let $\mathcal{M}(\mathbb{R})$ denote the collection of all subprobability distribution functions on $\mathbb{R}$. We say for $\{F_p\} \subset \mathcal{M}(\mathbb{R})$, F_p* **converges vaguely** *to $F \in \mathcal{M}(\mathbb{R})$ (written $F_p \xrightarrow{v} F$) if for all $[a, b]$, a, b continuity points of F, $\lim_{p\to\infty} F_p\{[a, b]\} = F\{[a, b]\}$. Equivalently, $F_p \xrightarrow{v} F$ if*

$$\int f(x)dF_p(x) \to \int f(x)dF(x)$$

for all continuous f vanishing at $\pm\infty$. When F_p and F are probability distribution functions, we say F_p **converges in distribution** *to F, denoted $F_p \xrightarrow{D} F$. Equivalently, $\lim_{p\to\infty} F_p(a) = F(a)$ for all continuity points a of F.*

Theorem 1.3 (Theorem B.9 in [7])**.** *Assume that $\{G_p\}$ is a sequence of functions of bounded variation and $G_p(-\infty) = 0$ for all p. Then,*

$$\lim_{p\to\infty} s_{G_p}(z) = s(z) \quad \forall z \in D$$

if and only if there is a function of bounded variation G with $G(-\infty) = 0$ and Stieltjes transform $s(z)$ and such that $G_p \to G$ vaguely.

Here we will use $\{G_p\} = F^{A_p}$ where F^{A_p} is the empirical spectral distribution of a $p \times p$ Hermitian matrix, A_p. Thus if we show the limit of the Stieltjes transforms of F^{A_p} converge to

the Stieltjes transform of the distribution, F, from the Marchenko-Pastur law almost surely, then using Theorem B.9 in [7] we get that $F^{A_p} \to F$ vaguely with probability one. It has been shown that the Marchenko-Pastur distribution, F, is indeed a distribution. Thus with probability one we have convergence in distribution of the empirical spectral distributions to the Marchenko-Pastur distribution.

Theorem 1.4 (Marchenko-Pastur [28], as stated in [36])**.** *Consider an $p \times p$ matrix, B_p. Assume that*

1. *X_p is a $p \times m$ matrix such that the matrix elements X_{ij} are i.i.d. complex random variables with mean zero and variance one.*
2. *$m = m(p)$ with $m/p \to \hat{\lambda} > 0$ as $p \to \infty$.*
3. *$T_p = diag(\tau_1^p, \tau_2^p, ..., \tau_m^p) \in \mathbb{R}^{m \times m}$ where $\tau_i^p \in \mathbb{R}$, and the ESD of $\{\tau_1^p, ..., \tau_m^p\}$ converges almost surely in distribution to a nonrandom probability distribution function $H(\tau)$ as $p \to \infty$.*
4. *$B_p = A_p + \frac{1}{p} X_p T_p X_p^*$, where A_p is a random Hermitian $p \times p$ matrix for which F^{A_p} converges vaguely to F^A almost surely, A being a possibly defective (i.e. with discontinuities) nonrandom distribution function.*
5. *X_p, T_p and A_p are independent.*

Then, almost surely, F^{B_p} converges vaguely as $p \to \infty$ to a nonrandom distribution function F^B whose Stieltjes transform $s_{F^B}(z)$, $z \in \mathbb{C}^+$ satisfies the canonical equation

$$s_{F^B}(z) = s_{F^A}\left(z - \hat{\lambda} \int \frac{\tau dH(\tau)}{1 + \tau s_{F^B}(z)}\right)$$

where s_{F^A} is the Stieltjes transform of A. For example, if $A_m = 0$, then $s_{F^A}(z) = \frac{1}{0-z} = -\frac{1}{z}$.

Remark. Setting $A_p = 0$ and $T_p = I_m$ gives the more common version of the Marchenko-Pastur law that was stated above. It looks a little different because here $m/p \to \hat{\lambda}$ and $B = \frac{1}{p}XX^*$, and above $p/m \to \lambda$ and we considered $\frac{1}{m}XX^*(= \frac{p}{m}B)$, but the two are equivalent. Indeed, Theorem 1.4 gives us that

$$s_{F^B}(z) = \frac{1}{-z + \frac{\hat{\lambda}}{1+s_{F^B}(z)}}$$

Using the definition of Stieltjes transform, we have $s_{F^{\frac{p}{m}B}}(z) = \frac{m}{p}s_{F^B}\left(\frac{m}{p}z\right)$, thus the line above is equivalent to

$$s_{F^B}\left(\frac{m}{p}z\right) = \frac{1}{-\frac{m}{p}z + \frac{\hat{\lambda}}{1+s_{F^B}\left(\frac{m}{p}z\right)}}$$

$$\implies \frac{p}{m}s_{F^{\frac{p}{m}B}}(z) = \frac{1}{-\frac{m}{p}z + \frac{\hat{\lambda}}{1+\frac{p}{m}s_{F^{\frac{p}{m}B}}(z)}}$$

$$\implies s_{F^{\frac{p}{m}B}}(z) = \frac{1}{-z + \frac{1}{1+\frac{p}{m}s_{F^{\frac{p}{m}B}}(z)}} \text{ using } m/p \to \hat{\lambda}$$

$$\implies s_{F^{\frac{p}{m}B}}(z) = \frac{1}{-z + \frac{1}{1+\lambda s_{F^{\frac{p}{m}B}}(z)}} \text{ using } p/m \to \lambda$$

For notational simplicity, let $S = s_{F^{\frac{p}{m}B}}(z)$, thus we have $S = \frac{1}{-z+\frac{1}{1+\lambda S}}$, which can be rearranged into the equation $S = \frac{1}{1-\lambda-z-z\lambda S}$. The distribution with a Stieltjes transform that satisfies this equation is the Marchenko-Pastur distribution from Theorem 1.1; this is proved on pages 51-52 in Chapter 3 Section 2 of [14].

Remark. In [6] Bai and Silverstein dropped the restriction on T_p being diagonal so that T_p can be any matrix whose ESD $F^{T_p} \to H$.

Outline of the proof of the Marchenko-Pastur law using Stieltjes transforms

We give an outline for the proof of the Marchenko-Pastur law using Stieltjes transforms. This outline is based on the proof given in the textbook *Random Matrix Methods for Wireless Communications* [14]. As done in the textbook, we break the proof into six key facts:

Fact 1: The (1,1) entry $[(XX^T - zI_p)^{-1}]_{1,1} = \frac{1}{-z - zy^T(Y^TY - zI_m)^{-1}y}$.

Proof of Fact 1: This comes from a classical matrix inversion lemma:

Lemma 1.5. *Let $A \in \mathbb{C}^{N\times N}$, $d \in \mathbb{C}^{n\times n}$ be invertible, and $B \in \mathbb{C}^{N\times n}, C \in \mathbb{C}^{n\times N}$. Then*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(A - BD^{-1}C)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

Setting

$$X = \begin{pmatrix} y^T \\ Y \end{pmatrix},$$

the result follows by applying this lemma to the matrix

$$(XX^T - zI_p)^{-1} = \begin{pmatrix} y^Ty - z & y^TY^T \\ Yy & YY^T - zI_{p-1} \end{pmatrix}^{-1}$$

and also using the identity

$$I_N - A_{(N\times n)}(B_{(n\times N)}A_{(N\times n)} - zI_n)^{-1}B_{(n\times N)} = -z(A_{(N\times n)}B_{(n\times N)} - zI_N)^{-1}.\square$$

Fact 2: $y^T(Y^TY - zI_m)^{-1}y \to \frac{1}{m}\operatorname{tr}((Y^TY - zI_m)^{-1})$ almost surely.

This is the heart of the proof and is essentially the part that we prove in our work for our specific types of matrices. We do not prove it here, but in [14], it is called the Theorem 3.4. and is proved on page 45.

Fact 3: $\frac{1}{m}\operatorname{tr}((Y^TY - zI_m)^{-1}) \to \frac{1}{m}\operatorname{tr}((X^TX - zI_m)^{-1})$ almost surely.

Proof of Fact 3: This is known as the rank-1 perturbation lemma. We will use the following theorem.

Theorem 1.6 (Silverstein and Bai [6])**.** *For $z \in \mathbb{C} \setminus \mathbb{R}^+$, we have the following quadratic form identities.*
(i) Let $z \in \mathbb{C} \setminus \mathbb{R}$, $A \in \mathbb{C}^{N\times N}$, $B \in \mathbb{C}^{N\times N}$ with B Hermitian, and $v \in \mathbb{C}^N$. Then

$$\left| \operatorname{tr}\left((B - zI_N)^{-1} - (B + vv^H - zI_N)^{-1}\right)A \right| \leq \frac{||A||}{|\Im[z]|}$$

with $||A||$ the spectral norm of A.

(ii) Moreover, if B is non-negative definite, for $z \in \mathbb{R}^-$

$$\left| \operatorname{tr}\left((B - zI_N)^{-1} - (B + vv^H - zI_N)^{-1}\right)A \right| \leq \frac{||A||}{|z|}.$$

Returning to the proof of Fact 3, we have that

$$\frac{1}{m}\operatorname{tr}((Y^TY - zI_m)^{-1}) - \frac{1}{m}\operatorname{tr}((X^TX - zI_m)^{-1})$$

$$= \frac{1}{m}\operatorname{tr}((Y^TY - zI_m)^{-1}) - \frac{1}{m}\operatorname{tr}((Y^TY + yy^T - zI_m)^{-1})$$

$$= \frac{1}{m}\operatorname{tr}\left((Y^TY - zI_m)^{-1} - (Y^TY + yy^T - zI_m)^{-1}\right)$$

Using Siverstein and Bai's theorem with $A = I_m, B = Y^T Y$, and $v = y$ we have for $z \in \mathbb{R}^-$

$$\frac{1}{m}\Big| \operatorname{tr}\left((Y^T Y - zI_m)^{-1} - (Y^T Y + yy^T - zI_m)^{-1}\right)\Big| \leq \frac{1}{m|z|} \to 0 \text{ as } m \to \infty$$

and for $z \in \mathbb{C} \setminus \mathbb{R}$

$$\frac{1}{m}\Big| \operatorname{tr}\left((Y^T Y - zI_m)^{-1} - (Y^T Y + yy^T - zI_m)^{-1}\right)\Big| \leq \frac{1}{m|\Im[z]|} \to 0 \text{ as } m \to \infty.$$

Since straight lines have measure zero in $\mathbb{R}^2$, we have shown that $\frac{1}{m}\Big| \operatorname{tr}\left((Y^T Y - zI_m)^{-1} - (Y^T Y + yy^T - zI_m)^{-1}\right)\Big|$ has limit zero with probability one. $\square$

Fact 4: $\frac{1}{m}\operatorname{tr}((X^T X - zI_m)^{-1}) = \frac{1}{m}\operatorname{tr}((XX^T - zI_p)^{-1}) + \frac{p-m}{m}\frac{1}{z}$

Proof of Fact 4: This follows from the fact that the non-zero eigenvalues of XX^T and $X^T X$ are the same. For further details, see Lemma 3.1 in the textbook *Random Matrix Methods for Wireless Communications* [14].

Putting together Facts 1 & 2, we have that

$$\left|[(XX^T - zI_p)^{-1}]_{1,1} - \frac{1}{-z - z\frac{1}{m}\operatorname{tr}((Y^T Y - zI_m)^{-1})}\right| \to 0 \text{ a.s. as } m \to \infty$$

Adding in Fact 3 gives

$$\left|[(XX^T - zI_p)^{-1}]_{1,1} - \frac{1}{-z - z\frac{1}{m}\operatorname{tr}((X^T X - zI_m)^{-1})}\right| \to 0 \text{ a.s. as } m \to \infty$$

Adding in Fact 4 gives

$$\left|[(XX^T - zI_p)^{-1}]_{1,1} - \frac{1}{1 - \frac{p}{m} - z - z\frac{p}{m}\frac{1}{p}\operatorname{tr}((XX^T - zI_p)^{-1})}\right| \to 0 \text{ a.s. as } m \to \infty$$

Written in terms of the Stieltjes transform, $s_{XX^T}(z)$, of XX^T, this is

$$\left| [(XX^T - zI_p)^{-1}]_{1,1} - \frac{1}{1 - \frac{p}{m} - z - z\frac{p}{m}s_{XX^T}(z)} \right| \to 0 \text{ a.s. as } m \to \infty$$

Fact 5: Due to the symmetric structure of XX^T, the result is also true for all diagonal entires (i, i), $i = 1, ..., p$. Summing them up and averaging, we conclude that

$$s_{XX^T}(z) - \frac{1}{1 - \frac{p}{m} - z - z\frac{p}{m}s_{XX^T}(z)} \to 0 \text{ a.s. as } m \to \infty.$$

Fact 6: $s_{XX^T}(z)$ given by the equation in Fact 5 converges to the Stieltjes transform of the probability measure in the Marchenko-Pastur law.

These facts together give the proof of the Marchenko-Pastur law. □

Historical advances of the Marchenko-Pastur law

Since Marchenko and Pastur's influencial paper [28], many generalizations have been made such as by Grenander and Silverstein [23], Wachter [38], Jonnson [25], Yin and Krishnaiah [43], Yin [44], Silverstein [35], and Silverstein and Bai [6]. A summary of the achievements of these is given in the introduction section of [6]. Some other papers have weakened the independence condition on the matrix X. Yin and Krishnaiah [43] require the column vectors of X, X_k, to come from a spherically symmetric distribution; specifically, they require the distribution of X_k to be the same as that of PX_k where P is an orthogonal matrix. Aubrun [5] proved the result with X having independent columns distributed uniformly on the l_p^m ball. That result was generalized by Pajor and Pastur [31] who showed the independent columns can be distributed according to any isotropic log-concave measure. Adamczak [2] gives three conditions and shows they suffice to giving the limiting eigenvalue distribution.

Those conditions are (1) the entries of X have finite moments of all orders, (2) for every m, i, j, $\mathbb{E}[X_{ij}^{(m)}|\mathcal{F}_{ij}] = 0$, where $\mathcal{F}_{ij}$ is the σ-field generated by $\{X_{kl}^{(m)} : (k,l) \neq (i,j)\}$, and (3) the Euclidean norm of a random row, normalized by $\sqrt{m}$, and the Euclidean norm of a random column, normalized by $\sqrt{p}$, both converge in probability to 1. O'Rourke [30] studies a class of matrices with weakly dependent entries that satisfy seven conditions. Additionally, he gives a rate of convergence of the expected empirical spectral distribution. Bai and Zhou [8] prove the limiting distribution for a class of matrices, and their work was extended by Yao [42] to give a time series dependence structure as well. We state Bai and Zhou's result as we will use it in our work to prove our results.

Theorem (Bai, Zhou) [8] *Let $\{X_{jk}^{(p)}, j = 1, ..., p, k = 1, ..., m\}$ be an array of complex random variables for each p. Write $X = X^{(p)} = (X_{jk})_{1 \leq j \leq p, 1 \leq k \leq m}$. Let $X_1, ..., X_m$ be the columns of X, which are assumed to be independent. As $p \to \infty$, assume the following.*

1. *For all k, $\mathbb{E}\overline{X}_{jk}X_{lk} = t_{lj}$, and for any non-random $p \times p$ matrix $A = A^{(p)} = (a_{jk})$ with bounded spectral norm, $\mathbb{E}|X_k^* A X_k - \mathrm{tr}(AT)|^2 = o(p^2)$, where $T = T^{(p)} = (t_{jl})$.*
2. *$\lambda_{(p)} := p/m \to \lambda \in (0, \infty)$.*
3. *The norm of the matrix $T = T^{(p)}$ is uniformly bounded and the empirical spectral distribution of T tends to a non-random probability distribution H.*

Then, with probability 1, the empirical spectral distribution of $\frac{1}{m}XX^*$ tends to a probability distribution, whose Stieltjes transform $s = s(z)$ $(z \in \mathcal{C}^+)$ satisfies

$$s = \int \frac{1}{t(1 - \lambda - \lambda z s) - z} dH(t).$$

When $T = \sigma^2 I$, this implies the the empirical spectral distribution of $\frac{1}{m}XX^*$ converges weakly in distribution to the simple Marchenko-Pastur distribution with ratio λ and variance σ^2.

Applications of the Marchenko-Pastur law in finance and biology

The Marchenko-Pastur law has been applied to both the fields of finance and biology. In this section we provide some references for reading about these applications.

The authors Laloux, Cizeau, Bouchaud, and Potters in [27] and [26] study financial price fluctuations. They find that the empirical correlation matrices from stocks of the S&P 500 (or other major markets) agree remarkably with correlation matrices from random matrices. These authors have many other papers relating finance to the Marchenko-Pastur law which the interested reader can look up. Authors Plerou, Gopikrishnan, Rosenow, Amaral, Guhr, and Stanley in [33] also study correlation matrices of stock returns. They find that the eigenvalues that lie outside of the range for random matrices tend to fall into groups that relate to different business sectors.

In biology, the Marchenko-Pastur law arises without being explicitly called "the Marchenko-Pastur law" in two papers that study HIV. The paper [15], published by PNAS, analyzed HIV sequences and found groups of amino acids whose mutations are collectively coordinated. Another paper [12] considers the question, "what are optimal vaccine targets for the HIV virus?"

Chapter 3 of this dissertation gives an application to dimensionality reduction for random vectors. Lastly, we point the interested reader to the paper Random Matrix Theory and its Innovative Applications [18] for some additional applications.

1.3.4 Related works

In this section, we present Wigner's semicircular law which is analogous to the Marchenko-Pastur law in that the Marchenko-Pastur law gave the limiting empirical spectral distribution (ESD) for covariance matrices $\frac{1}{m}XX^T$, and the semicircular law gives the limiting normalized

ESD for Hermitian matrices. Wigner's work from the 1950s was motivated by problems in physic, and random matrices still have applications in physics today.

Definition 1.22. *[37] For any $p \times p$ Hermitian matrix M_p, the* **normalized empirical spectral distribution** *of M_p is*

$$\mu_{\frac{1}{\sqrt{p}}M_p} := \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i(M_p)/\sqrt{p}},$$

where $\lambda_1(M_p) \geq ... \geq \lambda_p(M_p)$ are the (necessarily real) eigenvalues of M_p, counting multiplicity.

We see that $\mu_{\frac{1}{\sqrt{p}}M_p}$ is a probability measure.

Definition 1.23. *[37] A sequence of random ESDs $\mu_{\frac{1}{\sqrt{p}}M_p}$* **converge almost surely** *to a deterministic limit μ, which is a probability measure on $\mathbb{R}$, if for every test function $\varphi \in C_C(\mathbb{R})$, the quantities $\int_{\mathbb{R}} \varphi d\mu_{\frac{1}{\sqrt{p}}M_p}$ converge almost surely to $\int_{\mathbb{R}} \varphi d\mu$.*

Definition 1.24. *[37] A* **Wigner matrix** *is a random Hermitian matrix where the strictly upper triangular entries are i.i.d. real or complex valued random variables with mean zero and variance one, and the diagonal entries are i.i.d. real valued random variables, independent of the strictly upper triangular entries, with bounded mean and variance.*

Theorem 1.7. *(Wigner's semicircular law [39], as stated in [37]) Let M_p be the top left $p \times p$ minors of an infinite Wigner matrix $(\zeta_{i,j})_{i,j\geq 1}$. Then the ESDs $\mu_{\frac{1}{\sqrt{p}}M_p}$ converge almost surely to the* **Wigner semicircular distribution**

$$\mu_{sc} := \frac{1}{2\pi}(4 - |x|^2)_+^{1/2} dx$$

as $p \to \infty$, where $(x)_+ = x$ if $x > 0$ and 0 otherwise.

There have been many extensions to the semicircle law for the non-i.i.d. case such as by Pastur in [32], Götze and Tikhomirov in [22], Erdős, Yau, Yin, et al., which is summarized in

[19], and Götze, Naumov, and Tikhomirov in [21]. Additionally, there have been a number of extensions towards block random matrices using free probability theory. In chapter 9.1 of Mingo and Speicher's textbook [29], they consider block matrices where the random matrix is made up of blocks of Gaussian random matrices, where it is allowed for the blocks to repeat. Other works involving block matrices include the work of Shlyakhtenko [34] and Arizmendi, Nechita, and Vargas [4]. In Chapter 2, we consider block structured matrices for the Marchenko-Pastur law.

1.3.5 Vector and matrix norms

Our proofs rely heavily on understanding matrix norms, so we collect some definitions and basic facts here for the reader. Note: the definitions of vector and matrix norms can be extended to $\mathbb{C}$ and other fields, but we just consider real values here.

Definition 1.25. *A function* $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ *is a* **vector norm** *if for* $x, y \in \mathbb{R}^n$ *and* $\alpha \in \mathbb{R}$ *we have*

1. $\|x\| \geq 0$, *with* $\|x\| = 0 \iff x = 0$
2. $\|x + y\| \leq \|x\| + \|y\|$
3. $\|\alpha x\| = |\alpha|\|x\|$.

To distinguish between different norms, we adorn the function notation $\|\cdot\|$ with a subscript. The most common vector norms are the $p - norms$ defined by

$$\|x\|_p = (|x_1|^p + ... + |x_n|^p)^{\frac{1}{p}}, \text{ for } p \geq 1.$$

Theorem 1.8 (Holder's inequality)**.** *For $x, y \in \mathbb{R}^n$ and $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$,*

$$|x^T y| \leq \|x\|_p \|y\|_q.$$

The case where $p = q = 2$ in Holder's inequality is known as the Cauchy-Schwarz inequality, which is important enough to state on its own. We state it in its more well-known form for vectors in $\mathbb{C}$.

Theorem 1.9 (Cauchy-Schwarz inequality)**.** *For $x, y \in \mathbb{C}^n$,*

$$|\langle x, y \rangle| = |x^* y| \leq \|x\|_2 \|y\|_2$$

or

$$\left| \sum_{i=1}^{n} x_i \bar{v}_i \right|^2 \leq \sum_{j=1}^{n} |x_j|^2 \sum_{k=1}^{n} |y_k|^2.$$

Definition 1.26. *A function $\|\cdot\| : \mathbb{R}^{m \times n} \to \mathbb{R}$ is a* **matrix norm** *if for $A, B \in \mathbb{R}^{m \times n}$ and $\alpha \in \mathbb{R}$ we have*

1. $\|A\| \geq 0$, *with* $\|A\| = 0 \iff A = 0$
2. $\|A + B\| \leq \|A\| + \|B\|$
3. $\|\alpha A\| = |\alpha| \|A\|$.

Similarly, the most common matrix norms are the $p - norms$ which are defined in terms of the vector $p - norms$, via

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}.$$

Of particular interest are the $1 - norm$, $\infty - norm$, and $2 - norm$ (which is also called the

operator norm or **spectral norm**), which are given by

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}| = \text{ the max absolute column sum of the matrix;}$$

$$\|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}| = \text{ the max absolute row sum of the matrix;}$$

$$\|A\|_2 = \sigma_{max}(A) = \text{ the largest singular value of the matrix.}$$

Another common matrix norm is the Frobenius norm,

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}.$$

Theorem 1.10. *If S is a real or complex vector space of finite dimension, then any two norms, $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, on S are equivalent, meaning there exists positive real constants c_1 and c_2 such that $c_1\|A\|_\alpha \le \|A\|_\beta \le c_2\|A\|_\alpha$ for all $A \in S$. In particular for $S = \mathbb{R}^{m\times n}$ and $A \in \mathbb{R}^{m\times n}$ with rank r, we have*

1. $\frac{1}{\sqrt{n}}\|A\|_\infty \le \|A\|_2 \le \sqrt{m}\|A\|_\infty$
2. $\frac{1}{\sqrt{m}}\|A\|_1 \le \|A\|_2 \le \sqrt{n}\|A\|_1$
3. $\|A\|_2 \le \|A\|_F \le \sqrt{r}\|A\|_2$.

1.3.6 Asymptotic notation

Definition 1.27. *Let f be a real or complex valued function and g a real valued function, both defined on an unbounded subset of the real positive numbers, and $g(x)$ is strictly positive for large enough values of x. We say f is* **"big O"** *of g, written $f(x) = O\big(g(x)\big)$ as $x \to \infty$,*

if there exists a positive real number M and a real number x_0 such that

$$|f(x)| \leq Mg(x) \text{ for all } x \geq x_0.$$

In words, f is "big O" of g if for large values of x, f grows no faster than a constant times g. In other words, the order of f is at most the order of g.

Definition 1.28. *Let f be a real or complex valued function and g a real valued function, both defined on an unbounded subset of the real positive numbers, and $g(x)$ is strictly positive for large enough values of x. We say f is* **"little-o"** *of g, written $f(x) = o\big(g(x)\big)$ as $x \to \infty$, if for every $\varepsilon > 0$ there exists a constant, N, such that*

$$|f(x)| \leq \varepsilon g(x) \text{ for all } x \geq N.$$

Meaning, f is "little-o" of g if for large values of x, f grows much slower than g. In other words, the order of f is strictly less than the order of g.

For example the function $f(x) = 3x^2 + 2x + 20$ is big O of x^2, and little-o of $x^{2+\varepsilon}$ for $\varepsilon > 0$.

1.3.7 Binomial coefficients

One of our proofs requires bounding equations full of binomial coefficients, so here we collect the facts about binomial coefficients that we will need.

Fact 1: [Definition]

$$\binom{n}{t} = \frac{n!}{t!(n-t)!}$$

Fact 2:

$$\left(\frac{n}{t}\right)^t \leq \binom{n}{t} \leq \left(\frac{en}{t}\right)^t$$

Fact 3:

$$\binom{2n}{n} \leq 4^n$$

Proof.

$$\binom{2n}{n} = \sum_{k=0}^{n} \binom{n}{k}^2 \leq \left(\sum_{k=0}^{n} \binom{n}{k}\right)^2 = (2^n)^2 = 4^n,$$

where the inequality comes from the fact that the terms of the sum are positive. □

Fact 4: [Chu-Vandermonde identity]

$$\sum_{v=1}^{t} \binom{t}{v}\binom{n-t}{t-v} = \left(\binom{n}{t} - \binom{n-t}{t}\right)$$

Fact 5: [Pascal's identity]

$$\binom{n}{t} = \binom{n-1}{t-1} + \binom{n-1}{t}$$

Fact 6:

$$\operatorname*{argmax}_b \binom{a}{b}\binom{a}{c-b} = \frac{c}{2}$$

Proof.

$$\operatorname*{argmax}_b \binom{a}{b}\binom{a}{c-b} = \operatorname*{argmax}_b \frac{a!}{b!(a-b!)} \frac{a!}{(c-b)!(a-c+b)!)}$$

$$\operatorname*{argmin}_b b!(a-b!)(c-b)!(a-c+b)! = \operatorname*{argmin}_b \frac{c!}{\binom{c}{b}} \frac{(2a-c)!}{\binom{2a-c}{a-b}} = \operatorname*{argmax}_b \binom{c}{b}\binom{2a-c}{a-b}$$

and noticing that $\binom{c}{b}$ is maximized when $b = \frac{c}{2}$, as is $\binom{2a-c}{a-b}$. □

Fact 7:

$$\binom{a}{b-1} = \frac{b}{a+1-b}\binom{a}{b}$$

Chapter 2

Weakening Independence in the Marchenko-Pastur Law

2.1 Block version and vectorized tensor version for the Marchenko-Pastur law

In this chapter, we add to the collection of works that generalize the Marchenko-Pastur theorem with some dependence, by showing that the hypotheses from Bai and Zhou's work in [8] is satisfied for our specific random matrices. Specifically we consider two models of random matrices and show that they exhibit the Marchenko-Pastur law. We will first give the models and theorem statement, then discuss the motivation for these types of models. Here are our two models of the columns of random matrices:

Model 1: Consider a random vector $x \in \mathbb{R}^{nd}$ whose entries all have mean zero, variance one, and finite fourth moment. Assume further that the entries of x can be partitioned into

n blocks each of length d in such a way that the entries within a block are uncorrelated, and the entries in different blocks are independent.

Model 2: Consider a random vector $x \in \mathbb{R}^n$ whose entries are independent and all have mean zero, variance one, and finite fourth moment. Consider the random vector $\mathbf{x} \in \mathbb{R}^{\binom{n}{t}}$ obtained by vectorizing the symmetric tensor $x^{\otimes d}$. Thus, the entries of $\boldsymbol{x}$ are indexed by t-element subsets $\mathbf{i} \subset [n]$ and are defined as products of the entries of x over $\mathbf{i}$:

$$\boldsymbol{x}_{\mathbf{i}} = \prod_{i \in \mathbf{i}} x_i = x_{i_1} x_{i_2} ... x_{i_t}, \quad \mathbf{i} = \{i_1, \ldots, i_t\}.$$

Theorem 2.1 (Marchenko-Pastur law for new models)**.** *Let $X = X^{(p)}$, $p = 1, 2, ...,$ where p is a function of another variable n ($p = nd$ for Model 1 and $p = \binom{n}{t}$ for Model 2), be a sequence of $p \times m$ random matrices whose columns are independent and follow either Model 1 or Model 2* [1]*. Let $n, p \to \infty$ and $m = m(p)$ be so that $p/m \to \lambda \in (0, \infty)$. For Model 1, assume that $\max_\alpha \mathbb{E}[x_\alpha^4] = o(\frac{n}{d})$, and for Model 2, assume that $\max_\alpha \mathbb{E}[x_\alpha^4] = o(\frac{n^{2/3}}{t^2})$. Then the empirical spectral distribution of $\frac{1}{m} XX^T$ converges weakly in distribution to the Marchenko-Pastur distribution with ratio λ.*

Remark. For Model 1, by variance one and $\max_\alpha \mathbb{E}[x_\alpha^4] = o(\frac{n}{d})$ we necessarily have that $d = o(n)$. However, the next remark shows how we we remove this restriction on d. For Model 2, by variance one and $\max_\alpha \mathbb{E}[x_\alpha^4] = o(\frac{n^{2/3}}{t^2})$ we necessarily have that $t = o(n^{\frac{1}{3}})$.

Remark. If in addition to Model 1 we have exchangeability in the random variables in a block, meaning that $\mathbb{E}[x_i x_j x_k x_l] = \mathbb{E}[x_q x_r x_s x_t]$ when $x_i, x_j, x_k, x_l, x_q, x_r, x_s, x_t$ are all in the same block of uncorrelated random variables, then we can relax the condition on the 4^{th} moment to just require that the largest 4^{th} moment of the entries is $o(n)$. We still require the number of blocks, n, to go to infinity, but now d can be as large as we want. For example, we could have $d = n^2$.

[1] We allow the size of the block d in Model 1 and the degree t in Model 2 depend on p, i.e. they can be different for different matrices $X^{(p)}$ in the sequence.

The motivation for the block model is that we believe this model gives a more realistic interpretation which explains why the Marchenko-Pastur law has worked well in the applications where it's been used. For example, when used in finance, the analysts currently are assuming each day's stock values are independent and each stock price is independent, but it's more realistic to think the sectors of the stock market form groups and the stocks within a sector are not necessarily independent. The block model would also work well towards studying people in situations such as Netflix recommendations because it's fine to assume that each person's ratings are independent of each other, but for a specific person their ratings for two movies in the same genre are probably not independent. While our assumption of uncorrelated entries within a block can't handle correlated data yet, we believe our work can be extended in the future to allow for correlation within a block, provided the correlation has a limiting spectral distribution.

As for the vectorized tensor model, vectorized tensors come up in a variety of places and not too much theory has been developed yet for them. In statistics, suppose we wish to model lung capacity of a person as a linear function of their age, height, and gender. Often times, including interaction terms, such as age$*$gender may give better results. If we make a vector containing age, height, gender, and the number one, then when we tensor that with itself we end up with those features plus all pairwise products. Including all pairwise products of the features is essentially considering degree two polynomials, but still framing it as a linear problem. In a recent paper by Baldi and Vershynin [9], they linearized polynomial threshold functions by lifting them into the tensor product space. This reduced polynomial threshold functions to linear threshold functions and the corresponding hyperplane arrangements. Using theory on random tensors, much of which they had to create, they were able to solve their problem. Vectorized random tensors also arise in coding theory because they can save memory storage over storing a completely random vector of the same length, for example see [1].

2.1.1 Idea of the Proof

Theorem 2.1 will be proved by showing that the hypotheses from Bai and Zhou's work in [8] is satisfied for our specific random matrices. We state Bai and Zhou's theorem now:

Theorem (Bai, Zhou) [8] Let $\{X_{jk}^{(p)}, j = 1, ..., p, k = 1, ..., m\}$ be an array of complex random variables for each p. Write $X = X^{(p)} = (X_{jk})_{1 \leq j \leq p, 1 \leq k \leq m}$. Let $X_1, ..., X_m$ be the columns of X, which are assumed to be independent. As $p \to \infty$, assume the following.

1. For all k, $\mathbb{E}\overline{X}_{jk}X_{lk} = t_{lj}$, and for any non-random $p \times p$ matrix $A = A^{(p)} = (a_{jk})$ with bounded spectral norm, $\mathbb{E}|X_k^* A X_k - \mathrm{tr}(AT)|^2 = o(p^2)$, where $T = T^{(p)} = (t_{jl})$.
2. $p/m \to \lambda \in (0, \infty)$.
3. The norm of the matrix $T = T^{(p)}$ is uniformly bounded and the empirical spectral distribution of T tends to a non-random probability distribution H.

Then, with probability 1, the empirical spectral distribution of $\frac{1}{m}XX^*$ tends to a probability distribution, whose Stieltjes transform $s = s(z)$ $(z \in \mathcal{C}^+)$ satisfies

$$s = \int \frac{1}{t(1 - \lambda - \lambda z s) - z} dH(t).$$

When $T = I$, this implies the the empirical spectral distribution of $\frac{1}{m}XX^*$ converges weakly in distribution to the Marchenko-Pastur distribution with ratio λ.

For our models, we have isotropic vectors, meaning $T = I$. Since each column vector, X_k, is independent and of the same structure, we will drop the subscript k and use a lowercase x to represent an arbitrary column vector. We will use subscripts on x to represent the entries

within the vector, i.e. x_i is the i^{th} entry in the arbitrary column vector. We also consider real entries in X. Thus the property we need to show is

$$\mathbb{E}[|x^T Ax - \mathrm{tr}(A)|^2] = o(p^2)$$

where $x \in \mathbb{R}^p$ is an arbitrary column vector of X, $A \in \mathbb{R}^{p\times p}$ is any matrix with bounded spectral norm, and $\frac{p}{m} \to \lambda$. Here are the statements of our results, which when combined with Theorem (Bai, Zhou) give Theorem 2.1.

Theorem 2.2 (Concentration of quadratic forms in Model 1)**.** *Let $x \in \mathbb{R}^{nd}$ be a random vector that follows Model 1. Then, for any fixed matrix $A \in \mathbb{R}^{nd\times nd}$, we have*

$$\mathbb{E}\left[|x^{\mathsf{T}} Ax - \mathrm{tr}(A)|^2\right] \le CK\|A\|^2 nd^3.$$

Here C is an absolute constant and K is the largest fourth moment of the entries of x.

Moreover, if the entries of x within the same block are exchangeable,[2] *then the bound improves to $CK\|A\|^2 nd^2$.*

Theorem 2.3 (Concentration of quadratic forms in Model 2)**.** *Let $\boldsymbol{x} \in \mathbb{R}^{\binom{n}{d}}$ be a random vector that follows Model 2. Then, for any fixed matrix $A \in \mathbb{R}^{\binom{n}{t}\times\binom{n}{t}}$, we have*

$$\mathbb{E}[|\boldsymbol{x}^T A\boldsymbol{x} - \mathrm{tr}(A)|^2] \le C\|A\|^2 \binom{n}{t}^2 f\left(\frac{K^{\frac{3}{2}}t^3}{n}\right).$$

Here C is an absolute constant, K is the largest fourth moment of the entries of x, and $f\left(\frac{K^{\frac{3}{2}}t^3}{n}\right)$ is a function of $\frac{K^{\frac{3}{2}}t^3}{n}$ which is $o(1)$ provided $\frac{K^{\frac{3}{2}}t^3}{n} = o(1)$.

Remark. It is allowed for $t = t(n) \to \infty$ provided $\max_{\substack{\alpha\in\{1,\dots,n\}\\ \beta=3,4}} |\mathbb{E}[x_\alpha^\beta]| = o(\frac{n^{1/2}}{t^{3/2}})$, which necessarily implies $t = o(n^{1/3})$ since the entries of x have variance one.

[2]Exchangeability means that $\mathbb{E}[x_i x_j x_k x_l] = \mathbb{E}[x_q x_r x_s x_t]$ when i, j, k, l, q, r, s, t are all in the same block.

Combining Theorem 2.2 and Theorem 2.3 with Theorem (Bai, Zhou) [8] proves Theorem 2.1. The remainder of the chapter is organized as follows. We prove Theorem 2.2 in Section 2.2, and Theorem 2.3 in Section 2.3. Lastly, in Section 2.4 we give numerical experiments which show how well the results hold for a single realization and relatively small values of n.

2.2 Proof of Theorem 2.2

Without loss of generality, we may assume that $\|A\| = 1$ by rescaling. Expanding $x^{\mathsf{T}}Ax$ as a double sum of terms $A_{ij}x_ix_j$ and distinguishing the cases when $i = j$ or $i \neq j$, we have:

$$\mathbb{E}\left[|x^{\mathsf{T}}Ax - \operatorname{tr} A|^2\right] \leq 2\,\mathbb{E}\left[\Big(\sum_{i=1}^{nd} A_{ii}(x_i^2 - 1)\Big)^2\right] + 2\,\mathbb{E}\left[\Big(\sum_{i\neq j} A_{ij}x_ix_j\Big)^2\right] \tag{2.1}$$

where we used the inequality $(a+b)^2 \leq 2a^2 + 2b^2$.

2.2.1 Diagonal contribution

Let us start by considering the first expectation on the right-hand side of (2.1). Expanding the square, we can express it as

$$2\sum_{i,k=1}^{nd} A_{ii}A_{kk}\,\mathbb{E}(x_i^2 - 1)(x_k^2 - 1). \tag{2.2}$$

Now if i and k are in different blocks, then by independence and unit variance, all such terms have expectation zero and do not contribute anything in (2.2). So, the only contribution comes from the terms where i and k are in the same block. In such a case, we have the

bound

$$|\,\mathbb{E}(x_i^2-1)(x_k^2-1)| = |Cov(x_i^2,x_j^2)| \le \max_\alpha \mathbb{E}[x_\alpha^4] = K.$$

Thus, the contribution of the terms for which j and k are in the *first* block is

$$2\sum_{i,k=1}^{d} A_{ii}A_{kk}\,\mathbb{E}(x_i^2-1)(x_k^2-1) \le K\sum_{i,k=1}^{d} A_{ii}A_{kk} \le K\|A\|^2\sum_{i,k=1}^{d} 1 = Kd^2\|A\|^2 = Kd^2,$$

where we have used in the second inequality that every entry of A is bounded by its spectral norm.

Clearly, the same bound holds not only for the first block but for each of the n blocks. Summing up these bounds, we conclude that the expression in (2.2) is $\lesssim Knd$. Thus, we have shown that the first expectation on the right-hand side of (2.1) is $\lesssim Knd$.

2.2.2 Off-diagonal contribution: setting up partitions

We now consider the second term in the right-hand side of (2.1). Ignoring the 2 out front and expanding the square into

$$\mathbb{E}\sum_{i\neq j}\sum_{k\neq l} A_{ij}A_{kl}x_ix_jx_kx_l, \tag{2.3}$$

we can break this into cases considering partitions of 4, representing the powers on the entries of x. For example, the partition $(2,1,1)$ represents the indices (i,j,k,l) for which $i\neq j$, $k\neq l$ and such that exactly two among the four indices are the same. This comprises the terms of the form $A_{ij}A_{il}x_i^2x_jx_l$, $A_{ij}A_{ki}x_i^2x_jx_k$, $A_{ij}A_{jl}x_ix_j^2x_l$, and $A_{ij}A_{kj}x_ix_j^2x_k$.

Notice that in all partitions of 4 we have to consider, i.e. those for which $i\neq j$ and $k\neq l$, none of the powers can be greater than 2. (Indeed, in the partition $(3,1)$ three indices among i,j,k,l are the same, which violates either the constraint $i\neq j$ or the constraint $k\neq l$.) This

leaves us with partitions $(2,2)$, $(2,1,1)$ and $(1,1,1,1)$, which we shall consider one by one.

2.2.3 Partition $(2,2)$

The terms corresponding to this partition have the form $A_{ij}^2 x_i^2 x_j^2$ and $A_{ij}A_{ji}x_i^2x_j^2$. Notice that

$$|\,\mathbb{E}[x_i^2x_j^2]| \le \big(\mathbb{E}[x_i^4]\big)^{1/2}\big(\mathbb{E}[x_j^4]\big)^{1/2} \le K \tag{2.4}$$

Thus, we can bound the net contribution of the terms of the form $A_{ij}^2x_i^2x_j^2$ as follows:

$$\mathbb{E}\sum_{\substack{i,j=1\\ i\neq j}}^{nd} A_{ij}^2x_i^2x_j^2 \le K\sum_{i,j=1}^{nd} A_{ij}^2 \le Knd$$

where in the last step we used the bound

$$\sum_{i,j=1}^{nd} A_{ij}^2 = \|A\|_F^2 \le nd\|A\|^2 = nd. \tag{2.5}$$

Similarly, we can bound the the net contribution of the terms of the form $A_{ij}A_{ji}x_i^2x_j^2$:

$$\mathbb{E}\sum_{\substack{i,j=1\\ i\neq j}}^{nd} A_{ij}A_{ji}x_i^2x_j^2 \le K\sum_{i,j=1}^{nd}|A_{ij}||A_{ji}| \le K\Big(\sum_{i,j=1}^{nd}A_{ij}^2\Big)^{1/2}\Big(\sum_{i,j=1}^{nd}A_{ji}^2\Big)^{1/2} \le Knd$$

where we used the moment bound (2.4), Cauchy-Schwarz inequality, and (2.5).

Concluding, the net contribution to (2.3) of the terms comprising the partition $(2,2)$ is $\lesssim Knd$.

2.2.4 Partition $(1, 1, 1, 1)$

The terms corresponding to this partition have the form $A_{ij}A_{kl}x_ix_jx_kx_l$ where all indices i, j, k, l are distinct.

We claim that the expectation of such a term is zero unless all four indices i, j, k, l come from *the same block*. Indeed, suppose one of these indices – let's say i – comes from its own block, and none of the other three indices i, j, k are in the same block as i. Then $\mathbb{E}\, x_ix_jx_kx_l = 0$ since in this case x_i is independent of x_j, x_k and x_l and has zero mean. The remaining situation is where a pair of the indices – let's say i, j – are in one block, and the other pair k, l is in a different block. Then again $\mathbb{E}\, x_ix_jx_kx_l = \mathbb{E}[x_ix_j]\,\mathbb{E}[x_kx_l] = 0$ since x_ix_j is independent of x_kx_l, and moreover x_i and x_j are uncorrelated. This proves the claim.

So, let us assume that all indices i, j, k, l are in the same block. In such a case, we have the bound

$$|\,\mathbb{E}[x_ix_jx_kx_l]| \le \max_{\alpha} \mathbb{E}[x_\alpha^4] \le K. \tag{2.6}$$

Thus, the contribution of the terms for which i, j, k, l are in the *first* block can be bounded by

$$\mathbb{E} \sum_{\substack{i,j,k,l=1 \\ i,j,k,l \text{ distinct}}}^{d} A_{ij}A_{kl}x_ix_jx_kx_l \le K \sum_{i,j,k,l=1}^{d} |A_{ij}A_{kl}| = K\Big(\sum_{i,j=1}^{d} |A_{ij}|\Big)^2 \le Kd^2 \sum_{i,j=1}^{d} A_{ij}^2 \le Kd^3. \tag{2.7}$$

In the last step, we used that top-left $d \times d$ minor of A, which we denote by $A_{d\times d}$, satisfies

$$\sum_{i,j=1}^{d} A_{ij}^2 = \|A_{d\times d}\|_F^2 \le d\|A_{d\times d}\|^2 \le d\|A\|^2 = d.$$

Clearly, a bound similar to (2.7) holds not only for the first block but for each of the n blocks. Summing up these bounds, we conclude that the net contribution to (2.3) of the terms corresponding to the partition $(1,1,1,1)$ is $\lesssim Knd^3$.

2.2.5 Partition $(2,1,1)$

This partition comprises the terms of the form $A_{ij}A_{il}x_i^2x_jx_l$, $A_{ij}A_{ki}x_i^2x_jx_k$, $A_{ij}A_{jl}x_ix_j^2x_l$, and $A_{ij}A_{kj}x_ix_j^2x_k$. Just like in the previous case where we studied the partition $(1,1,1,1)$, we can argue that the only nonzero contribution comes from the terms where all three indices are *in the same block*.

Let us consider the terms of the form $A_{ij}A_{il}x_i^2x_jx_l$ first. In such a case, we have the bound

$$|\,\mathbb{E}[x_i^2x_jx_l]| \le \max_{\alpha} \mathbb{E}[x_\alpha^4] \le K.$$

Thus, the contribution of the terms for which i,j,l are in the *first* block can be bounded by

$$\mathbb{E} \sum_{\substack{i,j,l=1 \\ i,j,l \text{ distinct}}}^{d} A_{ij}A_{il}x_i^2x_jx_l \le K \sum_{i,j,l=1}^{d} |A_{ij}A_{il}| = K \sum_{i=1}^{d} \Bigg(\sum_{j=1}^{d} |A_{ij}| \Bigg)^2 \le Kd^2.$$

In the last step, we used that top-left $d \times d$ minor of A, which we denote by $A_{d\times d}$, satisfies

$$\sum_{j=1}^{d} |A_{ij}| \le \sqrt{d} \Bigg(\sum_{j=1}^{d} A_{ij}^2 \Bigg)^{1/2} \le \sqrt{d}\|A_{d\times d}\| = \sqrt{d} \quad \text{for every } i.$$

A similar result holds not only for the first block but for each of the n blocks. Summing up these bounds, we conclude that the net contribution to (2.3) of the terms of the form $A_{ij}A_{il}x_i^2x_jx_l$ is $\lesssim Knd^2$. Finally, we can repeat the above argument for the terms of the other three types, $A_{ij}A_{ki}x_i^2x_jx_k$, $A_{ij}A_{jl}x_ix_j^2x_l$, and $A_{ij}A_{kj}x_ix_j^2x_k$, and thus conclude that the

net contribution to (2.3) of the terms corresponding to the partition $(2,1,1)$ is $\lesssim Knd^2$.

Ultimately, adding the contributions of all partitions – $(2,2)$, $(1,1,1,1)$, and $(2,1,1)$, we see that the total off-diagonal contribution (2.3) is bounded by

$$O(Knd) + O(Knd^3) + O(Knd^2) = O(Knd^3).$$

Adding to this the the diagonal contribution $O(Knd)$, which we handled in Section 2.2.1, we conclude that (2.1) is bounded by $O(Knd^3)$. Theorem 2.2 is proved.

2.2.6 Exchangeable distributions

Here we prove the "moreover" part of the Theorem 2.2. Namely, we assume that the entries of x within the same block are exchangeable, and we seek to improve our previous bound on (2.1) from $O(Knd^3)$ to $O(Knd^2)$. A quick look at the previous paragraph reveals that the only part that needs to be strengthened is the partition $(1,1,1,1)$, where our bound was $O(Knd^3)$; it suffices to improve it to $O(Knd^2)$.

So let us focus on the partition $(1,1,1,1)$. Thus, we will have to bound the sum of the terms $A_{ij}A_{kl}x_ix_jx_kx_l$ over quadruples i,j,k,l of all distinct indices. As we argued in the beginning of Section 2.2.4, we can assume without generality that for each term, the indices i,j,k,l are in *the same block* (otherwise the expectation of such a term would be zero). Focusing on the *first* block for now, we are seeking to bound the quantity

$$\left|\mathbb{E}\sum_{\substack{i,j,k,l=1\\ i,j,k,l \text{ distinct}}}^{d} A_{ij}A_{kl}x_ix_jx_kx_l\right| = \left|\mathbb{E}[x_1x_2x_3x_4]\sum_{\substack{i,j,k,l=1\\ i,j,k,l \text{ distinct}}}^{d} A_{ij}A_{kl}\right| \le K\left|\sum_{\substack{i,j,k,l=1\\ i,j,k,l \text{ distinct}}}^{d} A_{ij}A_{kl}\right|, \tag{2.8}$$

where the equality follows from the exchangeability assumption, and the inequality follows

from (2.6). We reduced the problem to bounding the sum of $A_{ij}A_{kl}$ over quadruples i, j, k, l of all distinct indices in $\{1, \ldots, d\}$. The following lemma provides an adequate bound.

Lemma 2.4. *Any real $d \times d$ matrix B satisfies*

$$\Big| \sum_{\substack{i,j,k,l=1 \\ i,j,k,l \text{ distinct}}}^{d} B_{ij}B_{kl} \Big| \leq 10d^2\|B\|^2.$$

Proof. Without loss of generality, we can assume that $\|B\| = 1$. Denote by $\mathbf{1} \in \mathbb{R}^d$ the vector with all 1 coordinates and note that

$$\Big| \sum_{i,j=1}^{d} B_{ij} \Big| = |\mathbf{1}^\mathsf{T} B \mathbf{1}| \leq \|\mathbf{1}\|_2^2 \|B\| = d \quad \text{and} \quad \Big| \sum_{i=1}^{d} B_{ii} \Big| = |\operatorname{tr} B| \leq d\|B\| = d.$$

Subtracting one sum from the other and using triangle inequality, we get

$$\Big| \sum_{\substack{i,j=1 \\ i \neq j}}^{d} B_{ij} \Big| \leq 2d.$$

This yields

$$\Big| \sum_{\substack{i,j,k,l=1 \\ i \neq j,\ k \neq l}}^{d} B_{ij}B_{kl} \Big| = \Big(\sum_{\substack{i,j=1 \\ i \neq j}}^{d} B_{ij} \Big)^2 \leq 4d^2. \tag{2.9}$$

Thus, instead of the requirement that all four indices i, j, k, l be distinct, we were able to handle the weaker but simpler requirement that $i \neq j$ and $k \neq l$. This weaker requirement produces the sum over more terms. It remains to control the sum over the difference set E, i.e. over the set of quadruples of indices i, j, k, l not all of which are distinct but for which $i \neq j$ and $k \neq l$.

This set E can be expressed as the union of the following four sets:

$$E_1 := \big\{(i,j,k,l) : j \neq i = k \neq l\big\}, \qquad E_2 := \big\{(i,j,k,l) : j \neq i = l \neq k\big\},$$
$$E_3 := \big\{(i,j,k,l) : i \neq j = k \neq l\big\}, \qquad E_4 := \big\{(i,j,k,l) : i \neq j = l \neq k\big\}.$$

By the inclusion-exclusion principle, the sum of any terms w_{ijkl} over E can be expressed as

$$\sum_E w_{ijkl} = \sum_{E_1} w_{ijkl} + \sum_{E_2} w_{ijkl} + \sum_{E_3} w_{ijkl} + \sum_{E_4} w_{ijkl} - \sum_{E_1 \cap E_4} w_{ijkl} - \sum_{E_2 \cap E_3} w_{ijkl}, \quad (2.10)$$

since only two of all pairwise intersections are nonempty, namely $E_1 \cap E_4$ and $E_2 \cap E_3$, and all three-wise and four-wise intersections are empty.

The sum over E_α can be bounded by the same argument as in Section 2.2.5, which we repeat here for the reader's convenience. For example, let us bound the sum over E_1:

$$\Big| \sum_{\substack{i,j,l=1 \\ i \neq j}}^{d} B_{ij} B_{il} \Big| \leq \sum_{i,j,l=1}^{d} |B_{ij}||B_{il}| = \sum_{i=1}^{d} \Big(\sum_{j=1}^{d} |B_{ij}| \Big)^2 \leq d^2.$$

In the last step, we used that

$$\sum_{j=1}^{d} |B_{ij}| \leq \sqrt{d} \Big(\sum_{j=1}^{d} B_{ij}^2 \Big)^{1/2} \leq \sqrt{d} \|B\| = \sqrt{d} \quad \text{for every } i.$$

Similarly one can bound the sums over E_2, E_3, and E_4.

The sum over the pairwise intersections is simpler to bound. For example, let us bound the sum over $E_2 \cap E_3$:

$$\Big| \sum_{\substack{i,j=1 \\ i \neq j}}^{d} B_{ij} B_{ji} \Big| \leq \sum_{i,j=1}^{d} |B_{ij}||B_{ji}| \leq d^2.$$

In the last step, we used that the magnitude of each entry of B is bounded by the spectral norm of B, which we assumed to be 1. The sum over $E_1 \cap E_4$ can be handled similarly.

Thus, each of the six sums on the right in the inclusion-exclusion formula (2.10) is bounded by d^2. Hence the sum over E is bounded by $6d^2$. Thus, the difference of the sum over all-distinct i, j, k, l and over the larger set where $i \neq j$, $k \neq l$ can be at most $6d^2$. Using (2.9), this implies that sum over all-distinct i, j, k, l is at most $4d^2 + 6d^2 \leq 10d^2$. The lemma is proved. □

Apply Lemma 2.4 for the top-left $d \times d$ minor of A, and substitute into (2.8). We obtain

$$\left| \mathbb{E} \sum_{\substack{i,j,k,l=1 \\ i,j,k,l \text{ distinct}}}^{d} A_{ij} A_{kl} x_i x_j x_k x_l \right| \leq 10Kd^2.$$

A similar bound clearly holds not only when the indices i, j, k, l are in the first block, but also for each of the n blocks. Summing up these bounds, we conclude that the net contribution of the terms corresponding to the partition $(1, 1, 1, 1)$ is $\lesssim Knd^2$. As we explained in the beginning of this section, is completes the proof of the "moreover" part of Theorem 2.2. □

2.3 Proof of Theorem 2.3

The first step of the proof is the same as for Theorem 2.2. Without loss of generality, we may assume that $\|A\| = 1$ by rescaling. Expanding $\boldsymbol{x}^\mathsf{T} A \boldsymbol{x}$ as a double sum of terms $A_{\mathbf{ij}} \boldsymbol{x}_\mathbf{i} \boldsymbol{x}_\mathbf{j}$ and distinguishing the cases when $\mathbf{i} = \mathbf{j}$ or $\mathbf{i} \neq \mathbf{j}$, we have:

$$\mathbb{E}\left[|\boldsymbol{x}^\mathsf{T} A \boldsymbol{x} - \operatorname{tr} A|^2 \right] \leq 2\,\mathbb{E}\left[\Big(\sum_{\mathbf{i}} A_{\mathbf{ii}} (\boldsymbol{x}_\mathbf{i}^2 - 1) \Big)^2 \right] + 2\,\mathbb{E}\left[\Big(\sum_{\mathbf{i} \neq \mathbf{j}} A_{\mathbf{ij}} \boldsymbol{x}_\mathbf{i} \boldsymbol{x}_\mathbf{j} \Big)^2 \right] \tag{2.11}$$

2.3.1 Diagonal contribution

Let us start by considering the first expectation on the right-hand side of (2.11). Expanding the square, we can express it as

$$2\sum_{\mathbf{i},\mathbf{k}} A_{\mathbf{ii}} A_{\mathbf{kk}}\, \mathbb{E}(\boldsymbol{x}_{\mathbf{i}}^2 - 1)(\boldsymbol{x}_{\mathbf{k}}^2 - 1). \tag{2.12}$$

Both meta-indices $\mathbf{i}$ and $\mathbf{k}$ range in all $\binom{n}{t}$ subsets of $[n]$ of cardinality t. Let v denote the overlap between these two subsets, i.e.

$$v := |\mathbf{i} \cap \mathbf{k}|.$$

If $v = 0$, the subsets are disjoint, the random variables $\boldsymbol{x}_{\mathbf{i}}^2 - 1$ and $\boldsymbol{x}_{\mathbf{k}}^2 - 1$ are independent and have mean zero, and thus

$$\mathbb{E}(\boldsymbol{x}_{\mathbf{i}}^2 - 1)(\boldsymbol{x}_{\mathbf{k}}^2 - 1) = 0.$$

If $v \geq 1$, the monomial $\boldsymbol{x}_{\mathbf{i}}^2 \boldsymbol{x}_{\mathbf{k}}^2$ consists of v terms raised to the fourth power (coming from indices that are both in $\mathbf{i}$ and $\mathbf{k}$) and $2(t-v)$ terms raised to the second power (coming from the symmetric difference of $\mathbf{i}$ and $\mathbf{k}$). Thus,

$$\left|\, \mathbb{E}(\boldsymbol{x}_{\mathbf{i}}^2 - 1)(\boldsymbol{x}_{\mathbf{k}}^2 - 1)\right| \leq \mathbb{E}\, \boldsymbol{x}_{\mathbf{i}}^2 \boldsymbol{x}_{\mathbf{k}}^2 \leq \max_{\alpha} \left(\mathbb{E}\, x_\alpha^4\right)^v \cdot \max_{\beta} \left(\mathbb{E}\, x_\beta^2\right)^{2(t-v)} \leq K^v,$$

where we used the unit variance assumption.

There are $\binom{n}{t}$ ways to choose $\mathbf{i}$. Once we fix $\mathbf{i}$ and $v \in \{1, \ldots, t\}$, there are $\binom{t}{v}\binom{n-t}{t-v}$ ways to choose $\mathbf{k}$, since v indices must come from $\mathbf{i}$ and the remaining $t - v$ indices must come from

$[n] \setminus \mathbf{i}$. Therefore,

$$(2.12) \leq 2\binom{n}{t}\sum_{v=1}^{t}\binom{t}{v}\binom{n-t}{t-v}K^v. \tag{2.13}$$

To bound this sum, since K is a positive integer, then the following elementary inequality holds:

$$\binom{t}{v}K^v \leq \binom{Kt}{v},$$

and it can be quickly checked by writing the binomial coefficients in terms of factorials. Now, if we sum v from zero as opposed from 1 in (2.13), we can use Vandermonde's identity and get

$$\sum_{v=0}^{t}\binom{t}{v}\binom{n-t}{t-v}K^v \leq \sum_{v=0}^{t}\binom{Kt}{v}\binom{n-t}{t-v} = \binom{n-t+Kt}{t}.$$

Subtracting the zeroth term, we obtain

$$\sum_{v=1}^{t}\binom{t}{v}\binom{n-t}{t-v}K^v \leq \binom{n-t+Kt}{t} - \binom{n-t}{t}.$$

Now use a stability property of binomial coefficients (Lemma 2.5), which tells us that

$$\binom{n-t+Kt}{t} - \binom{n-t}{t} \leq \delta\binom{n-t}{t} \quad \text{where } \delta := \frac{2Kt^2}{n-2t+1},$$

as long as $\delta \leq 1/2$. According to our assumptions on the degree t, we do have $\delta \leq 1/2$.

Summarizing, we have shown that

$$(2.12) \leq 2\binom{n}{t}\cdot\delta\binom{n-t}{t} \lesssim \binom{n}{t}^2 \cdot \frac{Kt^2}{n}.$$

Lemma 2.5 (Stability of binomial coefficients)**.** *For any positive integers m, p and $t \leq m$,*

we have

$$\binom{m+p}{t} \leq (1+\delta)\binom{m}{t} \quad \textit{where } \delta := \frac{2tp}{m+1-t},$$

as long as $\delta \leq 1/2$.

Proof. Definition of binomial coefficients gives

$$\frac{\binom{m+p}{t}}{\binom{m}{t}} = \prod_{k=1}^{p}\left(1+\frac{t}{m-t+k}\right) \leq \left(1+\frac{t}{m-t+1}\right)^{p}.$$

From Bernoulli's inequality and linearizing the exponential function, we will use the bound $(1+\epsilon)^p \leq e^{\epsilon p} \leq 1+2\epsilon p$, which holds as long as $\epsilon p \in [0,1]$. □

2.3.2 Off-diagonal contribution

It remains to consider the second expectation on the right-hand side of (2.11). Ignoring the 2 and expanding the square, we have

$$\mathbb{E}\sum_{\mathbf{i}=1}^{\binom{n}{t}}\sum_{\substack{\mathbf{j}=1\\ \mathbf{j}\neq\mathbf{i}}}^{\binom{n}{t}}\sum_{\mathbf{k}=1}^{\binom{n}{t}}\sum_{\substack{\mathbf{l}=1\\ \mathbf{l}\neq\mathbf{k}}}^{\binom{n}{t}} A_{\mathbf{i},\mathbf{j}}A_{\mathbf{k},\mathbf{l}}\boldsymbol{x}_i\boldsymbol{x}_j\boldsymbol{x}_{\mathbf{k}}\boldsymbol{x}_{\mathbf{l}}. \tag{2.14}$$

Rewriting the $\boldsymbol{x}$'s in terms of x's via

$$\boldsymbol{x}_{\mathbf{i}}\boldsymbol{x}_{\mathbf{j}}\boldsymbol{x}_{\mathbf{k}}\boldsymbol{x}_{\mathbf{l}} = x_{i_1}...x_{i_t}x_{j_1}...x_{j_t}x_{k_1}...x_{k_t}x_{l_1}...x_{l_t} \tag{2.15}$$

where $i_1,...,i_t,j_1,...,j_t,k_1,...,k_t,l_1,...,l_t \in [n]$, where $|\{i_1,...,i_t\}\cap\{j_1,...,j_t\}| \neq t$ and $|\{k_1,...,k_t\}\cap\{l_1,...,l_t\}| \neq t$. Because of independence and mean zero, the expectation of these terms will be zero unless each $x_\cdot$ is of power at least two. There are $4t$ x's on the right-hand side of

(2.15), so if each $x_\cdot$ has power two exactly we would have $2t$ variables. For simplicity, let's consider the case when each $x_\cdot$ has power two exactly. The general case will be handled afterwards and it will also handle this simplified case, but the simplified case shows us that our bound cannot be improved.

For now, fix an $\mathbf{i} = \{i_1, ..., i_t\} \in \binom{n}{t}$. Now $\mathbf{j} = \{j_1, ..., j_t\}$ may have some overlap with $\mathbf{i} = \{i_1, ..., i_t\}$ or not. When $|\mathbf{i} \cap \mathbf{j}| = v$ for $0 \leq v \leq t-1$, then the number of options for $\mathbf{j}$ that we have is $\binom{t}{v}\binom{n-t}{t-v}$ because we select v of $\mathbf{j}$'s indices from the t indices of $\mathbf{i}$ and the remaining $t-v$ indices are selected from the numbers not in $\mathbf{i}$. Now the number of options for $\mathbf{k}$ will be $\binom{2(t-v)}{t-v}\binom{n-(2t-v)}{v}$ because we need to make sure the $2(t-v)$ single indices from $\mathbf{i}$ and $\mathbf{j}$ get a pair, so half of them must be part of $\mathbf{k}$ and the other half must be part of $\mathbf{l}$, and the remaining v indices of $\mathbf{k}$ will be numbers not used yet. Once $\mathbf{i}, \mathbf{j}$, and $\mathbf{k}$ are determined, then $\mathbf{l}$ is determined. Thus (2.14) with a fixed $\mathbf{i}$ and considering overlap of v between $\mathbf{i}$ and $\mathbf{j}$ becomes

$$\mathbb{E} \sum_{j\left(\binom{t}{v}\binom{n-t}{t-v}\right)} \sum_{k\left(\binom{2(t-v)}{t-v}\binom{n-(2t-v)}{v}\right)} \sum_{l\left(1\right)} A_{\mathbf{i},\mathbf{j}} A_{\mathbf{k},\mathbf{l}} \boldsymbol{x}_{\mathbf{i}} \boldsymbol{x}_{\mathbf{j}} \boldsymbol{x}_{\mathbf{k}} \boldsymbol{x}_{\mathbf{l}} \tag{2.16}$$

where the notation $\mathbf{j}(\cdot)$ means $\mathbf{j}$ is a sum over $\cdot$ many terms. Now because each $x_\cdot$ has power exactly two, mean zero and variance one, then $\mathbb{E}[\boldsymbol{x}_{\mathbf{i}} \boldsymbol{x}_{\mathbf{j}} \boldsymbol{x}_{\mathbf{k}} \boldsymbol{x}_{\mathbf{l}}] = 1$, so we have that

$$(2.16) \leq \sum_{\mathbf{j}\left(\binom{t}{v}\binom{n-t}{t-v}\right)} \sum_{\mathbf{k}\left(\binom{2(t-v)}{t-v}\binom{n-(2t-v)}{v}\right)} \sum_{\mathbf{l}\left(1\right)} |A_{\mathbf{i},\mathbf{j}}||A_{\mathbf{k},\mathbf{l}}|$$

$$\leq \|A\| \sum_{\mathbf{j}\left(\binom{t}{v}\binom{n-t}{t-v}\right)} |A_{\mathbf{i},\mathbf{j}}| \sum_{\mathbf{k}\left(\binom{2(t-v)}{t-v}\binom{n-(2t-v)}{v}\right)} 1$$

$$\leq \|A\|^2 \left(\binom{t}{v}\binom{n-t}{t-v}\right)^{1/2} \binom{2(t-v)}{t-v}\binom{n-(2t-v)}{v}, \tag{2.17}$$

where for the last inequality sign we have used that the max row sum for any submatrix, $\hat{A} \in \mathbb{R}^{\binom{n}{t}\times\binom{t}{v}\binom{n-t}{t-v}}$, of matrix A has the bound $\|\hat{A}\|_\infty \leq \sqrt{\binom{t}{v}\binom{n-t}{t-v}}\|A\|$.

Notice that

$$\binom{n-t}{t-v}\binom{n-(2t-v)}{v} = \frac{(n-t)!}{(t-v)!(n-2t+v)!}\frac{(n-2t+v)!}{v!(n-2t)!}$$

$$= \frac{(n-t)!}{(t-v)!v!(n-2t)!} = \binom{t}{v}\frac{(n-t)!}{t!(n-2t)!} = \binom{t}{v}\binom{n-t}{t}.$$

Therefore,

$$(2.17) = \|A\|^2\binom{n-t}{t}\frac{\left(\binom{t}{v}\right)^{3/2}\binom{2(t-v)}{t-v}}{\left(\binom{n-t}{t-v}\right)^{1/2}}. \tag{2.18}$$

Using the inequalities, $\left(\frac{n}{t}\right)^t \leq \binom{n}{t} \leq \left(\frac{en}{t}\right)^t$ and $\binom{2n}{n} \leq 2^{2n} = 4^n$, we have that

$$\frac{\left(\binom{t}{v}\right)^{3/2}\binom{2(t-v)}{t-v}}{\left(\binom{n-t}{t-v}\right)^{1/2}} = \frac{\left(\binom{t}{t-v}\right)^{3/2}\binom{2(t-v)}{t-v}}{\left(\binom{n-t}{t-v}\right)^{1/2}} \leq \left(\frac{(et)}{t-v}\right)^{\frac{3(t-v)}{2}} 4^{(t-v)}\left(\frac{(t-v)}{n-t}\right)^{\frac{(t-v)}{2}}$$

$$= \frac{4^{t-v}e^{\frac{3(t-v)}{2}}t^{\frac{3(t-v)}{2}}}{(t-v)^{t-v}(n-t)^{\frac{(t-v)}{2}}} = \left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(t-v)(n-t)^{\frac{1}{2}}}\right)^{t-v}$$

$$\leq \left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}}\right)^{t-v}. \tag{2.19}$$

Combining (2.19) with (2.17) and (2.18) we have that,

$$(2.16) \leq \|A\|^2 \binom{n-t}{t} \left(\frac{4e^{\frac{3}{2}} t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}} \right)^{t-v}. \tag{2.20}$$

It remains to sum (2.20) over the options for $\mathbf{i}$ and v, which will give us a bound on (2.14).

Notice

$$\sum_{v=0}^{t-1} \left(\frac{4e^{\frac{3}{2}} t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}} \right)^{t-v} = \sum_{q=1}^{t} \left(\frac{4e^{\frac{3}{2}} t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}} \right)^{q} = \left(\frac{4e^{\frac{3}{2}} t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}} \right) \frac{1 - \left(\frac{4e^{\frac{3}{2}} t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}} \right)^{t}}{1 - \left(\frac{4e^{\frac{3}{2}} t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}} \right)}. \tag{2.21}$$

Using (2.20), (2.21), and accounting for the $\binom{n}{t}$ possibilities of $\mathbf{i}$, we have that overall

$$(2.14) \leq \|A\|^2 \binom{n}{t} \binom{n-t}{t} \left(\frac{4e^{\frac{3}{2}} t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}} \right) \frac{1 - \left(\frac{4e^{\frac{3}{2}} t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}} \right)^{t}}{1 - \left(\frac{4e^{\frac{3}{2}} t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}} \right)}$$

This completes the simplified case where each $x_{\cdot}$ has power two exactly. The simplified case is nice because there are not many inequalities and we see $t = o(n^{\frac{1}{3}})$ is needed for this proof method.

We now consider the general case where $x_{\cdot}$ does not have to have power two exactly. If each $x_{\cdot}$ in (2.15) does not have power two exactly we will have at most $2t-1$ variables and at least $t+1$ variables, since $\mathbf{i} \neq \mathbf{j}$. Let w be such that the number of variables is $2t-w$ for $0 \leq w \leq t-1$ (the case $w=0$ is what we called the simplified case which we can include here). Again we will let v represent the number of overlapping indices in $\mathbf{i} = \{i_1, ..., i_t\}$ and $\mathbf{j} = \{j_1, ..., j_t\}$ for $w \leq v \leq t-1$. Note $v \geq w$ otherwise overlap of v would not be possible while having only $2t-w$ variables. Furthermore we will denote r as

the number of indices of $\mathbf{k} = \{k_1, ..., k_t\}$ that overlap with the v overlapping indices of $\mathbf{i}$ and $\mathbf{j}$ for $0 \leq r \leq \min\{v, t-v+w, 2w\}$ ($r \leq t-v+w$ since we first must put $v-w$ of $\mathbf{k}$'s indices towards getting the correct number of variables and then remaining indices could go towards r; and $r \leq 2w$ otherwise $\mathbf{l}$ would have to have more than t indices in order to provide pairs to every index which didn't yet have a pair, and this isn't possible - this is seen when we count the number of option for $\mathbf{l}$ below). For a pictorial view of the variables w, v, and r, see Figure 2.1. Now we have $\binom{n}{t}$ options for $\mathbf{i}$. For a fixed $\mathbf{i}$, we then have $\binom{t}{v}\binom{n-t}{t-v}$ choices for $\mathbf{j}$, since we need to pick the v overlapping indices from $\mathbf{i}$'s indices and the remaining $t-v$ indices from new indices. Now for a fixed $\mathbf{i}$ and $\mathbf{j}$, we have $\binom{v}{r}\binom{n-(2t-v)}{v-w}\binom{2(t-v)}{t-r-(v-w)}$ options for $\mathbf{k}$ because $v-w$ indices must be new to get the correct total number of variables, r of the indices come from the v overlapping indices of $\mathbf{i}$ and $\mathbf{j}$, and the remaining indices must come from the set $(\{i_1, ..., i_t\} \cap \{j_1, ..., j_t\})^C$. Lastly, the number of options for $\mathbf{l}$ will be $\binom{t+w-r}{2w-r}$ because the $v-w$ new indices of $\mathbf{k}$ will require a pair and the $t-v-w+r$ indices of $(\{i_1, ..., i_t\} \cap \{j_1, ..., j_t\})^C$ that didn't get a pair via $\mathbf{k}$ will require a pair, this means $t+r-2w$ indices of $\mathbf{l}$ are determined, and the remaining $2w-r$ indices are selected from $2t-w-(t+r-2w)$ many items. Note that actually this is potentially over counting of options on $\mathbf{l}$ since we haven't restricted $\mathbf{l} \neq \mathbf{k}$, but over counting is not a problem for the proof below. Thus (2.14) with a fixed $w, v, r, \mathbf{i}$ becomes bounded by

$$\sum_{\mathbf{j}\left(\binom{t}{v}\binom{n-t}{t-v}\right)} \sum_{\mathbf{k}\left(\binom{v}{r}\binom{n-(2t-v)}{v-w}\binom{2(t-v)}{t-r-(v-w)}\right)} \sum_{\mathbf{l}\left(\binom{t+w-r}{2w-r}\right)} |A_{\mathbf{i},\mathbf{j}} A_{\mathbf{k},\mathbf{l}}| |\mathbb{E}[\boldsymbol{x}_{\mathbf{i}} \boldsymbol{x}_{\mathbf{j}} \boldsymbol{x}_{\mathbf{k}} \boldsymbol{x}_{\mathbf{l}}]|.$$

Furthermore in this setting with $2t-w$ variables, the product

$$\boldsymbol{x}_{\mathbf{i}} \boldsymbol{x}_{\mathbf{j}} \boldsymbol{x}_{\mathbf{k}} \boldsymbol{x}_{\mathbf{l}} = x_{i_1} ... x_{i_t} x_{j_1} ... x_{j_t} x_{k_1} ... x_{k_t} x_{l_1} ... x_{l_t}$$

can have at most $2w$ x's that are a power higher than two. Indeed, each of the distinct $2t-w$ variables must have power at least two to avoid having expectation zero, which uses

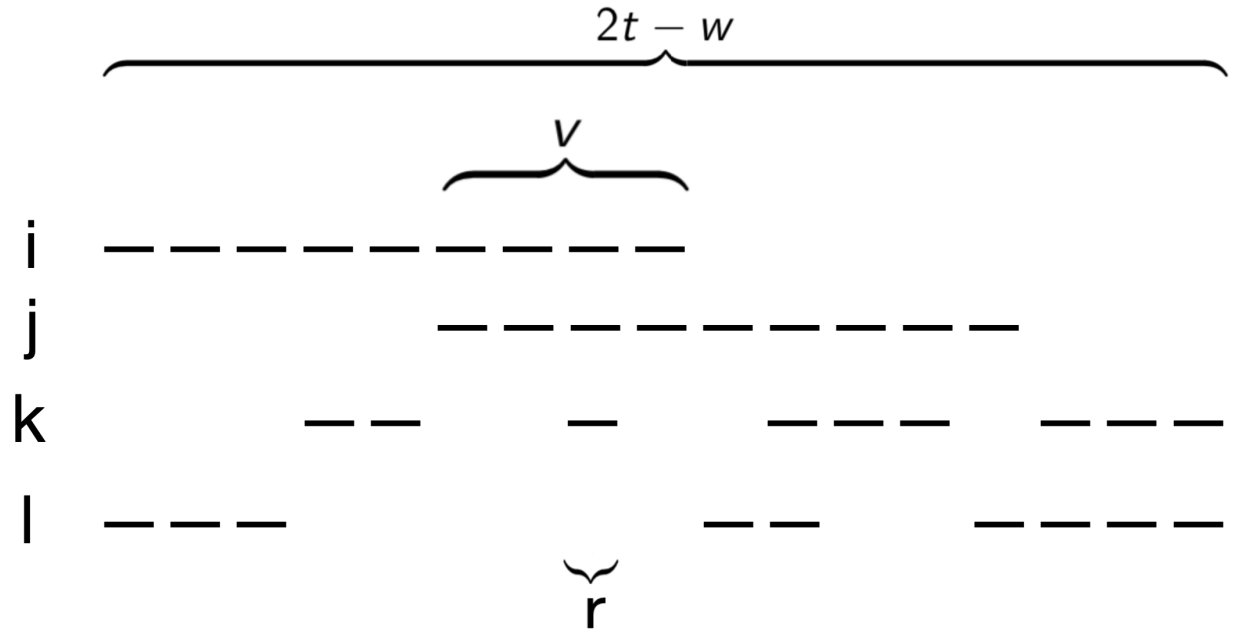


Figure 2.1: Pictorial view of the variables used for counting the off-diagonal terms.

$2(2t-w)$ indices up, and there are only $4t$ indices, so this leaves $2w$ extra indices that can go toward making some powers larger than two and not more than four. Suppose that we make f powers to be four, and thus $2w-2f$ powers to be three (and the rest are power two which we don't bother writing since they have variance one). Therefore using independence we have

$$\begin{aligned}|\mathbb{E}[\boldsymbol{x}_{\mathbf{i}}\boldsymbol{x}_{\mathbf{j}}\boldsymbol{x}_{\mathbf{k}}\boldsymbol{x}_{\mathbf{l}}]| &\leq \mathbb{E}[|x_{i_1}|^4]...\mathbb{E}[|x_{i_f}|^4]\mathbb{E}[|x_{j_1}|^3]...\mathbb{E}[|x_{j_{2w-2f}}|^3] \\ &\leq \left(\max_\alpha \mathbb{E}[|x_\alpha|^4]\right)^f \left(\max_\alpha \mathbb{E}[|x_\alpha|^3]\right)^{2w-2f}\end{aligned}$$

For absolute moments we have by Hölder's inequality that $E[|x_\alpha|^3])^{1/3} \leq E[|x_\alpha|^4])^{1/4}$, so we have

$$|\mathbb{E}[\boldsymbol{x}_{\mathbf{i}}\boldsymbol{x}_{\mathbf{j}}\boldsymbol{x}_{\mathbf{k}}\boldsymbol{x}_{\mathbf{l}}]| \leq \left(\max_\alpha \mathbb{E}[|x_\alpha|^4]\right)^f \left(\max_\alpha \mathbb{E}[|x_\alpha|^4]\right)^{\frac{3(2w-2f)}{4}} = \left(\max_\alpha \mathbb{E}[|x_\alpha|^4]\right)^{\frac{3w-f}{2}},$$

which is maximized when $f=0$. This finally gives that

$$|\mathbb{E}[\boldsymbol{x}_{\mathbf{i}}\boldsymbol{x}_{\mathbf{j}}\boldsymbol{x}_{\mathbf{k}}\boldsymbol{x}_{\mathbf{l}}]| \leq \left(\max_\alpha \mathbb{E}[|x_\alpha|^4]\right)^{\frac{3w}{2}} = K^{\frac{3w}{2}}$$

Thus (2.14) with a fixed $w, v, r, \mathbf{i}$ is bounded by

$$\leq C^2 K^{\frac{3w}{2}} \left(\binom{t}{v} \binom{n-t}{t-v} \right)^{1/2} \binom{v}{r} \binom{n-(2t-v)}{v-w} \binom{2(t-v)}{t-r-(v-w)} \left(\binom{t+w-r}{2w-r} \right)^{1/2} \tag{2.22}$$

Notice that

$$\binom{n-(2t-v)}{v-w} \leq \binom{n-(2t-v)}{v} \left(\frac{v}{n-2t+1} \right)^w \leq \binom{n-(2t-v)}{v} \left(\frac{t}{n-2t+1} \right)^w,$$

using the fact that $\binom{a}{b-1} = \frac{b}{a+1-b}\binom{a}{b}$.

Additionally, since $r \leq t - v + w$, we have that $v \leq t + w - r$, therefore

$$\binom{v}{r} \left(\binom{t+w-r}{2w-r} \right)^{1/2} \leq \binom{v}{r} \binom{t+w-r}{2w-r} \leq \binom{t+w-r}{r} \binom{t+w-r}{2w-r}$$

$$\leq \binom{t+w-r}{w} \binom{t+w-r}{w} \leq \left(\binom{t+w}{w} \right)^2 \leq \left(\binom{2t}{w} \right)^2 \leq (2et)^{2w}$$

where the third inequality can be seen with the following argument:

$$\operatorname*{argmax}_b \binom{a}{b} \binom{a}{c-b} = \operatorname*{argmax}_b \frac{a!}{b!(a-b!)} \frac{a!}{(c-b)!(a-c+b)!)}$$

$$= \operatorname*{argmin}_b b!(a-b!)(c-b)!(a-c+b)! = \operatorname*{argmin}_b \frac{c!}{\binom{c}{b}} \frac{(2a-c)!}{\binom{2a-c}{a-b}} = \operatorname*{argmax}_b \binom{c}{b} \binom{2a-c}{a-b} = \frac{c}{2}$$

since $\binom{c}{b}$ is maximized when $b = c/2$ and similarly $\binom{2a-c}{a-b}$ is maximized when $b = c/2$.

Therefore

$$(2.22) \leq \|A\|^2 K^{\frac{3w}{2}} \left(\binom{t}{v} \binom{n-t}{t-v} \right)^{1/2} \binom{n-(2t-v)}{v} \binom{2(t-v)}{t-v} (2et)^{2w} \left(\frac{t}{n-2t+1} \right)^w$$

$$= \|A\|^2 \binom{t}{v}\binom{n-t}{t-v}^{1/2} \binom{n-(2t-v)}{v}\binom{2(t-v)}{t-v}\left(\frac{4K^{\frac{3}{2}}e^2t^3}{n-2t+1}\right)^w$$

$$\leq \|A\|^2 \binom{n-t}{t}\left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}}\right)^{t-v}\left(\frac{4K^{\frac{3}{2}}e^2t^3}{n-2t+1}\right)^w,$$

where the last inequality used (2.18) and (2.19) from our work on the simplified case with exactly $2t$ distinct indices.

Now summing over all of the options of r, using that $0 \leq r \leq \min\{v, t-v+w, 2w\}$, in particular $0 \leq r \leq 2w$, we are bounded by

$$\|A\|^2 \binom{n-t}{t}\left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}}\right)^{t-v}(2w+1)\left(\frac{4K^{\frac{3}{2}}e^2t^3}{n-2t+1}\right)^w.$$

Summing over the possibilities for v, $w \leq v \leq t-1$, we are bounded by

$$\|A\|^2 \binom{n-t}{t}(2w+1)\left(\frac{4K^{\frac{3}{2}}e^2t^3}{n-2t+1}\right)^w \sum_{v=w}^{t-1}\left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}}\right)^{t-v}$$

$$\leq \|A\|^2 \binom{n-t}{t}(2w+1)\left(\frac{4K^{\frac{3}{2}}e^2t^3}{n-2t+1}\right)^w \sum_{v=0}^{t-1}\left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}}\right)^{t-v}$$

$$= \|A\|^2 \binom{n-t}{t}(2w+1)\left(\frac{4K^{\frac{3}{2}}e^2t^3}{n-2t+1}\right)^w \left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}}\right)\left(\frac{1-\left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}}\right)^t}{1-\left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}}\right)}\right),$$

since $\sum_{v=0}^{t-1} r^{t-v} = \sum_{q=1}^{t} r^q = r\sum_{q=0}^{t-1} r^q = r\frac{1-r^t}{1-r}$ for $r \neq 1$.

Now we will sum over the possibilities for w, $0 \leq w \leq t-1$, and for notation simplicity we will set

$$R := \frac{4K^{\frac{3}{2}}e^2t^3}{n-2t+1}$$

and since $K > 1$ (using Hölder's inequality and the variance one assumption), this gives that

$$\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}} \leq \frac{4K^{3/4}e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}} \leq \frac{4K^{3/4}e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-2t+1)^{\frac{1}{2}}} \leq 2e^{\frac{1}{2}}\frac{2K^{3/4}et^{\frac{3}{2}}}{(n-2t+1)^{\frac{1}{2}}} = 2e^{\frac{1}{2}}R^{\frac{1}{2}} < 4R^{\frac{1}{2}}$$

we have

$$\|A\|^2\binom{n-t}{t}\left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}}\right)\left(\frac{1-\left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}}\right)^t}{1-\left(\frac{4e^{\frac{3}{2}}t^{\frac{3}{2}}}{(n-t)^{\frac{1}{2}}}\right)}\right)\sum_{w=0}^{t-1}(2w+1)\left(\frac{4K^{\frac{3}{2}}e^2t^3}{n-2t+1}\right)^w$$

$$\leq \|A\|^2\binom{n-t}{t}\left(4R^{\frac{1}{2}}\right)\left(\frac{1-\left(4R^{\frac{1}{2}}\right)^t}{1-\left(4R^{\frac{1}{2}}\right)}\right)\sum_{w=0}^{t-1}(2w+1)R^w$$

$$= \|A\|^2\binom{n-t}{t}\left(4R^{\frac{1}{2}}\right)\left(\frac{1-\left(4R^{\frac{1}{2}}\right)^t}{1-\left(4R^{\frac{1}{2}}\right)}\right)\left(2\sum_{w=0}^{t-1}wR^w+\sum_{w=0}^{t-1}R^w\right)$$

$$\leq \|A\|^2\binom{n-t}{t}\left(4R^{\frac{1}{2}}\right)\left(\frac{1-\left(4R^{\frac{1}{2}}\right)^t}{1-\left(4R^{\frac{1}{2}}\right)}\right)\left(\frac{2R}{(1-R)^2}+\frac{1}{1-R}\right), \text{ provided } |R|<1$$

Finally, accounting for the $\binom{n}{t}$ options for $\mathbf{i}$ we have the final bound of

$$4\|A\|^2\binom{n}{t}\binom{n-t}{t}R^{\frac{1}{2}}\left(\frac{1-\left(4R^{\frac{1}{2}}\right)^t}{1-\left(4R^{\frac{1}{2}}\right)}\right)\left(\frac{2R}{(1-R)^2}+\frac{1}{1-R}\right), \text{ provided } |R|<1. \quad (2.23)$$

Looking at $R = \frac{4K^{\frac{3}{2}}e^2t^3}{n-2t+1} \lesssim 4e^2\frac{K^{\frac{3}{2}}t^3}{n}$, which indeed will be less than one in absolute value

when $\frac{K^{\frac{3}{2}}t^3}{n} = o(1)$, we finally have

$$(2.23) \leq C\|A\|^2 \binom{n}{t}\binom{n-t}{t} g\left(\frac{K^{\frac{3}{2}}t^3}{n}\right)$$

where $g(\frac{K^{\frac{3}{2}}t^3}{n})$ is a function of $\frac{K^{\frac{3}{2}}t^3}{n}$ that is $o(1)$ when $\frac{K^{\frac{3}{2}}t^3}{n} = o(1)$.

This completes the off-diagonal contribution thus bounding the right-hand side of (2.11). Putting together the diagonal bound and the off-diagonal bound we obtain the total bound of

$$C\|A\|^2 \left(\binom{n}{t}\right)^2 \frac{Kt^2}{n} + C\|A\|^2 \binom{n}{t}\binom{n-t}{t} g\left(\frac{K^{\frac{3}{2}}t^3}{n}\right)$$

$$\leq C\|A\|^2 \left(\binom{n}{t}\right)^2 \frac{K^{\frac{3}{2}}t^3}{n} + C\|A\|^2 \left(\binom{n}{t}\right)^2 g\left(\frac{K^{\frac{3}{2}}t^3}{n}\right)$$

$$\leq C\|A\|^2 \left(\binom{n}{t}\right)^2 f\left(\frac{K^{\frac{3}{2}}t^3}{n}\right)$$

where $f(\frac{K^{\frac{3}{2}}t^3}{n})$ is a function of $\frac{K^{\frac{3}{2}}t^3}{n}$ that is $o(1)$ when $\frac{K^{\frac{3}{2}}t^3}{n} = o(1)$.

This completes the proof of Theorem 2.3. □

2.4 Numerical experiments

We present a few numerical experiments to verify our empirical spectral density function tends to that from the Marchenko-Pastur law. In all our tests, the numerical results are computed from random vectors generated by one realization, i.e. we did not average over multiple trials.

Uncorrelated Blocks Experiments: In Figure 2.2 we show the empirical eigenvalue density for four experiments of matrices of the form of Model 1; in each case, they match up

very well with the corresponding Marchenko-Pastur density. In Figure 2.2a, the columns of matrix $X \in \mathbb{R}^{7000\times 21000}$ have $n = 10$ blocks, each block is length $d = 700$ and is of the form $\pm\sqrt{d}e_i$ for some $i \in \{1, ..., d\}$, where $\{e_i\}_{i=1}^d \in \mathbb{R}^d$ are the standard basis vectors in $\mathbb{R}^d$. This example shows that with the exchangeable criteria, it is possible for $n \ll d$. Additionally, our theorem holds as $n \to \infty$, but here we see the two densities agree extremely well and we only have $n = 10$. Similarly, in Figure 2.2b the columns of matrix $X \in \mathbb{R}^{6400\times 12800}$ have $n = 80$ blocks, each block is length $d = 80$ and is of the form $\pm\sqrt{d}e_i$ for some $i \in \{1, ..., d\}$. In Figure 2.2c, the columns of $X \in \mathbb{R}^{4000\times 16000}$ are $n = 2000$ blocks, each of length $d = 2$ where the first entry of the block is $z \sim N(0, 1)$ and the second entry is $\frac{1}{\sqrt{2}}(z^2 - 1)$. Thus the second entry is completely determined via a formula of the first entry. While this matrix has half the amount of randomness as an i.i.d. matrix of the same size, it still follows the same limiting distribution as the i.i.d. matrix. Furthermore, we see the densities match up very well even for these relatively small sized matrices. In Figure 2.2d, the columns of $X \in \mathbb{R}^{1800\times 12600}$ have $n = 600$ blocks each of length $d = 3$ where the first and second entry of the block are $\pm\frac{1}{2}$ each with probability $\frac{1}{2}$ and the third entry is a shifted XOR of the first and second (i.e. the third entry is $\frac{1}{2}$ if the first and second entries have opposite signs and it is $-\frac{1}{2}$ if the first and second entries have the same sign). In this case the variance of the entries is $\frac{1}{4}$, so it matches up with Marchenko-Pastur density with $\lambda = \frac{1}{7}$ and $\sigma^2 = \frac{1}{4}$. These figures and other experiments together suggest that having $n \geq 10$ and dimensions in the low thousands is enough for the empirical eigenvalue density of matrices of the form of Model 1 to match quite well with the corresponding Marchenko-Pastur density.

Vectorized Tensor Experiments: In Figures 2.3 and 2.4, we look at vectorized 2-tensors and 3-tensors. We see that the fourth moment of the entries is really important for the speed of convergence as $n \to \infty$. For both the 2-tensors and 3-tensors we consider three types of entries in the vector that we will tensor with itself: (1) the entries are Bernoulli ± 1 each with probability half - these entries have third moment zero and fourth moment of 1; (2) the entries are Uniform on $[-\sqrt{3}, \sqrt{3}]$ - these entries have third moment zero and fourth moment

of $\frac{9}{5}$; (3) the entries are standard normal - these entries have third moment zero and fourth moment of 3. For the 2-tensors ($t = 2$), our criteria of $\max_\alpha |\mathbb{E}[x_\alpha^4]| = o(\frac{n^{2/3}}{t^2})$ becomes: (1) $n \gg 8$; (2) $n \gg 20$; (3) $n \gg 42$. In Figure 2.3 we compare the the empirical eigenvalue density for 2-tensors with the corresponding Marchenko-Pastur density using $n = 145$. We see that the two densities match up quite well, and match up better when the entries had smaller fourth moments. For the 3-tensors ($t = 3$), that same criteria becomes: (1) $n \gg 27$; (2) $n \gg 66$; (3) $n \gg 141$. However, the computational cost is too high to make matrices with, say $\binom{141}{3} \approx 0.5$ million rows, so we had to consider much fewer rows. In Figure 2.4 we compare the the empirical eigenvalue density for 3-tensors with the corresponding Marchenko-Pastur density using $n = 45$. We see that the two densities match up quite well for the Bernoulli entry case, not very well for the uniform entry case, and very poorly for the standard normal case. This is not a surprise since the value of n we used was too small for the latter two cases. Using a super computer to test $n = 100$ does show that the empirical densities are getting closer to the Marchenko-Pastur density though, see Figure 2.5. These experiments show that while the limiting density does appear to go to the the Marchenko-Pastur density, they do not look that close for small values of n and indeed we wouldn't expect it to look close when our criteria, $\max_\alpha |\mathbb{E}[x_\alpha^4]| = o(\frac{n^{2/3}}{t^2})$, is violated.

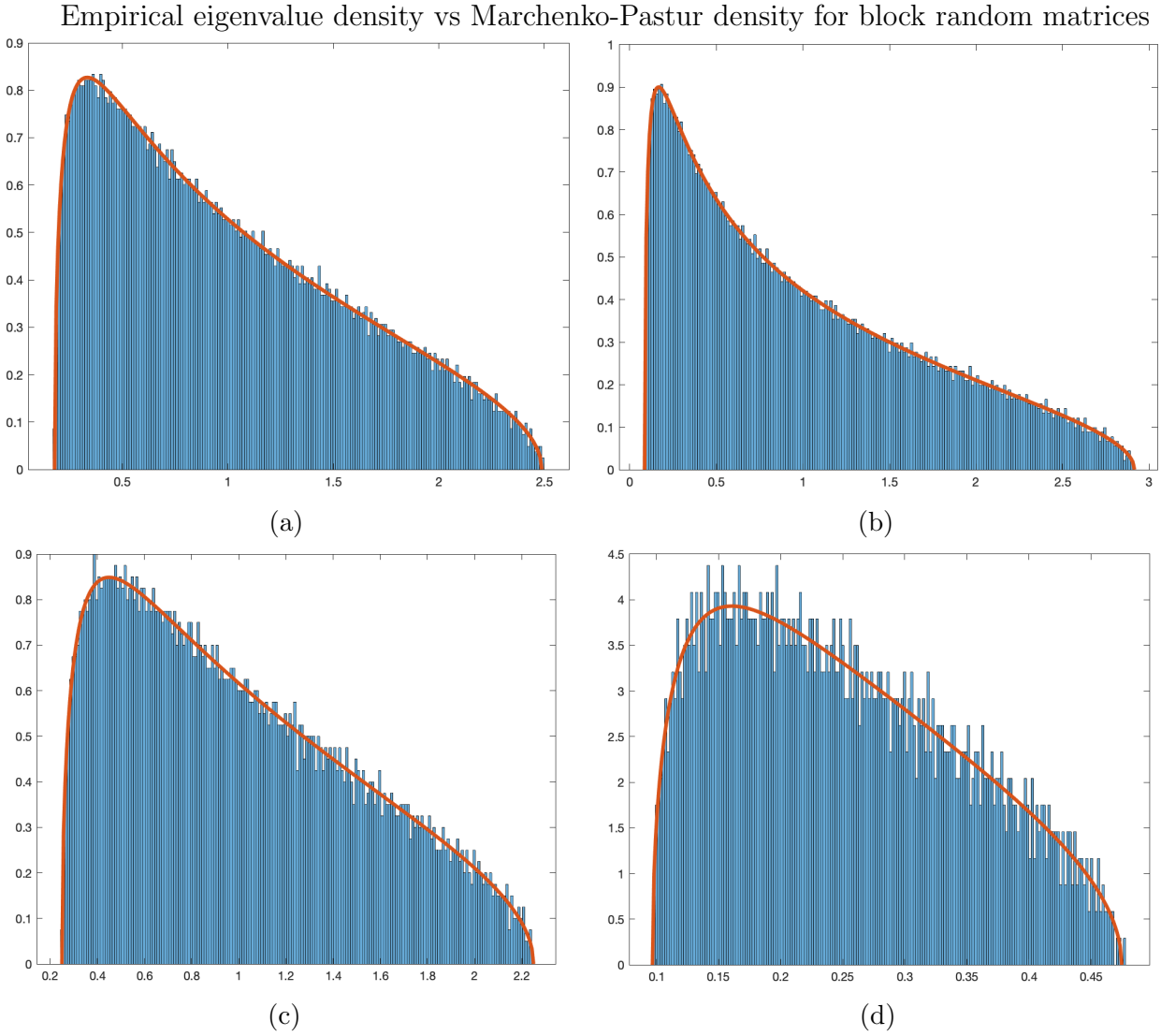
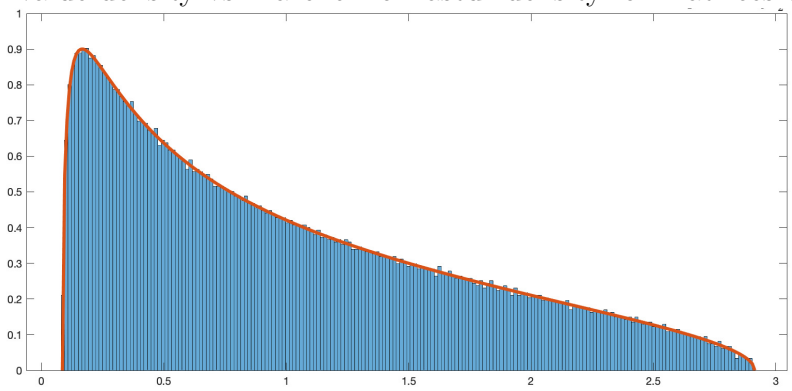
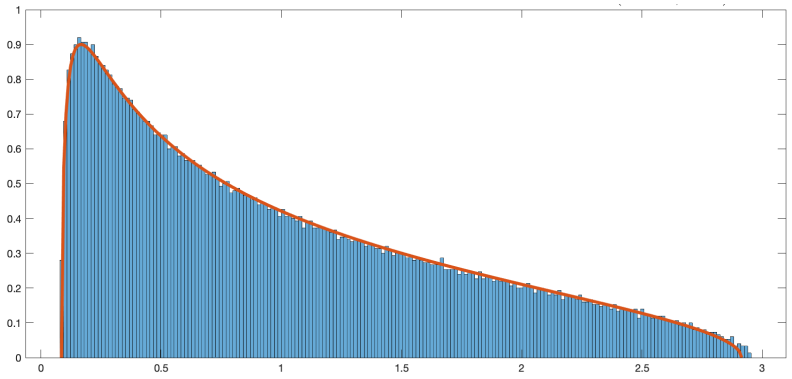


Figure 2.2: In (A), the columns of matrix $X \in \mathbb{R}^{7000\times 21000}$ have $n = 10$ blocks, each block is length $d = 700$ and is of the form $\pm\sqrt{d}e_i$ for some $i \in \{1, ..., d\}$, where $\{e_i\}_{i=1}^d \in \mathbb{R}^d$ are the standard basis vectors in $\mathbb{R}^d$. Furthermore X has three times as many columns as rows, which is why it matches up with the Marchenko-Pastur density with $\lambda = \frac{1}{3}$ and $\sigma^2 = 1$. Similarly, in (B) the columns of matrix $X \in \mathbb{R}^{6400\times 12800}$ have $n = 80$ blocks, each block is length $d = 80$ and is of the form $\pm\sqrt{d}e_i$ for some $i \in \{1, ..., d\}$. In (C), the columns of $X \in \mathbb{R}^{4000\times 16000}$ are $n = 2000$ blocks, each of length $d = 2$ where the first entry of the block is $z \sim N(0, 1)$ and the second entry is $\frac{1}{\sqrt{2}}(z^2 - 1)$. In (D), the columns of $X \in \mathbb{R}^{1800\times 12600}$ have $n = 600$ blocks each of length $d = 3$ where the first and second entry of the block are $\pm\frac{1}{2}$ each with probability $\frac{1}{2}$, and the third entry is $\frac{1}{2}$ if the first and second entries have opposite signs and $-\frac{1}{2}$ if they have the same sign.

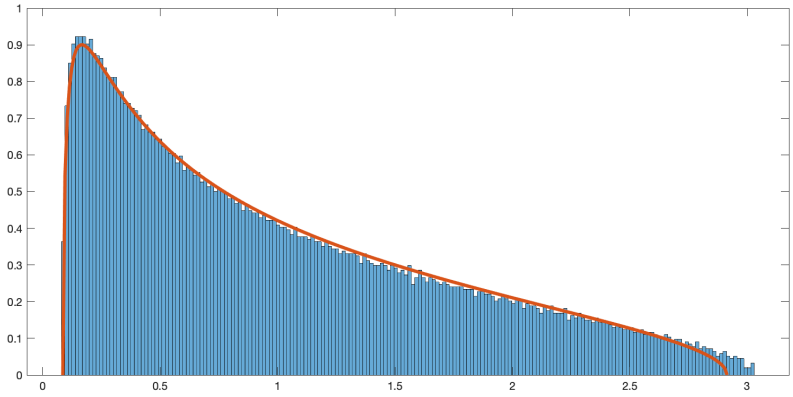
Empirical eigenvalue density vs Marchenko-Pastur density for matrices of vectorized 2-tensors



(a)



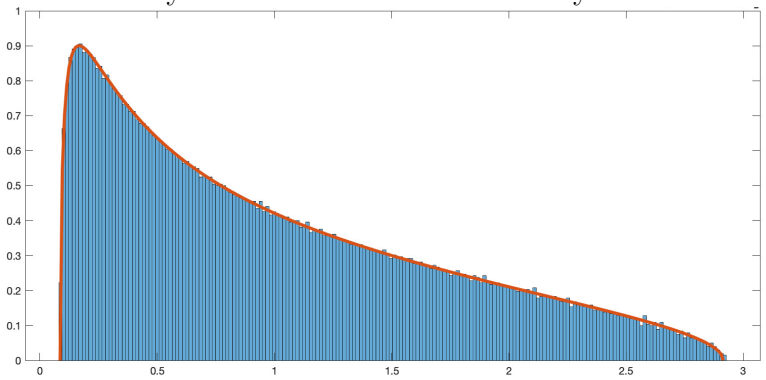
(b)



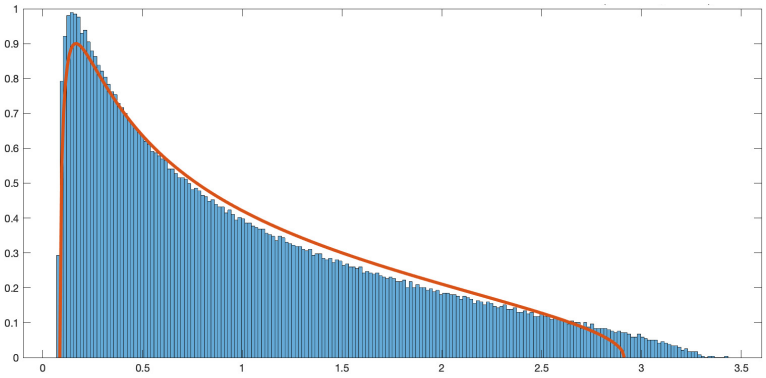
(c)

Figure 2.3: We consider vectorized 2-tensors, comparing three types of entries in the vector in $\mathbb{R}^{145}$ that we will tensor with itself: (A) the entries are Bernoulli ± 1 each with probability half; (B) the entries are Uniform on $[-\sqrt{3}, \sqrt{3}]$; (C) the entries are standard normal. In all three cases, our criteria of $\max_\alpha |\mathbb{E}[x_\alpha^4]| = o(\frac{n^{2/3}}{t^2})$ is satisfied and the empirical densities match up quite well with the Marchenko-Pastur density.

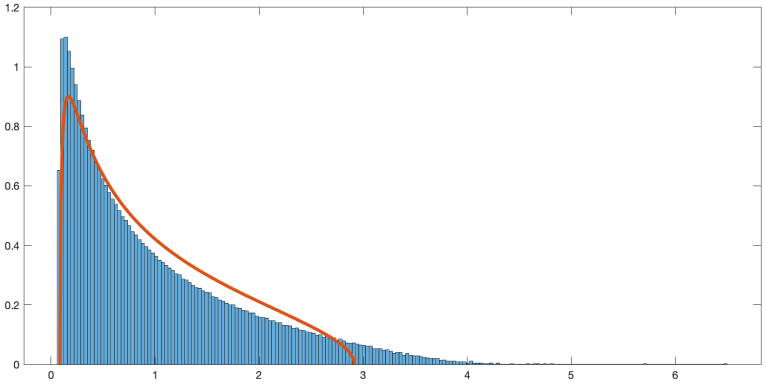
Empirical eigenvalue density vs Marchenko-Pastur density for matrices of vectorized 3-tensors



(a)



(b)



(c)

Figure 2.4: We consider vectorized 3-tensors, comparing three types of entries in the vector in $\mathbb{R}^{45}$ that we will tensor with itself: (A) the entries are Bernoulli ± 1 each with probability half; (B) the entries are Uniform on $[-\sqrt{3}, \sqrt{3}]$; (C) the entries are standard normal. Our criteria of $\max_\alpha |\mathbb{E}[x_\alpha^4]| = o(\frac{n^{2/3}}{t^2})$ is only satisfied in the first case, and is violated the most in the third case, which explains why the empirical densities in the second and third cases do not match up well with the Marchenko-Pastur density, with the third case being the worst.

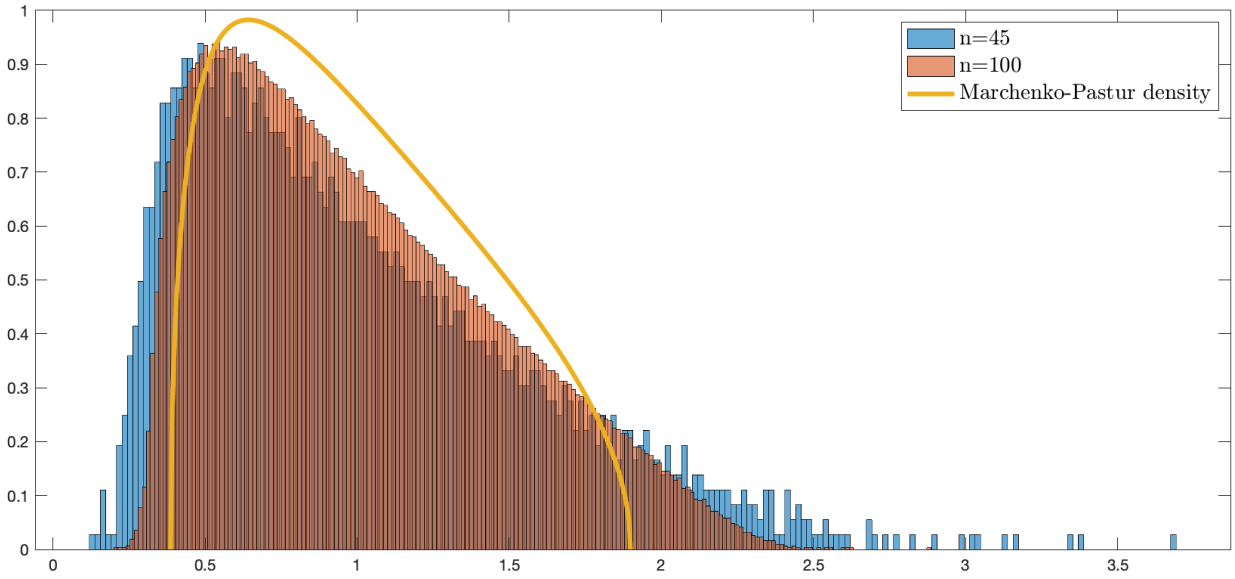


Figure 2.5: Let $V \in \mathbb{R}^{\binom{n}{3} \times \frac{1}{7}\binom{n}{3}}$ have columns from vectorized three tensors of a vector in $\mathbb{R}^n$ whose entries are Uniform on $[-\sqrt{3}, \sqrt{3}]$. We plot the ESD of $\frac{1}{\binom{n}{3}} V^T V$ for $n = 45$ and $n = 100$ and compare them to the Marchenko-Pastur density.

Chapter 3

Approximate Embeddings and the Marchenko-Pastur Law

In this chapter, we aim to find the least dimension required for approximately embedding vectors, which can be used for dimensionality reduction. Our notion of an approximate embedding will be made rigorous with the definition for vectors to be relatively root mean square (r.m.s) ε-embedded into a subspace. In the first section, we give a general lower bound for the least dimension for any collection of vectors. In the second section, we consider random vectors whose entries are i.i.d., mean 0, and variance 1. Here we give an asymptotic formula for the exact value of the least dimension. Some corollaries and the dual question, "Given a fixed dimension, k, what is the least amount of error one has to make when embedding these vectors in the best k-dimensional subspace?" will also be considered. In the third section, we show how the asymptotic formula can be determined in more general cases, where the vectors have a known covariance matrix which has a known limiting spectral distribution. In particular, we consider a matrix with a specific covariance structure of $C_{i,j} = \exp(-\frac{|i-j|}{\sigma})$ for some positive constant σ, since it was shown in [11] that the least dimension required for these vectors to be relatively r.m.s. ε-embedded in a subspace is

the discrete analog to the question of the number of terms needed in the Karhunen-Loéve expansion to approximate a random field within an ε tolerance. Lastly, we show numerical experiments which show the asymptotic formulas work very well even for a single instance and quite small dimensions. Because this works so well for small dimensions, this gives an easy numerical test that can help provide evidence for answering the question: "Does a specific covariance structure have a limiting spectral distribution or not?"

3.1 General lower bound on the least dimension required for relatively r.m.s. ε-embedding vectors

If we have p orthogonal vectors in $\mathbb{R}^m$ where $m > p$, then surely we need a p-dimensional space in order to embed them in. If we consider nearly orthogonal vectors, we can ask the same question of how many dimensions we need to embed them in. A lower bound and asymptotic lower bound on the number of dimensions was given in a paper by Alon [3]. Asymptotic upper bounds can be found using the Johnson-Lindenstrauss Lemma [24]. Suppose now that we weaken our requirement of embedding, so that the vectors don't have to be perfectly embedded, but rather we allow some specified error tolerance. We introduce the definition of relative root mean square (r.m.s.) ε-embedding of a set of vectors.

Definition 3.1. *[11] A set of vectors* $\{\mathbf{v}_i\}_{i=1}^p$ *are* **relatively root mean square (r.m.s) ε-embedded** *in a linear subspace, S, if*

$$\frac{\sum_{i=1}^p ||\mathbf{v}_i - P_S\mathbf{v}_i||_2^2}{\sum_{i=1}^p ||\mathbf{v}_i||_2^2} \leq \epsilon^2$$

where $P_S\mathbf{v}_i$ *denotes the projection of* $\mathbf{v}_i$ *in* S*.*

Definition 3.2. *[11] For a set of vectors, define* $\underline{\mathbf{N}}^\varepsilon$ *to be the least dimension of all subspaces* S *such that the vectors are relatively r.m.s.* ε*-embedded in* S*.*

The value $\underline{N}^{\varepsilon}$ is closely related to eigenvalues of a covariance matrix. Define the matrix $V = [\mathbf{v}_1, \mathbf{v}_2, ...\mathbf{v}_p] \in \mathbb{R}^{m\times p}$, and the matrix $A = V^T V \in \mathbb{R}^{p\times p}$. Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ be the eigenvalues of A. Then

$$\sum_{i=1}^{p} ||\mathbf{v}_i||_2^2 = tr(A) = \sum_{i=1}^{p} \lambda_i. \tag{3.1}$$

This gives us a new representation of the denominator in the definition of relatively r.m.s. ε-embedding. For the numerator, we know the best linear subspace, denoted by $\bar{S}_l$, of all linear subspaces of dimension l that approximates the set of vectors $\{\mathbf{v}_i\}_{i=1}^{p}$ in the least squares sense is the space spanned by the first l left singular vectors of V and satisfies

$$\sum_{i=1}^{p} ||\mathbf{v}_i - P_{\bar{S}_l}\mathbf{v}_i||_2^2 = \min_{S_l, dim(S_l)=l} \sum_{i=1}^{p} ||\mathbf{v}_i - P_{S_l}\mathbf{v}_i||_2^2 = \sum_{i=l+1}^{p} \lambda_i. \tag{3.2}$$

Combining Equations (3.1), (3.2), and the definitions of relatively r.m.s ϵ-embedding and $\underline{N}^{\varepsilon}$, we have

$$\underline{N}^{\varepsilon} = \min_{N\in\mathbb{Z}^+} \text{ such that } \frac{\sum_{i=N+1}^{p} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \leq \epsilon^2 \tag{3.3}$$

Equation (3.3) turns the problem of finding $\underline{N}^{\varepsilon}$ into an eigenvalue problem.

We now give a general lower bound for $\underline{N}^{\varepsilon}$ for any collection of vectors.

Theorem 3.1 (Bryson, Zhao, Zhong [11])**.** *Let $\{\mathbf{v}_i\}_{i=1}^{p}$ be a collection of vectors in $\mathbb{R}^m$. Let $\underline{N}^{\varepsilon}$ be the least dimension of a linear subspace such that the vectors $\{\mathbf{v}_i\}_{i=1}^{p}$ can be relatively r.m.s ε-embedded in that subspace. Then*

$$\underline{N}^{\varepsilon} \geq \frac{(\sum_{i=1}^{p} ||v_i||_2^2)^2(1-\epsilon^2)^2}{\sum_{i,j=1}^{p}(v_i \cdot v_j)^2} = \frac{tr(A)^2(1-\epsilon^2)^2}{||A||_F^2},$$

where $V = [\mathbf{v}_1, \mathbf{v}_2, ...\mathbf{v}_p] \in \mathbb{R}^{m\times p}$ and $A = V^T V$.

Proof. Notice

$$||A||_F^2 = tr(A^T A) = \sum_{i=1}^{p} \lambda_i^2 = \sum_{i=1}^{\underline{N}^\varepsilon} \lambda_i^2 + \sum_{i=\underline{N}^\varepsilon+1}^{p} \lambda_i^2 \geq \sum_{i=1}^{\underline{N}^\varepsilon} \lambda_i^2 \geq \frac{1}{\underline{N}^\varepsilon} \big(\sum_{i=1}^{\underline{N}^\varepsilon} \lambda_i \big)^2, \tag{3.4}$$

where the last inequality is from Cauchy-Schwartz.

Furthermore,

$$\sum_{i=1}^{\underline{N}^\varepsilon} \lambda_i = \sum_{i=1}^{p} \lambda_i - \sum_{i=\underline{N}^\varepsilon+1}^{p} \lambda_i \geq \sum_{i=1}^{p} \lambda_i - tr(A)\epsilon^2 = tr(A)(1-\epsilon^2).$$

Substituting this into (3.4) and solving for $\underline{N}^\varepsilon$ completes the proof. □

3.2 Asymptotic formulas for the least dimension required for relatively r.m.s. ε-embedding i.i.d. random vectors

In the special case where the vectors $\{\mathbf{v}_i\}_{i=1}^p \in \mathbb{R}^m$ have i.i.d. entries, we have a precise asymptotic formula for $\underline{N}^\varepsilon$ using the Marchenko-Pastur law (see Theorem 1.1 in Section 1.3.3). The Marchenko-Pastur law will give us the value of $\underline{N}^\varepsilon$ for a sequence of matrices that grow in size while keeping a fixed ratio of number of rows to number of columns. To be more precise, let $V_p \in \mathbb{R}^{m\times p}$ and consider a collection of matrices $\{V_p\}_{p\to\infty}$ which grow in size with a fixed ratio $p/m \to \lambda \in (0,\infty)$, then we can determine $\underline{N}^\varepsilon$ for $\lim_{p\to\infty} V_p$. However, numerical results show that the formula is very accurate even for a single realization and quite small p (and m).

Theorem 3.2 (Bryson, Zhao, Zhong [11])**.** *For a set of random vectors $\{\mathbf{v}_i\}_{i=1}^p \in \mathbb{R}^m$ whose entries are i.i.d. with mean zero and variance $= \sigma^2 < \infty$. Le $V = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p] \in \mathbb{R}^{m\times p}$ and*

let μ be the limit distribution of the eigenvalues of $\hat{A} = \frac{1}{m}V^TV$. Then the least dimension of a linear subspace, $\underline{N}^\varepsilon$, such that the vectors $\{\mathbf{v}_i\}_{i=1}^p$ can be relatively r.m.s. ε-embedded in, has the the following asymptotic formula:

$$\frac{\underline{N}^\varepsilon}{p} \to \int_y^{\lambda_+} d\mu(x), \text{ as } p \to \infty, \text{ where } y \text{ is such that } \int_{\lambda_-}^{y} x d\mu(x) = \sigma^2\epsilon^2. \tag{3.5}$$

Proof. Let $A = V^TV$ and let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ be the eigenvalues of A. Since $\sum_{i=1}^p \lambda_i = tr(A) = \sum_{i=1}^p v_i \cdot v_i = \sum_{i=1}^m \sum_{j=1}^p V_{i,j}^2 \to mp\sigma^2$ as $p, m \to \infty$, equation (3.3) becomes $\sum_{i=\underline{N}^\varepsilon+1}^p \lambda_i \leq mp\sigma^2\epsilon^2$. Let $\hat{A} = \frac{1}{m}A$ with eigenvalues $\hat{\lambda}_i = \frac{1}{m}\lambda_i$. Rewriting the previous equation in terms of $\hat{\lambda}_i$ gives $\sum_{i=\underline{N}^\varepsilon+1}^p \hat{\lambda}_i \leq p\sigma^2\epsilon^2$. Thus we would like to find the smallest integer, $\underline{N}^\varepsilon$, such that this equation holds.

Let μ be the limit measure for $\hat{A}$ as in the Marchenko-Pastur law (where $X = V^T$). We have that $\frac{1}{p}\sum_{i=\underline{N}^\varepsilon+1}^p \hat{\lambda}_i \to \int_{\lambda_-}^y x d\mu(x)$ where y is the value of the $(p - \underline{N}^\varepsilon)^{th}$-smallest eigenvalue. Therefore we would like to find the largest y such that $\int_{\lambda_-}^y x d\mu(x) \leq \sigma^2\epsilon^2$. Once we know y, we can find $\underline{N}^\varepsilon$ since $\underline{N}^\varepsilon \to p\int_y^{\lambda_+} d\mu(x)$. □

Remark. Let R^ε be the largest integer such that $\sqrt{\lambda_{R^\varepsilon}} \geq \varepsilon$ (the standard ε rank approximation of V). Under the same conditions in Theorem 3.2, using the Marchenko-Pastur law we have

$$\frac{p - R^\varepsilon + 1}{p} \to \int_{\lambda_-}^{\frac{\varepsilon^2}{m}} d\mu(x) \text{ as } p \to \infty.$$

Remark. The asymptotic formula in Theorem 3.2 says that, for any fixed tolerance $\varepsilon > 0$ in relative r.m.s. sense, the dimension, $\underline{N}^\varepsilon$, of the best linear subspace that can approximately embed a set of p random vectors in $\mathbb{R}^m$ ($m = O(p)$) with i.i.d. entries is proportional to p for $p \gg 1$. Let's denote the ratio, $\rho(\varepsilon) = \frac{\underline{N}^\varepsilon}{p}$ as a function of ε, one can compute the rate of change of $\rho(\varepsilon)$ with respect to ε. From (3.5) we have

$$\frac{d\rho(\varepsilon)}{d\varepsilon} = -\frac{\sqrt{(\lambda_+ - y)(y - \lambda_-)}}{2\pi\sigma^2\lambda y}\frac{dy}{d\varepsilon} = -\frac{2\sigma^2\varepsilon}{y}, \ \varepsilon \in (0, 1),$$

where the relation $\int_{\lambda_-}^{y} x d\mu(x) = \sigma^2\varepsilon^2$ is used for the last equality. Let $\varepsilon \to 0+$, which implies $y \to \lambda_-$, then

$$\frac{d\rho(\varepsilon)}{d\varepsilon} = \begin{cases} O(\varepsilon) & \text{if } lim_{p\to\infty}\frac{p}{m} = \lambda \neq 1 \\ O(\varepsilon^{-\frac{1}{3}}) & \text{if } lim_{p\to\infty}\frac{p}{m} = \lambda = 1 \end{cases}$$

as $\varepsilon \to 0^+$.

For $\lambda = 1$, we use the fact that

$$\frac{dy}{d\varepsilon} = \frac{4\pi\sigma^4\varepsilon}{\sqrt{y(\lambda_+ - y)}}.$$

Figure 3.1d shows numerical plots of $\rho(\varepsilon)$ for different $\lambda = \frac{p}{m}$, and we see the behavior of the slope of $\rho(\varepsilon)$ for $\lambda = 1$.

We can also ask the dual question: given k dimensions, what is the smallest possible value of ε? Meaning if we want to project our vectors onto the best k-dimensional subspace, then how much error in the relative r.m.s. sense does one have to make?

Corollary 3.3 (Bryson, Zhao, Zhong [11])**.** *Given vectors $\{\mathbf{v}_i\}_{i=1}^{p} \in \mathbb{R}^m$ whose entries are i.i.d. with mean zero and variance $= \sigma^2 < \infty$, and a fixed k, $k < m$. The relative r.m.s. error for the best k-dimensional subspace is asymptotically given by $\varepsilon = \sqrt{\frac{1}{\sigma^2}\int_{\lambda_-}^{y} x d\mu(x)}$, where y satisfies $\int_{\lambda_-}^{y} d\mu(x) = \frac{p-k}{p}$.*

Proof. Let $y = \hat{\lambda}_{k+1}$ be the value of the $(k+1)^{st}$ largest eigenvalue of $\hat{A} = \frac{1}{m}V^TV$, where $V = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p] \in \mathbb{R}^{m\times p}$. By the Marchenko-Pastur law, $p\int_{\lambda_-}^{y} d\mu(x) \to p - k$, which can be numerically solved for y. The Marchenko-Pastur law also gives that $\frac{1}{p}\sum_{i=k+1}^{p}\hat{\lambda}_i \to$

$\int_{\lambda_-}^{y} x d\mu(x)$. Since $\frac{1}{p}\sum_{i=1}^{p}\hat{\lambda}_i \to \sigma^2$, we have

$$\frac{\sum_{i=k+1}^{n}\lambda_i}{\sum_{i=1}^{p}\lambda_i} = \frac{\sum_{i=k+1}^{n}\hat{\lambda}_i}{\sum_{i=1}^{p}\hat{\lambda}_i} \to \frac{1}{\sigma^2}\int_{\lambda_-}^{y} x d\mu(x).$$

Setting this equal to ε^2 and solving for ε completes the proof. □

Remark. In practice, even if one does not have the knowledge of the probability distribution for a set of vectors, as long as the entries are approximately i.i.d., one can compute the empirical mean and variance from the data and use them to estimate $\underline{N}^{\varepsilon}$ from the above formulas.

3.3 ε-embedding random vectors with a known covariance

In this section, we study the scaling law of $\underline{N}^{\varepsilon}$ for a collection of random vectors $\mathbf{v}_i \in \mathbb{R}^m, i = 1, 2, ..., p$, with $[\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p] = V = XL^T$, where $X \in \mathbb{R}^{m\times p}$ is a random matrix whose entries are i.i.d. with mean 0 and variance 1, and $L \in \mathbb{R}^{p\times p}$ is the Cholesky decomposition of a covariance matrix $C = LL^T$. If the eigenvalues of the covariance matrix C have a limiting distribution as $p \to \infty$ and we know what it is, then we can use the general Marchenko-Pastur law (see Theorem 1.4 in Section 1.3.3) to find the empirical distribution of the eigenvalues of the sample covariance matrix V^TV. We can use this just as was done in Theorem 3.2 to obtain the explicit asymptotic scaling law for $\frac{\underline{N}^{\varepsilon}}{p}$.

Remark. While $V^TV = LX^TXL^T$ which does not look like the form for the general Marchenko-Pastur law, it is not a problem because the nonzero eigenvalues of V^TV are exactly equal to the nonzero eigenvalues of $VV^T = XL^TLX^T = XLL^TX = XCX^T$, which does look like the form for the general Marchenko-Pastur law. For more details see the work done by

Silverstein in [35].

In practice, the limiting eigenvalue distribution of C is difficult to find. This greatly restricts the number of matrices we can apply this to. Furthermore, the limiting eigenvalue distribution of C can make it impossible to find an analytical solution to the fixed point equation. In general, if the limiting eigenvalue distribution of C is diagonal with l many different values on the diagonal, then solving the fixed point equation will come down to finding the roots of an $l+1$ degree polynomial. Thus for matrices whose desired covariance matrix is diagonal with, for example, 25% of the diagonal elements are value α and 75% are value β, then we can use the general Marchenko-Pastur law as we did in Theorem 3.2. For an efficient computational algorithm to do this, see [16].

What about other matrices? Of particular interest is a random matrix with expected covariance given by C, where $C_{i,j} = \exp(\frac{-|i-j|}{\sigma})$ for any positive value of σ. Numerically calculating N^{ε} as a function of the dimension, p, for relatively small values of p, we saw that $\underline{N}^{\varepsilon}$ grows linearly with p. This led us to believe that the matrix $C \in R^{p\times p}$ given by $C_{i,j} = \exp(\frac{-|i-j|}{\sigma})$ has a limiting eigenvalue distribution function as $p \to \infty$. Indeed, it does, as was shown in [11], which led to the asymptotic scaling law

$$\frac{\underline{N}^{\varepsilon}}{p} \to \frac{2}{\pi} \arctan\left(\tanh\left(\frac{1}{2\sigma}\right) \tan\left(\frac{\pi}{2}(1-\varepsilon^2)\right)\right) \text{ as } p \to \infty.$$

This covariance matrix is of particular interest, since it was shown in [11] that $\underline{N}^{\varepsilon}$ is the discrete analog to the number of terms needed in the Karhunen-Loéve expansion to approximate a random field within tolerance ε.

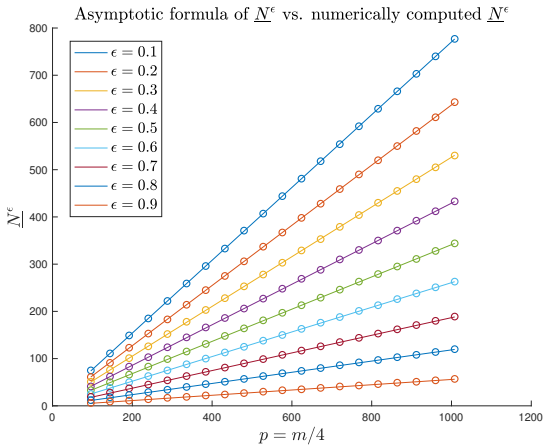
3.4 Numerical experiments

This section presents a few numerical experiments to verify the asymptotic formula for $\underline{N}^{\varepsilon}$ derived in this chapter for different scenarios. In all the tests, numerical results are computed from random vectors generated by one realization; no ensemble average is performed.

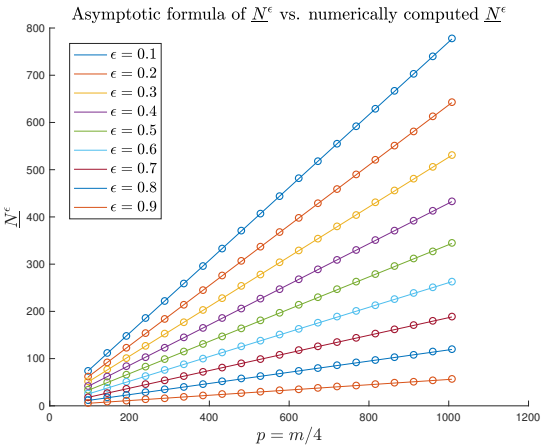
Example: random vectors with i.i.d. entries.. Figure 3.1(a), (b) plots the asymptotic formula of $\underline{N}^{\varepsilon}$ in terms of the total number of random vectors p (solid line) vs the numerically computed $\underline{N}^{\varepsilon}$ for random vectors with i.i.d. Gaussian entries and i.i.d. Bernoulli entries respectively, all with mean 0 and variance 1. In these tests, we set $p = \frac{m}{4}$ and show results for different ε. We see remarkable agreements between our asymptotic estimate in Theorem 3.2 and the numerical result even for a quite small number of random vectors in one realization. Figure 3.1(c) plots the asymptotic formula of R^{ε} (solid line) vs. the numerically computed results for different ε for random vectors with i.i.d. standard Gaussian entries. Figure 3.1(d) plots the ratio $\rho(\varepsilon) = \frac{\underline{N}^{\varepsilon}}{p}$ as a function of ε for random vectors with i.i.d. standard Gaussian entries. We can see the singularity of $\frac{d\rho(\varepsilon)}{d\varepsilon}\Big|_{\varepsilon=0}$ when $\frac{p}{m} = 1$.

Example: random vectors with given covariance.. In this test, we first generate random vectors with covariance matrix $C_{i,j} = \exp(-\frac{|i-j|}{\sigma})$ by $V = XL^T$ where X is a Gaussian matrix and $C = LL^T$. Figure 3.2(a) plots $\underline{N}^{\varepsilon}$ computed from the asymptotic formula given in section 3.3 vs. the numerically computed $\underline{N}^{\varepsilon}$ from C. Figure 3.2(b) plots $\underline{N}^{\varepsilon}$ vs. $\frac{1}{\sigma}$ for C. As we can see from the plot, there is a linear scaling regime for $\underline{N}^{\varepsilon}$ vs. $\frac{1}{\sigma}$ when σ is large compared to 1. However, when σ gets close to 1 and smaller, the entries of random matrix V become almost i.i.d. Hence $\frac{\underline{N}^{\varepsilon}}{p}$ should converge to the asymptotic estimate in Theorem 3.2 as $\sigma \to 0$. This is shown in Figure 3.3. Figure 3.2(c), (d) shows similar numerical tests for random vectors with covariance matrix $C_{i,j} = \exp(-\frac{|i-j|^2}{\sigma^2})$, although we do not have an analytical solution to compare. However, the numerical evidence suggest that $\rho(\varepsilon) = \lim_{p\to\infty} \frac{\underline{N}^{\varepsilon}}{p}$ does exist. From the relation stated in Theorem 3.2, one may deduce

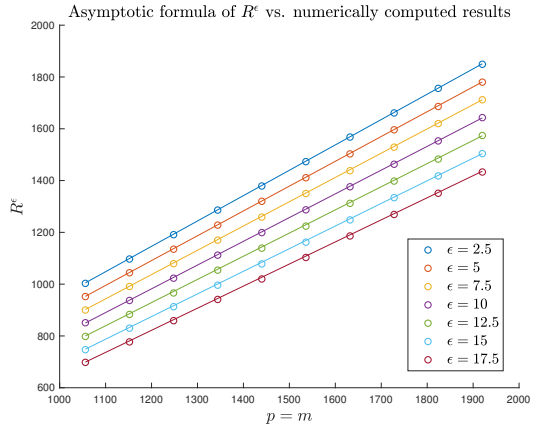
the limit distribution of the eigenvalues.



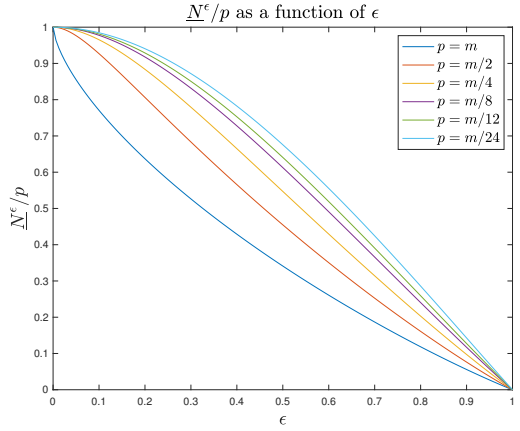
(a) random vectors with Gaussian i.i.d. entries



(b) random vectors with Bernoulli i.i.d. entries



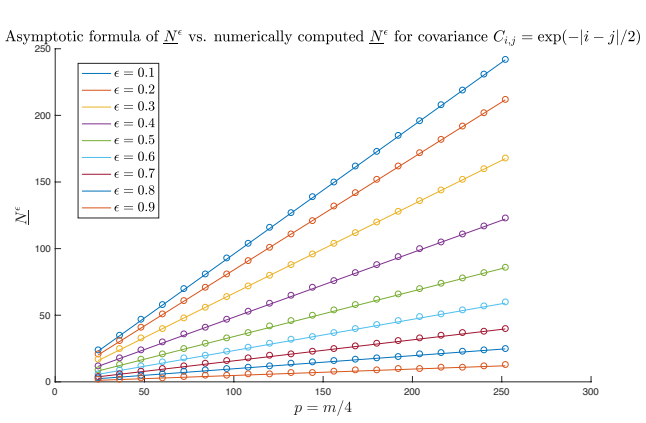
(c) R^ε



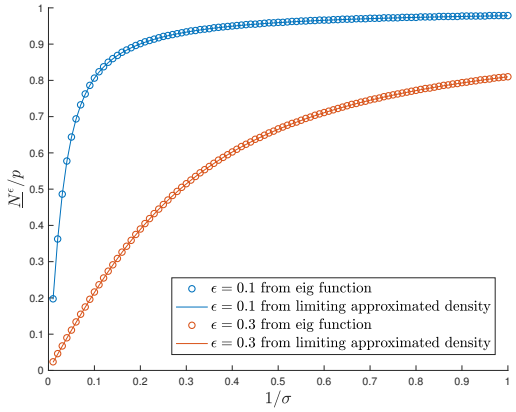
(d) $\rho(\varepsilon)$

Figure 3.1: Comparing the asymptotic formulas for N^ε and R^ε to their numerically computed values for random vectors with i.i.d. entries

for covariance matrix $C_{i,j} = \exp(\frac{-|i-j|}{\sigma})$

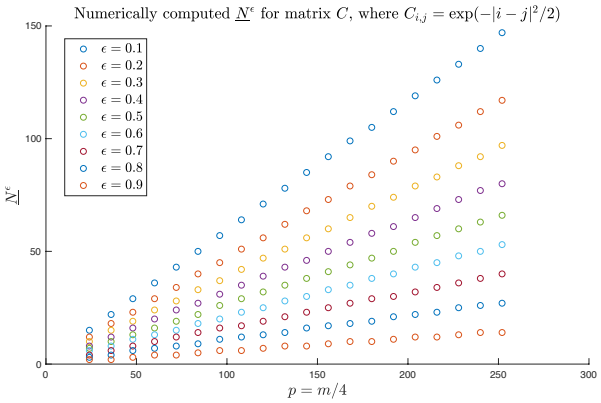


(a)

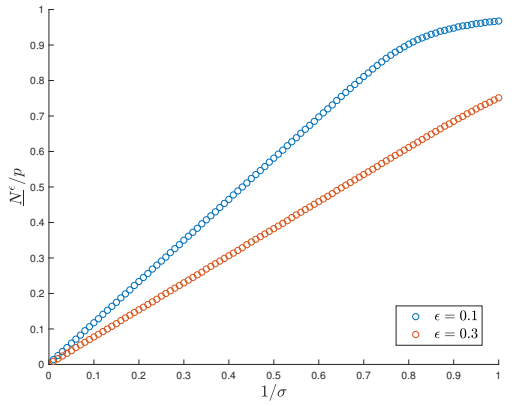


(b)

for covariance matrix $C_{i,j} = \exp(\frac{-(i-j)^2}{2\sigma^2})$



(c)



(d)

Figure 3.2: N^ε for random vectors with a given covariance

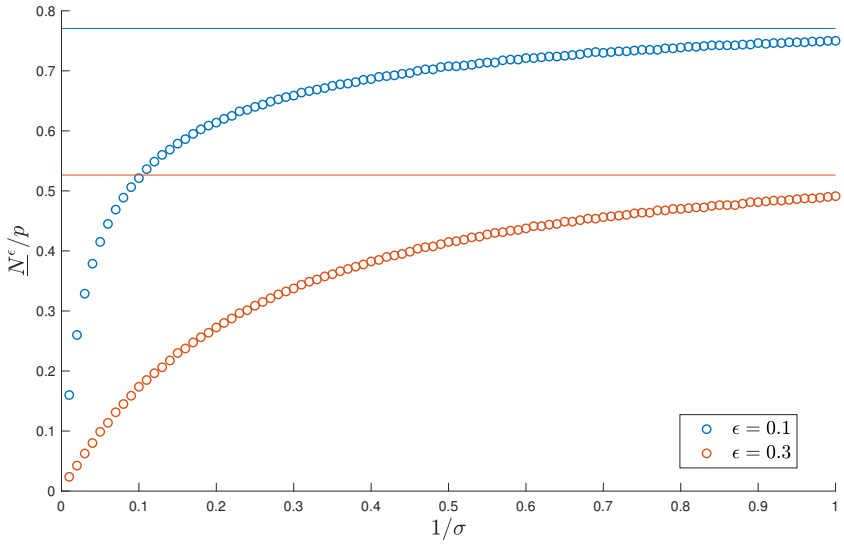


Figure 3.3: The asymptotic formula for N^{ε} for random vectors with covariance $C_{i,j} = \exp(-\frac{|i-j|}{\sigma})$ tends to the i.i.d. asymptotic estimate as $\sigma \to 0$

Bibliography

[1] E. Abbe, A. Shpilka and A. Wigderson, ReedMuller codes for random erasures and errors, *IEEE Transactions on Information Theory* 61, 5229-5252, 2015.

[2] R. Adamczak, On the Marchenko-Pastur and circular law for some classes of random matrices with dependent entries, *Electronic Journal of Probability* Vol. 16, Paper no. 37, 1068-1095, 2011.

[3] N. Alon, Problems and results in extremal combinatorics, I, Discrete Math. 273, 31-53, 2003.

[4] O. Arizmendi, I. Nechita and C. Vargas, On the asymptotic distribution of block-modified random matrices, *Journal of Mathematical Physics* 57, 015216, 2016.

[5] G. Aubrun, Random points in the unit ball of l_p^n. *Positivity,* 10(4):755?759, 2006.

[6] Z. D. Bai and J. W. Silverstein, On the empirical distribution of the eigenvalues of a class of large dimensional random matrices, *Journal of Multivariate Analysis,* 54(2):175-192, 1995.

[7] Z. D. Bai and J. W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, Springer-Verlag New York, 2010.

[8] Z. D. Bai and W. Zhou, Large sample covariance matrices without independence structures in columns, *Statistica Sinica,* 18, 425-442, 2008.

[9] P. Baldi, R. Vershynin, Polynomial threshold functions, hyperplane arrangements, and random tensors, *SIAM Journal on Mathematics of Data Science,* to appear.

[10] A. Blum, J. Hopcroft, and R. Kannan, Foundations of Data Science, Cambridge University Press, 2020.

[11] J. Bryson, H. Zhao and Y. Zhong, Intrinsic Complexity And Scaling Laws: From Random Fields to Random Vectors, *Multiscale Model. Simul.,* 17(1), 460-481, 2019.

[12] A. Chakraborty and J. Barton, Rational design of vaccine targets and strategies for HIV: a crossroad of statistical physics, biology, and medicine, *Rep. Prog. Phys.* Vol. 80, pp.32601-32620, 2017.

[13] K. L. Chung, A Course in Probability Theory, Academic Press, 2001.

[14] R. Couillet and M. Debbah, Random Matrix Methods for Wireless Communications, Cambridge University Press, 2011.

[15] V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T. Allen, M. Altfeld, M. Carrington, D. Irvine, B. Walker and A. Chakraborty, Coordinate linkage of HIV evolution reveals regions of immunological vulnerability, PNAS 108 (28) 11530-11535, 2011.

[16] E. Dobriban, Efficient computation of limit spectra of sample covariance matrices, *Random Matrices: Theory and Applications,* 04(04):1550019, 2015.

[17] R. Durrett,Probability: Theory and Examples,United Kingdom,Cambridge University Press,2019.

[18] A. Edelman and Y. Wang, Random matrix theory and its innovative appliations, *Advances in Applied Mathematics, Modeling, and Computational Science*, pp.91-116, Springer, 2013.

[19] L. Erdős, Universality of Wigner random matrices: a survey of recent results. arXiv:1004.0861.

[20] G. H. Golub and C. F. Van Loan, Matrix Computations, The Johns Hopkins University Press, 2013.

[21] F. Götze, A. Naumov and A. Tikhomirrov, Semicircle law for a class of random matrices with dependent entries, *arXiv:1211.0389v2.*

[22] F. Götze and A. Tikhomirrov, Limit theorems for spectra of random matrices with martingale structure, *Teor. Veroyatn. Primen.*, 51(1):171-192,2006.

[23] U. Greander and J. W. Silverstein, Spectral analysis of networks with random topologies, *SIAM J. Appl. Math.* 32 499-519, 1977.

[24] W.B. Johnson and J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, Contemporary mathematics, 26, 1, 1984.

[25] D. Jonnson, Some limit theorems for the eigenvalues of a sample covariance matrix, *J. Multivariate Anal.* 12 1-38, 1982.

[26] L. Laloux, P. Cizeau, J-P. Bouchaud and M. Potters, Noise dressing of financial correlation matrices, *Phys. Rev. Lett.* Vol. 83, pp.1467-1470, 1999.

[27] L. Laloux, P. Cizeau, M. Potters and J-P. Bouchaud, Random matrix theory and financial correlations, *International Journal of Theoretical and Applied Finance*, Vol. 01, No. 03, pp.391-397, 2000.

[28] V. A. Marčenko and L. A. Pastur, Distribution of eigenvalues for some sets of random matrices, *Math USSR Sbornik,* 1:457-483, 1967.

[29] J. Mingo and R. Speicher, Free Probability and Random Matrices, *Fields Institute Monographs, vol. 35, Springer* 2017.

[30] S. O'Rourke, A note on the Marchenko-Pastur law for a class of random matrices with dependent entries, *Electronic Communications in Probability* **17**, no. 28,1-13, 2012.

[31] A. Pajor and L. Pastur, On the limiting empirical measure of eigenvalues of the sum of rank one matrices with log-concave distribution, *Studia Math.*, 195(1):11?29, 2009.

[32] L. Pastur, Spectra of random self adjoint operators, *Uspehi Mat. Nauk*, 28(1(169)):3-64, 1973.

[33] V. Plerou, P. Gopikrishnan, B. Rosenow, L. Amaral, T. Guhr and H. Stanley, Random matrix approach to cross correlations in financial data, *Phys. Rev. E*, Vol 64, pp.66126-66144, 2002.

[34] D. Shlyakhtenko, Random Gaussian band matrices and freeness with amalgation, *Int. Math. Res. Not.* 20, 1013-1025, 1996.

[35] Jack W. Silverstein, Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices, *Journal of Multivariate Analysis,* 55(2):331-339, 1995.

[36] Jack W. Silverstein, The Stieltjes transform and its role in eigenvalue behavior of large dimensional random matrices, Random Matrix Theory and Its Applications, pp.1-25, 2009.

[37] T. Tao, Topics in random matrix theory, Graduate Studies in Mathematics, 132. American Mathematical Society, Providence, RI, 2012. MR-2906465

[38] K. W. Wachter, The strong limits of random matrix spectra for sample matrices of independent elements, *Ann. Probab.* 6, 1-18, 1978.

[39] E. P. Wigner, On the distribution of the roots of certain symmetric matrices, *Ann. of Math. (2)*, 67:325-327,1958.

[40] P. Yaskov, A short proof of the Marchenko-Pastur theorem, *C. R. Math. Acad. Sci. Paris*, 354:3, 319-322, 2016.

[41] P. Yaskov, Necessary and sufficient conditions for the Marchenko-Pastur theorem, *Electron. Commun. Probab.*, 21, 73-80, 2016.

[42] J. Yao, A note on a Marčenko-Pastur type theorem for time series, *Statist. and Probab. Letters* 82, 20-28, 2012.

[43] Y. Q. Yin and P. R. Krishnaiah, Limit theorems for the eigenvalues of product of large-dimensional random matrices when the underlying distribution is isotropic, *Teor. Veroyatnost. i Primenen.* 31, 394-398, 1986.

[44] Y. Q. Yin, Limiting spectral distribution for a class of random matrices, *J. Multivariate Anal.* 20, 50-68, 1986.

[45] B. Engquist and H. Zhao, Approximate separability of Green's function for high frequency Helmholtz equations, Tech. Report CAM 14-71, UCLA, 2014.

Appendix A

Matlab Code for Numerical Simulations

This appendix gives the key components of the Matlab code that was used to generate the figures of this dissertation.

A.1 Function for plotting the the Marchenko-Pastur density function

This code creates the function "MP_density" which takes in two inputs: the matrix aspect ratio $\lambda = \frac{\#rows}{\#columns} \in (0, 1]$, and the variance of the entries σ^2. The output is a plot with the corresponding Marchenko-Pastur density function.

```
function [mp_den] = MP_density(lam,var)
% Recreating the Marcenko-Pastur law:

%Input the matrix ratio (#rows/#columns) in (0,1] and the entry variance,
%then this function plots the MP density function

lam_plus = var*(1+sqrt(lam))^2;
lam_minus = var*(1-sqrt(lam))^2;
MP_den =@(x) 1/(2*pi*var)*sqrt((lam_plus-x)*(x-lam_minus))/(x*lam);
xx=linspace(lam_minus,lam_plus,300);
yy=zeros(1,length(xx));
for i = 1:length(xx)
    yy(1,i)=MP_den(xx(i));
end

plot(xx,yy, 'LineWidth', 3)
end
```

A.2 Plotting the empirical spectral density with the Marchenko-Pastur density

A.2.1 i.i.d. entries

This code plots the empirical spectral density of $\frac{1}{\#columns}XX^T$ on the same graph as the Marchenko-Pastur density with $\lambda = \frac{1}{2}$ and $\sigma^2 = 1$ where X $\in R^{1500\times3000}$ has standard Gaussian entries.

```
N=3000; %N = the number of columns
M=N/2; %M = the number of rows
X = randn(M,N);
EE=eig(1/N.*X*X');

figure();
numOfBins = 200;
histogram(EE,numOfBins, 'Normalization', 'pdf')

hold on
lam=M/N; % #rows/#cols of X
MP_density(lam,1)
```

A.2.2 Block uncorrelated entries

This code plots Figure 2.2a, which shows the remarkable accuracy the block version of the the Marchenko-Pastur law for such a small n.

```
n=10;     % n = number of blocks in a column
d=700;    % d= size of block
num_rows = n*d;
num_cols = num_rows * 3;

X = zeros(num_rows,num_cols);
for i=1:d:num_rows
    for j=1:num_cols
        basis_vec = randi(d); %gives 1,2,3,...,d with equal prob
        plus_or_minus = randi(2);
        if plus_or_minus == 1
            X(i-1+basis_vec,j) = sqrt(d);
        else
            X(i-1+basis_vec,j) = -sqrt(d);
        end
    end
end

EE=eig(1/num_cols.*X*X');
histogram(EE,200, 'Normalization', 'pdf') %200=number of bins
hold on
MP_density(num_rows/num_cols,1)
```

A.2.3 Vectorized tensor entries

This code plots the three subfigures in Figure 2.4, by changing x among the three options that appear as the first line inside the main for loop.

```
n=45;
t=3;
rows= nchoosek(n,t);
col=rows*2;
A = zeros(rows,col);
for l = 1:col
    %x=randn(n,1); % for standard normal random variables
    x=(rand(n,1)-1/2)*sqrt(12); % uniform mean 0 and variance 1 r.v.'s
    %x = (randi(2,n,1)-1)*2 -1;  % -1 and +1 each with prob 1/2 r.v.'s
    Y=zeros(rows,1);
    count=1;
    for i =1:n
        for j =i+1:n
            for k = j+1:n
                Y(count) = x(i)*x(j)*x(k);
                count=count+1;
            end
        end
    end
    A(:,l)=Y;
end
```

```
sample_cov = (1/col)*(A*A');
EE=eig(sample_cov);
histogram(EE,200, 'Normalization', 'pdf') %200=number of bins
hold on
lam=rows/col;
MP_density(lam,1)
```

A.3 Plotting the theoretically calculated N^{ϵ} vs. the actual N^{ϵ}, as done in Figure 3.1a

Here we give a Matlab function called "binary_to_find_y" which essentially does a binary search to find y, where y is the same y stated in Theorem 3.2. Then we give the code used to produce Figure 3.1a, which compares our calculated version of $\underline{N}^{\varepsilon}$ with true values of $\underline{N}^{\varepsilon}$ for a variety of values of ε.

```
function [y ] = binary_to_find_y( TenTarget,lam )
k = 1/lam;
funNox = @(x) sqrt(((1+sqrt(lam))^2-x).*(x-(1-sqrt(lam))^2))./(2*pi*lam);
y=(1-sqrt(1/k))^2;
target = 0.05*TenTarget;
for i = 1:6
    while sqrt(integral(funNox, (1-sqrt(1/k))^2, y)) < target && ...
            sqrt(integral(funNox, (1-sqrt(1/k))^2, y+1/10^i)) <= target
        y=y+1/10^i;
    end
end
```

```
lam=1/1;  %p/m=lam
k=20; %the number of times we change p and m
t=48; %step size for m
mvalues=zeros(1,k);
pvalues = zeros(1,k);
for j = 1:k
    mvalues(j)=48+t*j;
    pvalues(j)=lam*mvalues(j);
end


num_diff_epsilons = 9;
actualNeps = zeros(num_diff_epsilons, k);
calslope = zeros(num_diff_epsilons,1);
xcalc = zeros(num_diff_epsilons, 2);
ycalc = zeros(num_diff_epsilons, 2);
for h = 1:num_diff_epsilons
    epstol = 0.1*h;  %set the epsilon-embedding tolerance

    %For variance 1-this gives the 2 endpoints points (xcalc(h,1),ycalc(h,1)
    %and (xcalc(h,2),ycalc(h,2) for drawing the calculated slope
    fun = @(x) sqrt(((1+sqrt(lam))^2-x).*(x-(1-sqrt(lam))^2))./(2*lam*pi.*x);
    y=binary_to_find_y(epstol*20,lam);
    calslope(h)=integral(fun, y, (1+sqrt(lam))^2);
    xcalc(h,:) = [pvalues(1), pvalues(k)];
    ycalc(h,:) = [pvalues(1)*calslope(h,1) , pvalues(k)*calslope(h,1)];
    sumofsquares = zeros(1,k);
```

```
    for j = 1:k
        m = mvalues(j);
        p = pvalues(j);
        l=min(m,p);
        v = randn(m,p);       % Gaussian random iid entries
        A=1/m.*v'*v;
        frob = norm(A,'fro');
        sumofsquares(j)=frob^2;
        [U,S,V] = svds(v,l);
        singvaluessqu=zeros(1,l);
        svsqusum=0;
        for i = 1:l
            singvaluessqu(i)=(S(i,i))^2;
            svsqusum=svsqusum+singvaluessqu(i);
        end
        svsqusum2=svsqusum; %copy the sum of the squares
        count=0;
        for i = 1:l
            if svsqusum2/svsqusum > epstol^2
                svsqusum2=svsqusum2-singvaluessqu(i);
            else
                count=count+1;
            end
        end
        actualNeps(h,j) = l-count;
    end
end
```

```
%Plotting:
colors = get(gca,'colororder');
hold on;
ax = gca;
%plot the lines
for h = 1:num_diff_epsilons
    ax.ColorOrderIndex = h;
    plot(xcalc(h,:),ycalc(h,:));
end
%plot the true values of N^{\epsilon}
for h=1:num_diff_epsilons
    ax.ColorOrderIndex = h;
    scatter(pvalues,actualNeps(h,:));
end
```